ISSN 1847-3938

mipro 2016

organizer upro



39th

international convention

May 30 - June 03, 2016, Opatija – Adriatic Coast, Croatia

Lampadem tradere



mipro - path to knowledge and innovation

mipro proceedings

My profession.
My organization.
My IEEE.



Discover the benefits of IEEE membership.

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

Join today
www.ieee.org/join

IEEE

MIPRO 2016

39th International Convention

May 30 – June 03, 2016
Opatija, Croatia

Proceedings

Conferences:

Microelectronics, Electronics and Electronic Technology /MEET

Distributed Computing, Visualization and Biomedical Engineering /DC VIS

Telecommunications & Information /CTI

Special Session on Future Networks and Services /FNS

Computers in Education /CE

Computers in Technical Systems /CTS

Intelligent Systems /CIS

Special Session on Biometrics & Forensics & De-Identification and Privacy Protection /BiForD

Information Systems Security /ISS

Business Intelligence Systems /miproBIS

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

MIPRO Junior - Student Papers /SP

Edited by:
Petar Biljanović

International Program Committee

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

organized by

MIPRO Croatian Society

technical cosponsorship

IEEE Region 8

under the auspices of

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

patrons

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

sponsors

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

All papers are published in their original form

For Publisher:

Petar Biljanović

Publisher:

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

Printed by:

GRAFIK, Rijeka

ISBN 978-953-233-087-8

Copyright © 2016 by MIPRO

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

CONTENTS

LIST OF PAPER REVIEWERS

LIST OF AUTORS

FOREWORD

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

INVITED PAPER

(Si)GeSn Nanostructures for Optoelectronic Device Applications 5
I.A. Fischer, F. Oliveira, A. Benedetti, S. Chiussi, J. Schulze

PAPERS

Thermoelectric Properties of Polycrystalline WS_2 and Solid Solutions of $WS_{2-y}Se_y$ Types 11
G.E. Yakovleva, A.I. Romanenko, A.S. Berdinsky, A.Yu. Ledneva, V.A. Kuznetsov, M.K. Han, S.J. Kim, V.E. Fedorov

Piezoresistive Effect in Polycrystalline Bulk and Film Layered Sulphide $W_{0.95}Re_{0.05}S_2$ 16
V.A. Kuznetsov, A.I. Romanenko, A.S. Berdinsky, A.Yu. Ledneva, S.B. Artemkina, V.E. Fedorov

Luminescent Diagnostics in the NIR-region on a Base of Yb-porphyrin Complexes 20
V.D. Rumyantseva, I.P. Shilov, Yu.V. Alekseev, A.S. Gorshkova

Simulation Study of the Composite Silicon Solar Cell Efficiency Sensitivity to the Absorption Coefficients and the Thickness of intrinsic Absorber Layer 24
V. Tudić, N. Posavec

The Investigation of Influence of Localized States on a-Si:H p-i-n Photodiode Transient Response to Blue Light Impulse with Blue Light Optical Bias 30
M. Čović, V. Gradišnik, Ž. Jeričević

Analysis of Electrical and Optical Characteristics of InP/InGaAs Avalanche Photodiodes in Linear Regime by a New Simulation Environment 34
T. Knežević, T. Suligoj

Design of Passive-Quenching Active-Reset Circuit with Adjustable Hold-Off Time for Single-Photon Avalanche Diodes 40
I. Berdalović, Ž. Osrečki, F. Šegmanović, D. Grubišić, T. Knežević, T. Suligoj

Impact of the Emitter Polysilicon Thickness on the Performance of High-Linearity Mixers with Horizontal Current Bipolar Transistors 46
J. Žilak, M. Koričić, H. Mochizuki, S. Morita, T. Suligoj

Fully-integrated Voltage Controlled Oscillator in Low-cost HCBT Technology 51
M. Koričić, J. Žilak, H. Mochizuki, S. Morita, T. Suligoj

Variable-Gain Amplifier for Ultra-Low Voltage Applications in 130nm CMOS Technology 57
D. Arbet, M. Kováč, L. Nagy, V. Stopjaková, M. Šovčík

Relaxation Oscillator Calibration Technique with Comparator Delay Regulation 63
J. Mikulić, G. Schatzberger, A. Barić

A Bootstrap Circuit for DC–DC Converters with a Wide Input Voltage Range in HV-CMOS 68
N. Mitrović, R. Enne, H. Zimmermann

A Fractional-N Subsampling PLL based on a Digital-to-Time Converter 72
N. Markulic, K. Raczkowski, P. Wambacq, J. Craninckx

Infrared Protection System for High-Voltage Testing of SiC and GaN FETs used in DC-DC Converters 78
F. Hormot, J. Bačmaga, A. Barić

Optimal Conduction Angle of an E-PHEMT Harmonic Frequency Multiplier 82
K. Martinčić

Ultra-Wideband Transmitter Based on Integral Pulse Frequency Modulator
T. Matić, M. Herceg, J. Job, L. Šneler 86

Design of a Transmitter for High-Speed Serial Interfaces in Automotive Micro-Controller 90
A. Bandiziol, W. Grollitsch, F. Brandonisio, R. Nonis, P. Palestri

Application of the Calculation-Experimental Method in the Design of Microwave Filters 95
A.S. Geraskin, A.N. Savin, I.A. Nakrap, V.P. Meshchanov

Minimax Design of Multiplierless Sharpened CIC Filters Based on Interval Analysis 100
G. Molnar, A. Dudarin, M. Vučić

Minimization of Maximum Electric Field in High-Voltage Parallel-Plate Capacitor 105
R. Blečić, Q. Diduck, A. Barić

Modelling SMD Capacitors by Measurements 110
R. Mišlov, M. Magerl, S. Fratte-Sumper, B. Weiss, C. Stockreiter, A. Barić

Impact of Capacitor Dielectric Type on the Performance of Wireless Power Transfer System 116
D. Vinko, P. Oršolić

Switching Speed and Stress Analysis for Fixed-fixed Beam Based Shunt Capacitive RF MEMS Switches 120
A. Kumar A., R. R

Performance Analysis of Micromirrors - Lift-off and von Mises Stress 126
S. Finny, R. R

Material and Orientation Optimization for Quality Factor Enhancement of BAW Resonators 130
R. Raj R.S., R. R

Impact of Propagation Medium on Link Quality for Underwater and Underground Sensors 135
G. Horvat, D. Vinko, J. Vlaović

Electrical Field Intensity Model on the Surface of Human Body for Localization of Wireless Endoscopy Pill 141
B. Lukovac, A. Koren, A. Marinčić, D. Šimunić

Wide Band Current Transducers in Power Measurment Methods - an Overview 146
R. Malarić, Ž. Martinović, M. Dadić, P. Mostarac, Ž. Martinović

Laboratory Model for Design and Verification of Synchronous Generator Excitation Control Algorithms 152
S. Tusun, I. Erceg, I. Sirotić

The European Project SolarDesign Illustrating the Role of Standardization in the Innovation System 158
W. Brenner, N. Adamovic

Open Public Design Methodology and Design Process 164
D. Rembold, S. Jovalekic

DISTRIBUTED COMPUTING, VISUALIZATION AND BIOMEDICAL ENGINEERING

INVITED PAPER

Views on the Role and Importance of Dew Computing in the Service and Control Technology 175
Z. Šojat, K. Skala

PAPERS

DISTRIBUTED COMPUTING AND CLOUD COMPUTING

Parameters That Affect the Parallel Execution Speed of Programs in Multi-Core Processor Computers 185
V. Xhafa, F. Dika

Federated Computing on the Web: the UNICORE Portal 190
M. Petrova-El Sayed, K. Benedyczak, A. Rutkowski, B. Schuller

Problem-Oriented Scheduling of Cloud Applications: PO-HEFT Algorithm Case Study 196
E.A. Nepovinnykh, G.I. Radchenko

Towards a Novel Infrastructure for Conducting High Productive Cloud-Based Scientific Analytics 202
P. Brezany, T. Ludescher, T. Feilhauer

An OpenMP Runtime Profiler/Configuration Tool for Dynamic Optimization of the Number of Threads 208
T. Dancheva, M. Gusev, V. Zdravevski, S. Ristov

An Effective Task Scheduling Strategy in Multiple Data Centers in Cloud Scientific Workflow 214
E.I. Djebbar, G. Belalem

Visualisation in the ECG QRS Detection Algorithms 218
A. Ristovski, A. Guseva, M. Gusev, S. Ristov

Analysis and Comparison of Algorithms in Advanced Web Clusters Solutions 224
D. Alagić, K. Arbanas

Metamodeling as an Approach for Better Computer Resources Allocation in Web Clusters 230
D. Alagić, D. Maček

Showers Prediction by WRF Model above Complex Terrain 236
T. Davitashvili, N. Kutaladze, R. Kvatadze, G. Mikuchadze, Z. Modebadze, I. Samkharadze

Methods and Tools to Increase Fault Tolerance of High-Performance Computing Systems 242
I.A. Sidorov

Logical-Probabilistic Analysis of Distributed Computing Reliability 247
A.G. Feoktistov, I.A. Sidorov

Distributed Graph Reduction Algorithm with Parallel Rigidity Maintenance 253
D. Sušanj, D. Arbula

Architecture of Virtualized Computational Resource Allocation on SDN-enhanced Job Management System Framework 257
Y. Watashiba, S. Date, H. Abe, K. Ichikawa, Y. Kido, H. Yamanaka, E. Kawai, S. Shimojo

Near Real-time Detection of Crisis Situations 263
S. Girtelschmid, A. Salfinger, B. Pröll, W. Retschitzegger, W. Schwinger

Automatic Protocol Based Intervention Plan Analysis in Healthcare 269
M. Kozlovszky, L. Kovács, K. Batbayar, Z. Garaguly

Using Fourier and Hartley Transform for Fast, Approximate Solution of Dense Linear Systems 274
Ž. Jeričević, I. Kožar

Procedural Generation of Mediterranean Environments 277
N. Mikuličić, Ž. Mihajlović

Energy-Aware Power Management of Virtualized Multi-core Servers through DVFS and CPU Consolidation 283
H. Rostamzadeh Hajilari, M.M. Talebi, M. Sharifi

Human Posture Detection Based on Human Body Communication with Muti-carriers Modulation 289
W. Ni, Y. Gao, Ž. Lučev Vasić, S.H. Pun, M. Cifrek, M.I. Vai, M. Du

SAT-Based Search for Systems of Diagonal Latin Squares in Volunteer Computing Project SAT@home 293
O. Zaikin, S. Kochemazov, A. Semenov

Architectural Models for Deploying and Running Virtual Laboratories in the Cloud 298
E. Afgan, A. Lonie, J. Taylor, K. Skala, N. Goonasekera

A CAD Service for Fusion Physics Codes 303
M. Telenta, L. Kos

Correlation between Attenuation of 20 GHz Satellite Communication Link and Liquid Water Content in the Atmosphere 308
M. Kolman, G. Kosec

Practical Implementation of Private Cloud with Traffic Optimization 314
D.G. Grozev, M.P. Shopov, N.R. Kakanakov

Improving Data Locality for NUMA-Agnostic Numerical Libraries 320
P. Zinterhof

Use Case Diagram Based Scenarios Design for a Biomedical Time-Series Analysis Web Platform 326
A. Jović, D. Kukolja, K. Jozić, M. Cifrek

Augmented Reality for Substation Automation by Utilizing IEC 61850 Communication 332
M. Antonijević, S. Sučić, H. Keserica

Innovation of the Campbell Vision Stimulator with the Use of Tablets 337
J. Brozek, M. Jakes, V. Svoboda

Classification of Scientific Workflows Based on Reproducibility Analysis 343
A. Bánáti, P. Kacsuk, M. Kozlovszky

Dynamic Execution of Scientific Workflows in Cloud 348
E. Kail, J. Kovács, M. Kozlovszky, P. Kacsuk

FPGA Kernels for Classification Rule Induction 353
P. Škoda, B. Medved Rogina

VISUALIZATION SYSTEMS

Prototyping of Visualization Designs of 3D Vector Fields Using POVRay Rendering Engine 361
J. Opiła

New Cybercrime Taxonomy of Visualization of Data Mining Process 367
M. Babič, B. Jerman-Blažič

Visual Representation of Predictions in Software Development Based on Software Metrics History Data 370
B. Popović, A. Balota, Dž. Strujić

Interaction with Virtual Objects in a Natural Way 376
I. Prazina, K. Balić, K. Pršeš, S. Rizvić, V. Okanović

Bone Shape Characterization Using the Fourier Transform and Edge Detection in Digital X-Ray Images 380
D. Sušanj, G. Gulan, I. Kožar, Ž. Jeričević

GIS in the e-Government Platform to Enable State Financial Subsidies Data Transparency 383
M. Kranjac, U. Sikimić, I. Simić, M. Paroški, S. Tomić

Evaluation of Caching Techniques for Video on Demand in Named Data Networks 388
K. Jakimoski, S. Arsenovski, L. Gorachinova, S. Chungurski, O. Iliev, L. Djinevski, E. Kamcheva

BIOMEDICAL ENGINEERING

Diagnostic of Asthma Using Fuzzy Rules Implemented in Accordance with International Guidelines and Physicians Experience 395
A. Badnjević, L. Gurbeta, M. Cifrek, D. Marjanović

Robust Beat Detection on Noisy Differential ECG 401
P. Lavrič, M. Depolli

Classification of Asthma Using Artificial Neural Network 407
A. Badnjević, L. Gurbeta, M. Cifrek, D. Marjanović

Brain-Computer Interface Based on Steady-State Visual Evoked Potentials 411
K. Friganović, M. Medved, M. Cifrek

Comparison of Wireless Electrocardiographic Monitoring and Standard ECG in Dogs 416
A. Krvavica, Š. Likar, M. Brložnik, A. Domanjko-Petrič, V. Avbelj

A Medical Cloud 420
J. Tasič, M. Gusev, S. Ristov

A Hospital Cloud-Based Open Archival Information System for the Efficient Management of HL7 Big Data 426
A. Celesti, M. Fazio, A. Romano, M. Villari

Recognition and Adjustment for Strip Background Baseline in Fluorescence Immuno-chromatographic Detection System 432
Y. Gao, C. Lin, S.H. Pun, M.I. Vai, M. Du

Agile Development of a Hospital Information System 436
S.L.R. Vrhovec

SOA Based Interoperability Component for Healthcare Information System 442
D. Kučak, G. Đambić, V. Kokanović

Wireless Intrabody Communication Sensor Node Realized Using PSoC Microcontroller 446
F. Grilec, Ž. Lučev Vasić, W. Ni, Y. Gao, M. Du, M. Cifrek

Detection of Heart Rate Variability from a Wearable Differential ECG Device 450
J. Slak, G. Kosec

Penetration of the ICT Technology to the Health Care Primary Sector – Ljubljana PILOT 456
T. Poplas Susič, U. Stanič

Image-Based Metal Artifact Reduction in CT Images 462
A. Šerifović-Trbalić, A. Trbalić

New Algorithm for Automatic Determination of Systolic and Diastolic Blood Pressures in Oscillometric Measurements 467
V. Jazbinšek

TGTP-DB – a Database for Extracting Genome, Transcriptome and Proteome Data Using Taxonomy 472
K. Križanović, M. Marinović, A. Bulović, R. Vaser, M. Šikić

Development and Perspectives of Biomedical Engineering in South East European Countries 477
A. Badnjević, L. Gurbeta

Clustering of Heartbeats from ECG Recordings Obtained with Wireless Body Sensors 481
A. Rashkovska, D. Kocev, R. Trobec

Heart Rate Analysis with NevroEkg 487
M. Mohorčič, M. Depolli

TELECOMMUNICATIONS & INFORMATION

FNS • SPECIAL SESSION ON FUTURE NETWORKS AND SERVICES

PAPERS

A Survey of IoT Cloud Providers 497
T. Pflanzner, A. Kertesz

QoS-Aware Deployment of Data Streaming Applications over Distributed Infrastructures 503
M. Nardelli

QoS-Aware Application Placement Over Distributed Cloud 509
F. Bianchi, F. Lo Presti

Energy-Aware Control of Server Farms 515
M.E. Gebrehiwot, S. Aalto, P. Lassila

SDN Based Service Provisioning Management in Smart Buildings 521
M. Tošić, O. Iković, D. Bošković

TELECOMMUNICATIONS & INFORMATION

INVITED PAPER

Time Series Analysis and Possible Applications 531
M. Ivanović, V. Kurbalija

PAPERS

WIRELESS COMMUNICATIONS AND TECHNOLOGIES

Wireless Resonant Power Transfer – An Overview 543
Ž. Martinović, M. Dadić, R. Malarić, Ž. Martinović

Investigation of a Small Handheld PCB Nesting Two Antennas NFC 13.56 MHz and to RF 868 MHz 550
L.A. Iliev, I.S. Stoyanov, T.B. Iliev, E.P. Ivanova, Gr.Y. Mihaylov

The Coverage Belt for Low Earth Orbiting Satellites 554
S. Cakaj

The Investigation of the Effect of the Carrier Frequency Offset (CFO) in SC-FDMA System 558
N. Taşpinar, M. Balki

Performance Analysis of Low Density Parity Check Codes Implemented in Second Generations of Digital Video Broadcasting Standards 562
Gr.Y. Mihaylov, T.B. Iliev, E.P. Ivanova, I.S. Stoyanov, L.A. Iliev

DATA AND IMAGE ANALYSIS

Iterative Denoising of Sparse Images 569
I. Stanković, I. Orović, S. Stanković, M. Daković

Compressive Sensing Based Image Processing in TrapView Pest Monitoring System 574
M. Marić, I. Orović, S. Stanković

Big Data Analytics for Communication Service Providers 579
D. Šipuš

Role of Data Analytics in Utilities Transformation 584
V. Čačković, Ž. Popović

Using MEAN Stack for Development of GUI in Real-Time Big Data Architecture 590
M. Štajcer, M. Štajcer, D. Oreščanin

NETWORK TECHNOLOGIES

A Survey on Physical Layer Impairments Aware Routing and Wavelength Assignment Algorithms in Transparent Wavelength Routed Optical Networks 599
H. Dizdarević, S. Dizdarević, M. Škrbić, N. Hadžiahmetović

A Survey on Transition from GMPLS Control Plane for Optical Multilayer Networks to SDN Control Plane 606
S. Dizdarević, H. Dizdarević, M. Škrbić, N. Hadžiahmetović

About the Telco Cloud Management Architectures 614
I. Nenadić, D. Kobal, D. Palata

CPE Virtualization by Unifying NFV, SDN and Cloud Technologies 622
P. Cota, J. Šabec

Soft Sensors in Wireless Networking as Enablers for SDN Based Management of Content Delivery 628
M. Tošić, O. Ikković, D. Bošković

A FIRM Approach for Software-Defined Service Composition 634
P. Kathiravelu, T. Galinac Grbac, L. Veiga

Test Environment & Application as a Service 640
T. Žitnik, M. Galin, G. Pauković, I. Dević, R. Čižmar, Z. Bosić

Workaround Solutions Used During PSTN Migration of Customers to IMS Network 644
N. Štokić

Development of the Generic OFDM Based Transceiver in the LabView Software Environment 650
D. Hamidović, N. Suljanović

The Challenge of Cellular Cooperative ITS Services Based on 5G Communications Technology 656
Z. Kljaić, P. Škorput, N. Amin

NETWORKS PERFORMANCES

Performance Evaluation of Different Scheduling Algorithms in LTE Systems 667
A. Marinčić, D. Šimunić

Performance Analysis of LTE Networks with Random Linear Network Coding 673
T.D. Assefa, K. Kralevska, Y. Jiang

VoLTE E2E Performance Management 679
D. Klobučarević, Ž. Klobučarević, D. Belošić

Balancing Security and Blocking Performance with Reconfiguration of the Elastic Optical Spectrum 684
S. Kumar Singh, W. Bziuk, A. Jukan

Ensuring Continuous Operation of Critical Process of Remote Control System at the Level of Network Connectivity 690
I. Fosić, D. Budiša

IoT PLATFORM AND APPLICATIONS

Requirements and Challenges in Wireless Network's Performance Evaluation in Ambient Assisted Living Environments 699
A. Koren, D. Šimunić

Long Term Evolution as a Precondition for Internet of Postal Things 703
A. Kosovac, A. Veispahić, M. Berković

Advanced Sensing and Internet of Things in Smart Cities 707
D. Capeska Bogatinoska, R. Malekian, J. Trengoska, W. Asiama Nyako

Security Challenges of the Internet of Things 713
M. Weber, M. Boban

Promoting Health for Chronic Conditions: a Novel Approach That Integrates Clinical and Personal Decision Support 719
I. Lasorsa, M. Ajčević, P. D'Antrassi, G. Carlini, A. Accardo, S. Marceglia

A Taxonomy of Localization Techniques Based on Multidimensional Scaling 724
B. Risteska Stojkoska

Distributed Real-Time Lift Kinematic Monitoring Using COTS Smartphones 730
N. Miškić-Pletenac, K. Lenac

Mobile Devices as Authentic and Trustworty Sources in Multi-Agent Systems 736
V. Vyroubal, A. Stančić, I. Grgurević

SOFTWARE ENGINEERING

Comparative Analysis of Functional and Object-Oriented Programming 745
D. Alić, S. Omanović, V. Giedrimas

Improving the Composition and Assembly of APIs in Service Dominant Ecosystem Environments 751
D. Ramljak

Service Level Agreement - SLA, raspoloživost servisa i kvalitet sistema u telekomunikacijama 755
D. Glamočanin

Challenges of a Service Transition in Multi Domain Environment 761
I. Golub, B. Radojević

Teaching "Ten Commandments" of Software Engineering 766
Z. Putnik, M. Ivanović, Z. Budimac, K. Bothe

Methodologies for Development of Mobile Applications 772
Z. Stapić, M. Mijač, V. Strahonja

Upravljanje promjenama na primjeru telekom operatera u Jugoistočnoj Evropi 777
A. Gabela

Tehnologije integracije informacijskih sustava 783
A. Stojanović, N. Lazić, Ž. Kovačević

TELECOM PRODUCTS, SERVICES AND MARKET

Restructuring of Telco Products 791
I. Vrbovčan, T. Pavić, M. Šoša Anić

Moving from Network-Centric toward Customer-Centric CSPs in Bosnia and Herzegovina 794
N. Banović-Ćurguz, D. Ilišević

Future Communication Model: Challenges and Opportunities for Society as a Whole 800
D. Ilišević, N. Banović-Ćurguz

Some Aspects of Network Management System for Video Service 805
O. Jukić, I. Heđi

Influence of OTT Service Providers on Croatian Telecomunication Market 809
I. Dražić Lutilsky, M. Ivić

Implementing Shared Service Center in Telecom Environment as More Efficient and More Cost Effective Business Model 814
T. Žilić, V. Čošić

ICT APPLICATIONS

Usluga mobilnog plaćanja računa m:Pay 821
V. Žlof, S. Salapura

Praćenje imovinsko pravnih poslova na elektroničkoj komunikacijskoj infrastrukturi putem Web GIS aplikacije 826
D. Salopek, T. Đigaš, F. Ambroš, M. Štimac

Fieldbus Diagnostic Online Solution Program Establishment at Rijeka Oil Refinery 831
B. Žeželj, H. Hajdo

Automatic Communication System Ship to Shipping Terminal, for Reporting Potential Malfunctions of a Ballast Water Treatment System Operation 836
G. Bakalar, M. Baggini

AdriaHUB ICT platform 841
T. Škorjanc, R. Žigulić, N. Anđelić

Mjerenje kvaliteta usluge mobilnog plaćanja m:Pay 847
S. Salapura, V. Žlof

Uspostava sustava upravljanja identitetima u Carinskoj upravi 852
M. Hajnić, D. Cmuk

COMPUTERS IN EDUCATION

INVITED PAPER

New Informatics Curriculum - Croatian Tradition with World Trends 863
L. Kralj

PAPERS

Creativity, Communication and Collaboration: Grading with Open Badges 869
I. Salopek Čubrić, G. Čubrić

A Study of Students' Attitudes and Perceptions of Digital Scientific Information Landscape 875
R. Vrana

Researcher Measured - Towards a Measurement-driven Academia 881
H. Jaakkola, J. Henno, J. Mäkelä, K. Ahonen

Use of 'Learning Analytics' 888
J. Henno, H. Jaakkola, J. Mäkelä

Smart Immersive Education for Smart Cities with Support via Intelligent Pedagogical Agents 894
M. Soliman, A. Elsaadany

Review of Source-Code Plagiarism Detection in Academia 901
M. Novak

The Comparison of Impact Offline and Online Presentation on Student Achievements: A Case Study 907
P. Esztelecki, G. Kőrösi, Z. Námestovski, L. Major

Digital Competences for Teachers: Classroom Practice 912
M. Filipović Tretinjak, V. Anđelić

Introducing Collaborative e-Learning Activities to the e-Course "Information Systems" 917
M. Ašenbrener Katić, S. Čandrlić, M. Holenko Dlab

A Curriculum for Unified Embedded Engineering Education 923
I. Kaštelan, M. Temerinac

Individual versus Collaborative Learning in a Virtual World 929
P. Pürcher, M. Höfler, J. Pirker, L. Tomes, A. Ischebeck, C. Gütl

Preparation of a Hybrid e-Learning Course for Gamification 934
D. Kermek, D. Strmečki, M. Novak, M. Kaniški

Implementation of Fundamental Ideas into the Future Managers´ Informatics Education 940
L. Révészová

Fostering Creativity in Technology Enhanced Learning 946
A. Žižić, A. Granić, I. Šitin

Teaching Physics in Primary Schools with Tablet Computers: Key Advantages 952
V. Grubelnik, L. Grubelnik

Project Based Learning (PBL) in the Teachers' Education 957
M. Krašna

Didactical Suitability of e-Generated Drill Tests for Physics 962
R. Repnik, M. Sovič

Utilizing MOOCs in the Development of Education and Training Programs 966
P. Linna, T. Mäkinen, H. Keto

Distance Delivery and Technology-Enhanced Learning in Information Technology and Programming Courses at RIT Croatia 970
K. Marasović, B. Mihaljević, I. Bačić

Overview of IT Solutions for Career Services and Quality Assurance at Higher Education 976
E. Gjorgjevska, P. Tonkovikj, M. Gusev

Selecting the Most Appropriate Web IDE for Learning Programming Using AHP 982
I. Škorić, B. Pein, T. Orehovački

Using Real Projects as Motivators in Programming Education 988
M. Konecki, S. Lovrenčić, M. Kaniški

Making Programming Education More Accessible for Visually Impaired 992
M. Konecki, N. Ivković, M. Kaniški

Use of Computer Programs in Teaching Photography Courses at Schools of Applied Arts and Design in Croatia 996
Z. Prohaska, Z. Prohaska, I. Uroda

Universtiy Search Engine 1002
Ž. Knok, M. Marčec

Experience with Usage of LMS Moodle not Only for the Educational Purposes at the Educational Institution 1006
D. Paľová

Using Robot Simulation Applications at the University – Experiences with the KUKA Sim 1012
D. Lukac

Implementation and Analysis of Open Source Information Systems in Electronic Business Course for Economy Students 1017
H. Jerković, P. Vranešić, G. Slamić

Virtual Firms as Education Tool in the Field of eCommerce 1023
M. Vejačka

Systems and Software Assurance - A Model Cyber Security Course 1028
V. Jovanović, J. Harris

Analysis of Learning Management Systems Features and Future Development Challenges in Modern Cloud Environment 1033
H. Jerković, P. Vranešić, A. Radan

Markov Model of Mathematical Competences in Elementary Education 1039
G. Paić, B. Tepeš, K. Pavlina

PYTHON as Pseudo Language for Formal Language Theory 1045
Z. Dovedan Han, K. Kocijan, V. Lopina

Croatian Students' Attitudes Towards Technology Usage in Teaching Asian Languages – a Field Research 1051
M. Janjić, S. Librenjak, K. Kocijan

Adaptive e-Learning System for Language Learning: Architecture Overview 1056
V. Slavuj, B. Kovačić, I. Jugo

L2L – Learn to Learn: Teach to Learn: CARTOON ENGLISH (A constructivist approach to teaching and learning) 1061
K. Bedi

Facilitating Mobile Learning by Use of Open Access Information Resources 1067
R. Vrana

Work-Based Learning: New Skills for New Technologies 1072
M. Lamza Maronić, I. Ivančić

Creating Assets as a Part of Tertiary Education of Technical Domains 1078
J. Brozek, D. Hamernik, Z. Kopecky

Software Solution Incorporating the Steganographic Principle for Hiding Pictures within Pictures 1084
J. Brozek, J. Marek, V. Svoboda

A Platform Independent Tool for Programming, Visualization and Simulation of Simplified FPGAs 1091
M. Čupić, K. Brkić, Ž. Mihajlović

Digital Risks and Experiences of Future Teachers 1097
T. Bratina

A Study of Factors Influencing Higher Education Teachers' Intention to Use e-Learning in Hybrid Environments 1103
S. Babić, M. Čičin-Šain, G. Bubaš

Development and Implementation of E-Learning System in Smart Educational Environment 1109
A. Elsaadany, K. Abbas

Introducing Inquiry-Based Learning to Estonian Teachers: Experiences from the Creative Classroom Project 1115
N. Hoić-Božić, M. Laanpere, K. Pata, I. Franković, S. Teder

Mobile Robots Approach for Teaching Programming Skills in Schools 1121
W. Werth, C. Ungermanns

Age Independent Examination of Algorithm Creating Abilities 1125
Z.A. Godó, D. Kocsis, G. Kiss, G. Stóka

The Digitalization Push in Universities 1130
H. Jaakkola, H. Aramo-Immonen, J. Henno, J. Mäkelä

Toby the Explorer – an Interactive Educational Game for Primary School Pupils 1137
N. Kaevikj, A. Kostadinovska, B. Risteska Stojkoska, M. Mihova, K. Trivodaliev

The Use of Contemporary e-Services and e-Contents at Mother Tongue Classes 1142
V. Jesenek

Migration from in-House LMS to Google Classroom: Case of SEEU 1145
L. Abazi Bexheti, A. Kadriu, M. Apostolova Trpkovska

Survey Analyses of Impacting Factors in ICT Usage in School Management: Case Study 1149
B. Fetaji, M. Fetaji, R. Azemi, M. Ebibi

Case Study Analyses of Semantic Security Using SQL Injection in Web Enabled ORACLE Database 1155
M. Fetaji, B. Fetaji, M. Ebibi

Using Web Applications in Education 1161
A. Babić, S. Vukmirović, Z. Čapko

Qualitative Approach to Determining the Relevant Facets of Mobile Quality of Educational Social Web Applications 1165
T. Orehovački, S. Babić

Međukurikularni projekti u nastavi informatike u Ekonomskoj školi – primjeri dobre prakse 1171
S. Bulešić Milić

Tradicionalni ili hibridni model nastave računalstva 1174
M. Sertić, K. Šolić

Provjere znanja pomoću Classroom Managera u učionicama budućnosti 1180
M. Korać

Primjena e-učenja u hrvatskom vojnom obrazovanju 1184
D. Možnik

Prezentacijski alati za prikaz matematičkih sadržaja 1190
M. Štefan Trubić, I. Radošević

Siguran put do škole 1196
D. Šokac, I. Biuklija

Primjena obrazovne društvene mreže Edmodo u nastavi III. osnovne škole Čakovec 1199
N. Boj

Digitalni scenariji učenja 1203
M. Mirković

Detekcija najčešćih sintaktičkih i logičkih grešaka učenika kod stvaranja programa u početnim godinama učenja programiranja 1209
K. Blažeka

Nastava matematike na SageMathCloud platformi 1215
Ž. Tutek

Uvod u robotiku - Arduino platforma i web aplikacija 1218
A. Lacković, B. Fulanović

Informacijski sustav visokih učilišta - analiza slučaja za Veleučilište u Šibeniku 1222
S. Krajačić, L. Topolčić, F. Urem

Mobilne aplikacije u visokom obrazovanju 1225
M. Blašković, M. Fumić, F. Urem

Metodologija izrade *E – learning* sadržaja za edukaciju o izradi Standarda zanimanja 1230
I. Vunarić, S. Grgić, T. Babić

Uloga IKT u razvoju financijske pismenosti djece 1235
I. Ružić

Informacijsko-komunikacijske znanosti u nastavi - digitalizirani materijali za učenje 1239
T. Babić, A. Ogrin, M. Babić

Istraživanje stavova i očekivanja studenata prilikom upisa na studij kao metoda povećanja kvalitete usluge u visokom obrazovanju 1245
T. Babić, S. Grgić, E. Rajković

E-obrazovanjem do fleksibilnog modela učenja 1250
M. Božurić, R. Bogut, M. Tretinjak

Preporuke i primjeri dobre prakse e-učenja u hrvatskom visokom školstvu 1254
D. Junaković, I. Paćelat, F. Urem

Ilustracija primjene novog Kurikuluma iz predmeta Informatika i to domene - Računalno razmišljanje i programiranje na primjeru metode Početnica Mema za prvi razred osnovne škole 1258
M. Čičin-Šain, S. Babić, L. Kralj

Izloženost i navike korištenja medija i računala kod djece u razrednoj nastavi 1262
T. Pavičić, J. Šurić

COMPUTERS IN TECHNICAL SYSTEMS

INVITED PAPERS

Architecture and Application of Virtual Desk and 3D Process Simulation for Wire Rod Rolling Mills 1271
A. Venuti

Use of Offline Computational Tools for Plant Data Analysis and Setup Model Calibration: a Perspective in the Industry of Flat Metal Production 1276
C. Aurora, F.A. Cuzzola

Architecture and Implementation of a MES System in a Large Scale Steel Plant: Severstal Cherepovets Success Story 1280
G. Brunetti

PAPERS

Anfis as a Method for Determinating MPPT in the Photovoltaic System Simulated in Matlab/Simunlink 1289
D. Mlakić, S. Nikolovski

Linear Motion Calculation of the High Voltage Circuit Breaker Contacts Using Rotary Motion Measurement with Nonlinear Transfer Function 1294
K. Obarčanin, R. Ostojić

Robot Arm Teleoperation via RGBD Sensor Palm Tracking 1300
F. Marić, I. Jurin, I. Marković, Z. Kalafatić, I. Petrović

A Proposal for a Fully Distributed Flight Control System Design 1306
M. Šegvić, K. Krajček Nikolić, E. Ivanjko

Control of Thermal Process with Simulink and NI USB-6211 in Real Time 1311
I. Tikvić, G. Vujisić, M. Fruk

Stabilization of Multi-AUV Formation with Digital Control 1315
S.A. Ul'yanov, N.N. Maksimkin

A Hybrid Approach to Solve the Dynamic Patrol Routing Problem for Group of Underwater Robots 1321
M.Yu. Kenzin, I.V. Bychkov, N.N. Maksimkin

Multi - Heater Induction Heating System with Sandwich Material Heater 1327
A. Smrke

Two-Rate Motion Control of VTAV by NARMA-L2 Controller for Enhanced Situational Awareness 1333
I. Astrov

LADDER Program Solution for Multi-probe Monitoring and Control in Simple Cooling Process 1339
T. Špoljarić, M. Špoljarić

An M2M Solution for Smart Metering in Electrical Power Systems 1348
M.P. Shopov

Noise within a Data Center 1352
D. Miljković

Active Noise Control: From Analog to Digital – Last 80 Years 1358
D. Miljković

Responding to Stakeholders' Resistance to Change in Software Projects – A Literature Review 1364
S.L.R. Vrhovec

Object-Oriented Programming Model for Synthesis of Domain-Specific Application Development Environment 1369
T. Lugarić, Z. Pavlić, D. Škvorc

Logistic and Production Computer Systems in Small-Medium Enterprises 1375
M. Pighin

The Implications of Employing Component Based Software Design in Non-Commercial Applications 1380
B. Zorić, G. Martinović, I. Crnković

Extended Approach to Selecting a Project-specific Reliability Growth Model 1386
J. Krini, A. Krini, O. Krini, J. Börcsök

Embedded Linux Controlled Sensor Network 1392
M. Saari, A.M. Baharudin, P. Sillberg, P. Rantanen, J. Soini

Portable Sensor System for Reliable Condition Measurement 1397
J. Soini, P. Sillberg, P. Rantanen, J. Nummela

Architecture of an Interoperable IoT Platform Based on Microservices 1403
T. Vresk, I. Čavrak

Performance Estimation in Heterogeneous MPSoC Based on Elementary Operation Cost 1409
N. Frid, D. Ivošević, V. Sruk

Sustav za lociranje atmosferskih pražnjenja u identifikaciji kvarova TK mreže uzrokovanih atmosferskim prenaponima 1413
V. Milardić, B. Franc, M. Budimirović

SNUPI - Sustav za nadzor i upravljanje procesima infrastrukture podatkovnog centra 1420
M. Zmijanac

INTELLIGENT SYSTEMS

BiForD • SPECIAL SESSION ON BIOMETRICS & FORENSICS & DE-IDENTIFICATION AND PRIVACY PROTECTION

KEYNOTE SPEECH

Face Alignment: Addressing Pose Variability in Face Recognition Systems 1433
V. Štruc

PAPERS

Shape and Texture Combined Face Recognition for Detection of Forged ID Documents 1437
D. Sáez-Trigueros, H. Hertlein, L. Meng, M. Hartnett

Simple Method Based on Complexity for Authorship Detection of Text 1443
L. Meluch, I. Tokárová, P. Farkaš, F. Schindler

Privacy Protection Performance of De-identified Face Images with and without Background 1448
Z. Sun, L. Meng, A. Ariyaeeinia, X. Duan, Z.-H. Tan

Deep Metric Learning for Person Re-Identification and De-Identification 1454
I. Filković, Z. Kalafatić, T. Hrkać

Deformable Part-Based Robust Face Detection under Occlusion by Using Face Decomposition into Face Components 1459
D. Marčetić, S. Ribarić

Creating a Face Database for Age Estimation and Classification 1465
P. Grd, M. Bača

Forensic Anthropometry from Voice: An Articulatory-Phonetic Approach 1469
R. Singh, B. Raj, D. Gencaga

INTELLIGENT SYSTEMS

PAPERS

Computer Vision for the Blind: a Dataset for Experiments on Face Detection and Recognition 1479
S. Carrato, S. Marsi, E. Medvet, F.A. Pellegrino, G. Ramponi, M. Vittori

Impact of Light Conditions on the Vertical Traffic Signs Detection in Vertical Traffic Signs Recognition System 1485
D. Solus, Ľ. Ovseník, J. Turán

Wound Detection and Reconstruction Using RGB-D Camera 1490
D. Filko, E.K. Nyarko, R. Cupec

Clustering of Affective Dimensions in Pictures: An Exploratory Analysis of the NAPS Database 1496
M. Horvat, K. Jednoróg, A. Marchewka

Challenges in Adopting Big Data Strategies and Plans in Organizations 1502
A. Budin, S. Krajnović

A Survey of Intelligent System Techniques for Indian Stock Market Forecasting 1508
S. Panwar, V.P. Upadhyay, S.K. Bishnoi

The Effect of Class Distribution on Classification Algorithms in Credit Risk Assessment 1514
K. Andrić, D. Kalpić

Software Solution for Optimal Planning of Sales Persons Work Based on Depth-First Search and Breadth-First Search Algorithms 1521
E. Žunić, A. Djedović, B. Žunić

Iterated Local Search Algorithm for Planning the Sequence of Arrivals and Departures at Airport Runways 1527
E. Bytyçi, K. Sylejmani, A. Dika

Energy Efficiency with Intelligent Light Management Systems 1532
I. Britvić, A. Nikitović

Adaptive and Modular Urban Smart Infrastructure 1538
M. Klarić, I. Kuzle, I. Livaja

Automatic Pathole and Speed Breaker Detection Using Android System 1543
V. Rishiwal, H. Khan

The Influence of the CAPTCHA Types to Its Solving Times 1547
D. Brodić, S. Petrovska, M. Jevtić, Z.N. Milivojević

Techniques and Applications of Emotion Recognition in Speech 1551
S. Lugović, I. Dunđer, M. Horvat

Word Occurrences and Emotions in Social Media: Case Study on a Twitter Corpus 1557
I. Dunđer, M. Horvat, S. Lugović

The Application of Parameterized Algorithms for Solving SAT to the Study of Several Discrete Models of Collective Behavior 1561
S. Kochemazov, A. Semenov, O. Zaikin

Logical-Algebraic Equations Application in Discrete-Event Systems Studying 1566
N. Nagul

An Evaluation Framework and a Brief Survey of Decision Tree Tools 1572
N. Vlahović

Positive Constructed Formulas Preprocessing for Automatic Deduction 1578
E. Cherkashin, A. Davydov, A. Larionov

Monte-Carlo Randomized Algorithm: Empirical Analysis on Real-World Information Systems 1582
R. Kudelić, D. Oreški, M. Konecki

Control Flow Graph Visualization in Compiled Software Engineering 1586
A. Mikhailov, A. Hmelnov, E. Cherkashin, I.V. Bychkov

Bottom-Left and Sequence Pair for Solving Packing Problems 1591
T. Rolich, D. Domović, M. Golub

Automatic Image Annotation Refinement 1597
M. Pobar, M. Ivašić-Kos

Defining Ontology Combining Concepts of Massive Multi-Player Online Role Playing Games and Organization of Large-Scale Multi-Agent Systems 1603
B. Okreša Đurić, M. Schatten

Comparison of Solution Representations for Scheduling in the Unrelated Machines Environment 1609
M. Đurasević, D. Jakobović

INFORMATION SYSTEMS SECURITY

PAPERS

TECHNICAL TRACK

Technical Recommendations for Improving Security of Email Communications 1623
A. Malatras, I. Coisel, I. Sanchez

Performance Analysis of Two Open Source Intrusion Detection Systems 1629
B. Brumen, J. Legvart

Challenges of Mobile Device Use in Healthcare 1635
S.L.R. Vrhovec

Safe Use of Mobile Devices in the Cyberspace 1639
S.L.R. Vrhovec

Securing Web Content and Services in Open Source Content Management Systems 1644
H. Jerković, P. Vranešić, S. Dadić

Can Malware Analysts be Assisted in Their Work Using Techniques from Machine Learning? 1650
I. Novković, S. Groš

Performance Evaluation of a Rule-Based Access Control Framework 1656
S.A. Afonin

SOCIAL ENGINEERING TRACK

Going White Hat: Security Check by Hacking Employees Using Social Engineering Techniques 1663
Z. Lovrić Švehla, I. Sedinić, L. Pauk

Analysis of Phishing Attacks against Students 1667
J. Andrić, D. Oreški, T. Kišasondi

What Do Students Do with Their Assigned Default Passwords? 1674
L. Bošnjak, B. Brumen

Analysing Real Students' Passwords and Students' Passwords Characteristics Received From a Questionnaire 1680
V. Taneski, M. Heričko, B. Brumen

MISC TRACK

Using DEMF in Process of Collecting Volatile Digital Evidence 1689
M. Bača, J. Ćosić, P. Grd

From Safe Harbour to European Data Protection Reform 1694
T. Katulić, G. Vojković

Information Security Assessment in Nature Parks 1699
S. Aksentijević, T. Đugum, K. Šakić

Clustering Approach for User Location Data Privacy in Telecommunication Services 1706
M. Vuković, M. Kordić, D. Jevtić

Analiza sigurnosnih ranjivosti inteligentnih sučelja za upravljanje podatkovnim centrom 1711
M. Ramljak

BUSINESS INTELLIGENCE SYSTEMS

PAPERS

Analyzing Air Pollution on the Urban Environment 1723
E. Baralis, T. Cerquitelli, S. Chiusano, P. Garza, M.R. Kavoosifar

Application of Model Driven Architecture for Development of Data Consolidation Web-System 1729
A.A. Korobko, L.F. Nozhenkova

Business Process Management Systems Selection Guiedelines: Theory and Practice 1735
V. Bosilj Vukšić, L. Brkić, M. Baranović

Organization of Tax Data Warehouse for Legal Entities 1741
M. Sretenović, B. Kovačić, V. Jovanović

Predictive Analytics in Big Data Platforms – Comparison and Strategies 1747
M. Zekić-Sušac, A. Has

The Analysis of CSFs in Stages of ERP Implementation - Case Study in Small and Medium - Sized (SME) Companies in Croatia 1753
M. Nikitović, V. Strahonja

Model optimizacije procesa s primjenom na punjenju bankomata 1759
I. Osman, K. Bokulić

DIGITAL ECONOMY AND GOVERNMENT, LOCAL GOVERNMENT, PUBLIC SERVICES

PAPERS

The Modern Approach to the Analysis of Logistics Information Systems 1769
A. Iskra, E. Tijan, S. Aksentijević

Development of the Data Warehouse Model for Public Authorities Accounts in Croatia 1774
M. Sretenović, B. Kovačić, V. Jovanović

The Future of Digital Economy in Some SEE Countries (Case study: Croatia, Macedonia, Montenegro, Serbia, Bosnia and Herzegovina) 1780
M. Vidas-Bubanja, I. Bubanja

Effects and Evaluation of Open Government Data Initiative in Croatia 1786
T. Vračić, M. Varga, K. Ćurko

ICT Technologies and Structured Dialogue: Experience of "Go, go, NGO!" Project 1792
N. Kadoić

Using ICT Tools for Decision Making Support in Local Government Units 1798
N. Kadoić, I. Kedmenec

The Conceptual Risk Management Model - A Case Study of Varazdin County 1804
R. Kelemen, M. Biškup, N. Begičević Ređep

Electronic Commerce in Croatia and a Comparison of Open Source Tools for the Development of Electronic Commerce 1811
J. Tomljanović, T. Turina, E. Krelja Kurelović

The Social Marketing as Prerequisite for the Competitiveness of South-East European Companies 1817
I. Bubanja

Homeostasis and Collaborative Decision Making for Smart and Cognitive Cities 1822
J. Klasinc

Can the Bank Payment Obligation Replace the International Documentary Letter of Credit? 1828
R. Bergami

Implementation and Design of Cool'n'Project - Web-Based Project Management Software 1834
I. Špeh

Analysis of ICT Use in Private Accommodation Rentals in Croatia 1841
Lj. Zekanović-Korona, J. Grzunov

Records Management Challenges and Opportunities: An Australian Perspective 1847
A. Davies, R. Bergami

Effectiveness Analysis of Using Solid State Disk Technology 1852
A. Skendžić, B. Kovačić, E. Tijan

Information and Communication Technologies and the New Forms of Organized Crime in Network Society 1857
M. Boban

Digitalizacija lokalne uprave na primjeru Istarske županije 1863
L. Ordanić, N. Šarić-Kekić

Digitalna ekonomija – rezultanta disruptivnih tehnologija 1869
M. Mauher

MIPRO Junior – STUDENT PAPERS

PAPERS

Technical Diagnosis of Basic Logic Gates 1879
Z. Tucaković

Developing a Parking Monitoring System Based on the Analysis of Images from an Outdoor Surveillance Camera 1884
I.V. Sukhinskiy, E.A. Nepovinnykh, G.I. Radchenko

Laboratory Model of an Elevator: Control with Three Speed Profiles 1889
A. Jozić, T. Špoljarić, D. Gadže

Security and Privacy in an IT Context – a Low-Cost WIDS Employed against MITM Attacks (concept) 1895
N. Poljak, M. Ševo, I. Livaja

Use of HLA During Customer Flow Simulation in a Polyclinic 1899
J. Brozek, J. Fikejz, V. Samotan, L. Gago

Revealing the Structure of Domain Specific Tweets via Complex Networks Analysis 1904
E. Močibob, S. Martinčić-Ipšić, A. Meštrović

Counting Prime Numbers in Paralell - Faster by Reducing the Synchronization Overhead 1909
A. Duraković, E. Pajić, I. Branković, E. Kušundžija, S. Karkelja

Parallelization Challenges of BFS Traversal on Dense Graphs Using the CUDA Platform 1914
H. Milišić, D. Ahmić, H. Sinanović, E. Šarić, A. Asotić, A. Huseinović

Buck Converter Controlled by Arduino Uno 1919
H. Kovačević, Ž. Stojanović

Audio Phonebook for the Blind People 1924
G. Popović, U. Pale

Heart Rate Variability Analysis Using Different Wavelet Transformations 1930
U. Pale, F. Thürk, E. Kaniusas

Istraživanje *ransomware* napada i prijedlozi za bolju zaštitu 1936
M. Rak, M. Žagar

LIST OF PAPER REVIEWERS

Aksentijević, S.	(Croatia)
Alexin, Z.	(Hungary)
Antolić, Ž.	(Croatia)
Antonić, A.	(Croatia)
Aramo-Immonen, H.	(Finland)
Arbula, D.	(Croatia)
Ašenbrener Katić, M.	(Croatia)
Avbelj, V.	(Slovenia)
Babić, D.	(Croatia)
Babić, S.	(Croatia)
Bačmaga, J.	(Croatia)
Bako, N.	(Croatia)
Balaž, A.	(Serbia)
Banek, M.	(Croatia)
Banek Zorica, M.	(Croatia)
Barić, A.	(Croatia)
Basch, D.	(Croatia)
Bebel, B.	(Croatia)
Begušić, D.	(Croatia)
Bellatreche, L.	(France)
Bibuli, M.	(Italy)
Bilas, V.	(Croatia)
Blažević, D.	(Croatia)
Blažević, Z.	(Croatia)
Blečić, R.	(Croatia)
Bogunović, N.	(Croatia)
Bonastre, J.	(France)
Bosiljevac, M.	(Croatia)
Brčić, M.	(Croatia)
Bregar, K.	(Slovenia)
Brestovec, B.	(Croatia)
Brezany, P.	(Austria)
Britvić, I.	(Croatia)
Brkić, K.	(Croatia)
Brkić, L.	(Croatia)
Brkić, M.	(Croatia)
Broz, I.	(Croatia)
Budin, A.	(Croatia)
Budin, L.	(Croatia)
Bujan, I.	(Croatia)
Bujas, G.	(Croatia)
Buković, M.	(Croatia)
Butković, Ž.	(Croatia)
Car, Ž.	(Croatia)
Cifrek, M.	(Croatia)
Crnković Stumpf, B.	(Croatia)
Čačković, V.	(Croatia)
Čandrlić, S.	(Croatia)
Čeperić, V.	(Croatia)
Čičin-Šain, M.	(Croatia)
Čubrilo, M.	(Croatia)
Čupić, M.	(Croatia)
Davidović, M.	(Croatia)
Delač, G.	(Croatia)
Depolli, M.	(Slovenia)
Dešić, S.	(Croatia)
Dobrijević, O.	(Croatia)
Domazet-Lošo, M.	(Croatia)
Duarte, M.	(Portugal)
Džanko, M.	(Croatia)
Džapo, H.	(Croatia)
Đerek, V.	(Croatia)
Erceg, I.	(Croatia)
Eškinja, Z.	(Croatia)
Fertalj, K.	(Croatia)
Filjar, R.	(Croatia)
Fischer, D.	(Croatia)
Frid, N.	(Croatia)
Galinac Grbac, T.	(Croatia)
Gamulin, O.	(Croatia)
Garza, P.	(Italy)
Glavaš, G.	(Croatia)
Glavaš, J.	(Croatia)
Gojanović, D.	(Croatia)
Golfarelli, M.	(Italy)
Golub, M.	(Croatia)
Golubić, S.	(Croatia)
Gomez Chavez, A.	(Germany)
Gracin, D.	(Croatia)
Granić, A.	(Croatia)
Grd, P.	(Croatia)
Grgić, K.	(Croatia)
Grgurić, A.	(Sweden)
Groš, S.	(Croatia)
Grubišić, D.	(United States)
Gržinić, T.	(Croatia)
Gulić, M.	(Croatia)
Hadjina, T.	(Croatia)
Henno, J.	(Estonia)
Hoić-Božić, N.	(Croatia)
Holenko Dlab, M.	(Croatia)
Horvat, G.	(Croatia)
Horvat, M.	(Croatia)
Hrabar, S.	(Croatia)
Hrkać, T.	(Croatia)
Humski, L.	(Croatia)
Hure, N.	(Croatia)
Ilić, Ž.	(Croatia)
Inkret, R.	(Croatia)
Ipšić, I.	(Croatia)
Ivanjko, E.	(Croatia)
Ivašić-Kos, M.	(Croatia)
Ivković, N.	(Croatia)
Ivošević, D.	(Croatia)
Jaakkola, H.	(Finland)
Jakobović, D.	(Croatia)
Jakopović, Ž.	(Croatia)
Jakupović, A.	(Croatia)
Jardas, M.	(Croatia)
Jarm, T.	(Slovenia)
Jelenković, L.	(Croatia)
Jevtić, D.	(Croatia)
Ježić, G.	(Croatia)
Joler, M.	(Croatia)
Jovanovic, V.	(United States)
Jović, A.	(Croatia)
Kalafatić, Z.	(Croatia)
Kalpić, D.	(Croatia)

Kapus-Kolar, M. (Slovenia)
Karan, M. (Croatia)
Kaštelan, I. (Serbia)
Katanić, N. (Croatia)
Kaučič, B. (Slovenia)
Keto, H. (Croatia)
Kišasondi, T. (Croatia)
Klemenc-Ketiš, Z. (Croatia)
Kocev, D. (Slovenia)
Kocijan, K. (United States)
Kopčak, G. (Sweden)
Koričić, M. (Croatia)
Kosec, G. (Slovenia)
Kovačić, A. (Croatia)
Kovačić, B. (Croatia)
Krašna, M. (Slovenia)
Krhen, M. (Croatia)
Krivec, S. (Croatia)
Krois, I. (Croatia)
Krpić, Z. (Croatia)
Kudelić, R. (Croatia)
Kunda, I. (Croatia)
Kušek, M. (Croatia)
Lacković, I. (Croatia)
Lipovac, A. (Croatia)
Lo Presti , F. (Italy)
Lončarić, S. (Croatia)
Lovrenčić, A. (Croatia)
Lučev Vasić, Ž. (Croatia)
Lučić, D. (Croatia)
Lugarić, T. (Croatia)
Lukac, D. (Germany)
Ljubić, S. (Croatia)
Maček, M. (Slovenia)
Magdalenić, I. (Croatia)
Malarić, R. (Croatia)
Mandić, F. (Croatia)
Mandić, T. (Croatia)
Maračić, M. (Croatia)
Marčetić, D. (Croatia)
Marinović, I. (Croatia)
Marinović, M. (Croatia)
Marjanović, M. (Croatia)
Markuš, N. (Croatia)
Martinčić-Ipšić, S. (Croatia)
Matić, T. (Croatia)
Mauša, G. (Croatia)
Mekovec, R. (Croatia)
Mekterović, I. (Croatia)
Meng, L. (United Kingdom)
Mezak, J. (Croatia)
Mihajlović, Ž. (Croatia)
Mikac, B. (Croatia)
Mikuc, M. (Croatia)
Milanović, I. (Serbia)
Miličević, K. (Croatia)
Mišković, N. (Croatia)
Mlinarić, H. (Croatia)
Močinić, D. (Croatia)
Modlic, B. (Croatia)
Molnar, G. (Croatia)
Mošmondor, M. (Croatia)
Mrakovčić, T. (Croatia)
Mrković, B. (Croatia)
Nađ, Đ. (Croatia)
Nikitović, M. (Croatia)
Očko, M. (Croatia)
Oletić, D. (Croatia)
Orsag, M. (Croatia)
Pale, P. (Croatia)
Palestri, P. (Italy)
Paspallis, N. (United Kingdom)
Pavlić, Z. (Croatia)
Pečar-Ilić, J. (Croatia)
Pelin, D. (Croatia)
Perić Hadžić, A. (Croatia)
Perkovac, M. (Croatia)
Perković, T. (Croatia)
Petrović, G. (Croatia)
Pintar, D. (Croatia)
Pivac, B. (Croatia)
Pobar, M. (Croatia)
Pocta, P. (Slovakia)
Poljak, M. (Croatia)
Poplas Susič, T. (Slovenia)
Pribanić, T. (Croatia)
Pripužić, K. (Croatia)
Ptiček, M. (Croatia)
Rashkovska, A. (Slovenia)
Repnik, R. (Slovenia)
Resnik, D. (Slovenia)
Ribarić, S. (Croatia)
Rimac-Drlje, S. (Croatia)
Ristić, D. (Croatia)
Rupčić, S. (Croatia)
Seva, J. (United Kingdom)
Sillberg, P. (Finland)
Skala, K. (Croatia)
Skočir, P. (Slovenia)
Skorin-Kapov, L. (Croatia)
Soini, J. (Finland)
Soler, J. (Denmark)
Sorić, K. (Croatia)
Sruk, V. (Croatia)
Stanič, J. (Slovenia)
Stanič, U. (Slovenia)
Stapić, Z. (Croatia)
Stojković, N. (Croatia)
Stupar, I. (Croatia)
Sučić, S. (United States)
Suligoj, T. (Croatia)
Sužnjević, M. (Croatia)
Sviličić, B. (Croatia)
Šarolić, A. (Croatia)
Šegvić, S. (Croatia)
Ševrović, M. (Croatia)
Šikić, M. (Croatia)
Šilić, M. (Croatia)
Škvorc, D. (Croatia)
Štajduhar, I. (Croatia)
Štih, Ž. (Croatia)
Šunde, V. (Croatia)
Švedek, T. (Croatia)
Švogor, I. (Croatia)
Tanković, N. (Croatia)
Tijan, E. (Croatia)

Tomczak, J.	(Austria)	Vrdoljak, B.	(Croatia)
Tomić, M.	(Croatia)	Vrhovec, S.	(Slovenia)
Tralić, D.	(Croatia)	Vrlika, V.	(Croatia)
Trancoso, I.	(Portugal)	Vukadinović, D.	(Croatia)
Trobec, R.	(Slovenia)	Vuković, M.	(Croatia)
Tržec, K.	(Croatia)	Weber, M.	(Croatia)
Tuomi, P.	(Finland)	Werth, W.	(Austria)
Uroda, I.	(Croatia)	Zaluški, D.	(Croatia)
Varga, M.	(Croatia)	Zereik, E.	(Italy)
Vasić, D.	(Croatia)	Zinner, T.	(Germany)
Vidaček-Hainš, V.	(Croatia)	Žonja, S.	(Croatia)
Vladimir, K.	(Croatia)	Zulim, I.	(Croatia)
Vlahović, N.	(Croatia)	Žgank, A.	(Slovenia)
Vojković, G.	(Croatia)	Žilak, J.	(Croatia)
Vrančić, K.	(Croatia)	Živković, M.	(Croatia)
Vranić, M.	(Croatia)	Žulj, S.	(Croatia)

AUTHOR INDEX

Aalto, S. 515
Abazi Bexheti, L. 1145
Abbas, K. 1109
Abe, H. 257
Accardo, A. 719
Adamovic, N. 158
Afgan, E. 298
Afonin, S.A. 1656
Ahmić, D. 1914
Ahonen, K. 881
Ajčević, M. 719
Aksentijević, S. 1699, 1769
Alagić, D. 224, 230
Alekseev, Yu.V. 20
Alić, D. 745
Ambroš, F. 826
Amin, N. 656
Andrić, J. 1667
Andrić, K. 1514
Anđelić, N. 841
Anđelić, V. 912
Antonijević, M. 332
Apostolova Trpkovska, M. 1145
Aramo-Immonen, H. 1130
Arbanas, K. 224
Arbet, D. 57
Arbula, D. 253
Ariyaeeinia, A. 1448
Arsenovski, S. 388
Artemkina, S.B. 16
Asiama Nyako, W. 707
Asotić, A. 1914
Assefa, T.D. 673
Astrov, I. 1333
Ašenbrener Katić, M. 917
Aurora, C. 1276
Avbelj, V. 416
Azemi, R. 1149
Babič, M. 367
Babić, A. 1161
Babić, M. 1239
Babić, S. 1103
Babić, S. 1165, 1258
Babić, T. 1230, 1239, 1245
Bača, M. 1465, 1689
Bačić, I. 970
Bačmaga, J. 78
Badnjević, A. 395, 407, 477
Baggini, M. 836
Baharudin, A.M. 1392
Bakalar, G. 836
Balić, K. 376
Balki, M. 558
Balota, A. 370
Bánáti, A. 343
Bandiziol, A. 90
Banović-Ćurguz, N. 794, 800
Baralis, E. 1723
Baranović, M. 1735
Barić, A. 63, 78, 105, 110
Batbayar, K. 269
Bedi, K. 1061
Begičević Ređep, N. 1804
Belalem, G. 214
Belošić, D. 679
Benedetti, A. 5
Benedyczak, K. 190
Berdalović, I. 40
Berdinsky, A.S. 11, 16
Bergami, R. 1828, 1847
Berković, M. 703
Bianchi, F. 509
Bishnoi, S.K. 1508
Biškup, M. 1804
Biuklija, I. 1196
Blašković, M. 1225
Blažeka, K. 1209
Blečić, R. 105
Boban, M. 713, 1857
Bogut, R. 1250
Boj, N. 1199
Bokulić, K. 1759
Börcsök, J. 1386
Bosić, Z. 640
Bosilj Vukšić, V. 1735
Bošković, D. 521, 628
Bošnjak, L. 1674
Bothe, K. 766
Božurić, M. 1250
Brandonisio, F. 90
Branković, I. 1909
Bratina, T. 1097
Brenner, W. 158
Brezany, P. 202
Britvić, I. 1532
Brkić, K. 1091
Brkić, L. 1735
Brložnik, M. 416
Brodić, D. 1547
Brozek, J. 337, 1078, 1084, 1899
Brumen, B. 1629, 1674, 1680
Brunetti, G. 1280
Bubanja, I. 1780, 1817
Bubaš, G. 1103
Budimac, Z. 766
Budimirović, M. 1413
Budin, A. 1502
Budiša, D. 690
Bulešić Milić, S. 1171
Bulović, A. 472
Bychkov, I.V. 1321, 1586
Bytyçi, E. 1527
Bziuk, W. 684
Cakaj, S. 554
Capeska Bogatinoska, D. 707
Carlini, G. 719
Carrato, S. 1479
Celesti, A. 426

Cerquitelli, T. 1723
Cherkashin, E. 1578, 1586
Chiusano, S. 1723
Chiussi, S. 5
Chungurski, S. 388
Cifrek, M. 289, 326, 395, 407, 411, 446
Cmuk, D. 852
Coisel, I. 1623
Cota, P. 622
Craninckx, J. 72
Crnković, I. 1380
Cupec, R. 1490
Cuzzola, F.A. 1276
Čačković, V. 584
Čandrlić, S. 917
Čapko, Z. 1161
Čavrak, I. 1403
Čičin-Šain, M. 1103, 1258
Čižmar, R. 640
Čošić, V. 814
Čović, M. 30
Čubrić, G. 869
Čupić, M. 1091
Ćosić, J. 1689
Ćurko, K. 1786
D'Antrassi, P. 719
Dadić, M. 146, 543
Dadić, S. 1644
Daković, M. 569
Dancheva, T. 208
Date, S. 257
Davies, A. 1847
Davitashvili, T. 236
Davydov, A. 1578
Depolli, M. 401, 487
Dević, I. 640
Diduck, Q. 105
Dika, A. 1527
Dika, F. 185
Dizdarević, H. 599, 606
Dizdarević, S. 599, 606
Djebbar, E.I. 214
Djedović, A. 1521
Djinevski, L. 388
Domanjko-Petrič, A. 416
Domović, D. 1591
Dovedan Han, Z. 1045
Dražić Lutilsky, I. 809
Du, M. 289, 432, 446
Duan, X. 1448
Dudarin, A. 100
Dunđer, I. 1551, 1557
Duraković, A. 1909
Đambić, G. 442
Đigaš, T. 826
Đugum, T. 1699
Đurasević, M. 1609
Ebibi, M. 1149, 1155
Elsaadany, A. 894, 1109
Enne, R. 68
Erceg, I. 152
Esztelecki, P. 907
Farkaš, P. 1443
Fazio, M. 426
Fedorov, V.E. 11, 16
Feilhauer, T. 202
Feoktistov, A.G. 247
Fetaji, B. 1149, 1155
Fetaji, M. 1149, 1155
Fikejz, J. 1899
Filipović Tretinjak, M. 912
Filko, D. 1490
Filković, I. 1454
Finny, S. 126
Fischer, I.A. 5
Fosić, I. 690
Franc, B. 1413
Franković, I. 1115
Fratte-Sumper, S. 110
Frid, N. 1409
Friganović, K. 411
Fruk, M. 1311
Fulanović, B. 1218
Fumić, M. 1225
Gabela, A. 777
Gadže, D. 1889
Gago, L. 1899
Galin, M. 640
Galinac Grbac, T. 634
Gao, Y. 289, 432, 446
Garaguly, Z. 269
Garza, P. 1723
Gebrehiwot, M.E. 515
Gencaga, D. 1469
Geraskin, A.S. 95
Giedrimas, V. 745
Girtelschmid, S. 263
Gjorgjevska, E. 976
Glamočanin, D. 755
Godó, Z.A. 1125
Golub, I. 761
Golub, M. 1591
Goonasekera, N. 298
Gorachinova, L. 388
Gorshkova, A.S. 20
Gradišnik, V. 30
Granić, A. 946
Grd, P. 1465, 1689
Grgić, S. 1230, 1245
Grgurević, I. 736
Grilec, F. 446
Grollitsch, W. 90
Groš, S. 1650
Grozev, D.G. 314
Grubelnik, L. 952
Grubelnik, V. 952
Grubišić, D. 40
Grzunov, J. 1841
Gulan, G. 380
Gurbeta, L. 395, 407, 477
Gusev, M. 208, 218, 420, 976
Guseva, A. 218
Gütl, C. 929
Hadžiahmetović, N. 599, 606
Hajdo, H. 831
Hajnić, M. 852

Hamernik, D. 1078
Hamidović, D. 650
Han, M.K. 11
Harris, J. 1028
Hartnett, M. 1437
Has, A. 1747
Heđi, I. 805
Henno, J. 881, 888, 1130
Herceg, M. 86
Heričko, M. 1680
Hertlein, H. 1437
Hmelnov, A. 1586
Höfler, M. 929
Hoić-Božić, N. 1115
Holenko Dlab, M. 917
Hormot, F. 78
Horvat, G. 135
Horvat, M. 1496, 1551, 1557
Hrkać, T. 1454
Huseinović, A. 1914
I. Orović, 574
Ichikawa, K. 257
Iković, O. 521, 628
Iliev, L.A. 550, 562
Iliev, O. 388
Iliev, T.B. 550, 562
Ilišević, D. 794, 800
Ischebeck, A. 929
Iskra, A. 1769
Ivančić, I. 1072
Ivanova, E.P. 550, 562
Ivanović, M. 531, 766
Ivanjko, E. 1306
Ivašić-Kos, M. 1597
Ivić, M. 809
Ivković, N. 992
Ivošević, D. 1409
Jaakkola, H. 881, 888, 1130
Jakes, M. 337
Jakimoski, K. 388
Jakobović, D. 1609
Janjić, M. 1051
Jazbinšek, V. 467
Jednoróg, K. 1496
Jeričević, Ž. 30, 274, 380
Jerković, H. 1017, 1033, 1644
Jerman-Blažič, B. 367
Jesenek, V. 1142
Jevtić, D. 1706
Jevtić, M. 1547
Jiang, Y. 673
Job, J. 86
Jovalekic, S. 164
Jovanović, V. 1028, 1741, 1774
Jović, A. 326
Jozić, A. 1889
Jozić, K. 326
Jugo, I. 1056
Jukan, A. 684
Jukić, O. 805
Junaković, D. 1254
Jurin, I. 1300
Kacsuk, P. 343, 348
Kadoić, N. 1792, 1798
Kadriu, A. 1145
Kaevikj, N. 1137
Kail, E. 348
Kakanakov, N.R. 314
Kalafatić, Z. 1300, 1454
Kalpić, D. 1514
Kamcheva, E. 388
Kaniški, M. 934, 988, 992
Kaniusas, E. 1930
Karkelja, S. 1909
Kaštelan, I. 923
Kathiravelu, P. 634
Katulić, T. 1694
Kavoosifar, M.R. 1723
Kawai, E. 257
Kedmenec, I. 1798
Kelemen, R. 1804
Kenzin, M.Yu. 1321
Kermek, D. 934
Kertesz, A. 497
Keserica, H. 332
Keto, H. 966
Khan, H. 1543
Kido, Y. 257
Kim, S.J. 11
Kiss, G. 1125
Kišasondi, T. 1667
Klarić, M. 1538
Klasinc, J. 1822
Klobučarević, D. 679
Klobučarević, Ž. 679
Kljaić, Z. 656
Knežević, T. 34, 40
Knok, Ž. 1002
Kobal, D. 614
Kocev, D. 481
Kochemazov, S. 293, 1561
Kocijan, K. 1045, 1051
Kocsis, D. 1125
Kokanović, V. 442
Kolman, M. 308
Konecki, M. 988, 992, 1582
Kopecky, Z. 1078
Korać, M. 1180
Kordić, M. 1706
Koren, A. 141, 699
Koričić, M. 46, 51
Korobko, A.A. 1729
Kőrösi, G. 907
Kos, L. 303
Kosec, G. 308, 450
Kosovac, A. 703
Kostadinovska, A. 1137
Kovács, J. 348
Kovács, L. 269
Kováč, M. 57
Kovačević, H. 1919
Kovačević, Ž. 783
Kovačić, B. 1056, 1741, 1774, 1852
Kozlovszky, M. 269, 343, 348
Kožar, I. 274, 380
Krajačić, S. 1222

Krajček Nikolić, K. 1306
Krajnović, S. 1502
Kralevska, K. 673
Kralj, L. 863, 1258
Kranjac, M. 383
Krašna, M. 957
Krelja Kurelović, E. 1811
Krini, A. 1386
Krini, J. 1386
Krini, O. 1386
Križanović, K. 472
Krvavica, A. 416
Kučak, D. 442
Kudelić, R. 1582
Kukolja, D. 326
Kumar A., A. 120
Kumar Singh, S. 684
Kurbalija, V. 531
Kušundžija, E. 1909
Kutaladze, N. 236
Kuzle, I. 1538
Kuznetsov, V.A. 11, 16
Kvatadze, R. 236
Laanpere, M. 1115
Lacković, A. 1218
Lamza Maronić, M. 1072
Larionov, A. 1578
Lasorsa, I. 719
Lassila, P. 515
Lavrič, P. 401
Lazić, N. 783
Ledneva, A.Yu. 11, 16
Legvart, J. 1629
Lenac, K. 730
Librenjak, S. 1051
Likar, Š. 416
Lin, C. 432
Linna, P. 966
Livaja, I. 1538, 1895
Lo Presti, F. 509
Lonie, A. 298
Lopina, V. 1045
Lovrenčić, S. 988
Lovrić Švehla, Z. 1663
Lučev Vasić, Ž. 289, 446
Ludescher, T. 202
Lugarić, T. 1369
Lugović, S. 1551, 1557
Lukac, D. 1012
Lukovac, B. 141
M. Marić, 574
Maček, D. 230
Magerl, M. 110
Major, L. 907
Mäkelä, J. 881, 888, 1130
Mäkinen, T. 966
Maksimkin, N.N. 1315, 1321
Malarić, R. 146, 543
Malatras, A. 1623
Malekian, R. 707
Marasović, K. 970
Marceglia, S. 719
Marchewka, A. 1496
Marčec, M. 1002
Marčetić, D. 1459
Marek, J. 1084
Marić, F. 1300
Marinčić, A. 141, 667
Marinović, M. 472
Marjanović, D. 407, 395
Marković, I. 1300
Markulic, N. 72
Marsi, S. 1479
Martinčić, K. 82
Martinčić-Ipšić, S. 1904
Martinović, G. 1380
Martinović, Ž. 146, 543
Martinović, Ž. 146, 543
Matić, T. 86
Mauher, M. 1869
Medved Rogina, B. 353
Medved, M. 411
Medvet, E. 1479
Meluch, L. 1443
Meng, L. 1437, 1448
Meshchanov, V.P. 95
Meštrović, A. 1904
Mihajlović, Ž. 277, 1091
Mihaljević, B. 970
Mihaylov, Gr.Y. 550, 562
Mihova, M. 1137
Mijač, M. 772
Mikhailov, A. 1586
Mikuchadze, G. 236
Mikuličić, N. 277
Mikulić, J. 63
Milardić, V. 1413
Milišić, H. 1914
Milivojević, Z.N. 1547
Miljković, D. 1352, 1358
Mirković, M. 1203
Miškić-Pletenac, N. 730
Mišlov, R. 110
Mitrović, N. 68
Mlakić, D. 1289
Mochizuki, H. 46, 51
Močibob, E. 1904
Modebadze, Z. 236
Mohorčič, M. 487
Molnar, G. 100
Morita, S. 46, 51
Mostarac, P. 146
Možnik, D. 1184
Nagul, N. 1566
Nagy, L. 57
Nakrap, I.A. 95
Námestovski, Z. 907
Nardelli, M. 503
Nenadić, I. 614
Nepovinnykh, E.A. 196, 1884
Ni, W. 289, 446
Nikitović, A. 1532
Nikitović, M. 1753
Nikolovski, S. 1289
Nonis, R. 90
Novak, M. 901, 934

Novković, I. 1650
Nozhenkova, L.F. 1729
Nummela, J. 1397
Nyarko, E.K. 1490
Obarčanin, K. 1294
Ogrin, A. 1239
Okanović, V. 376
Okreša Đurić, B. 1603
Oliveira, F. 5
Omanović, S. 745
Opiła, J. 361
Ordanić, L. 1863
Orehovački, T. 982, 1165
Oreščanin, D. 590
Oreški, D. 1582, 1667
Orović, I. 569
Oršolić, P. 116
Osman, I. 1759
Osrečki, Ž. 40
Ostojić, R. 1294
Ovseník, Ľ. 1485
Paćelat, I. 1254
Paić, G. 1039
Pajić, E. 1909
Palata, D. 614
Pale, U. 1924, 1930
Palestri, P. 90
Paľová, D. 1006
Panwar, S. 1508
Paroški, M. 383
Pata, K. 1115
Pauk, L. 1663
Pauković, G. 640
Pavičić, T. 1262
Pavić, T. 791
Pavlić, Z. 1369
Pavlina, K. 1039
Pein, B. 982
Pellegrino, F.A. 1479
Petrova-El Sayed, M. 190
Petrović, I. 1300
Petrovska, S. 1547
Pflanzner, T. 497
Pighin, M. 1375
Pirker, J. 929
Pobar, M. 1597
Poljak, N. 1895
Poplas Susič, T. 456
Popović, B. 370
Popović, G. 1924
Popović, Ž. 584
Posavec, N. 24
Prazina, I. 376
Prohaska, Z. 996
Prohaska, Z. 996
Pröll, B. 263
Pršeš, K. 376
Pun, S.H. 289, 432
Pürcher, P. 929
Putnik, Z. 766
R, R. 120, 126, 130
Raczkowski, K. 72
Radan, A. 1033
Radchenko, G.I. 196, 1884
Radojević, B. 761
Radošević, I. 1190
Raj R.S., R. 130
Raj, B. 1469
Rajković, E. 1245
Rak, M. 1936
Ramljak, D. 751
Ramljak, M. 1711
Ramponi, G. 1479
Rantanen, P. 1392, 1397
Rashkovska, A. 481
Rembold, D. 164
Repnik, R. 962
Retschitzegger, W. 263
Révészová, L. 940
Ribarić, S. 1459
Rishiwal, V. 1543
Risteska Stojkoska, B. 724, 1137
Ristov, S. 208, 218, 420
Ristovski, A. 218
Rizvić, S. 376
Rolich, T. 1591
Romanenko, A.I. 11, 16
Romano, A. 426
Rostamzadeh Hajilari, H. 283
Rumyantseva, V.D. 20
Rutkowski, A. 190
Ružić, I. 1235
Saari, M. 1392
Sáez-Trigueros, D. 1437
Salapura, S. 821, 847
Salfinger, A. 263
Salopek Čubrić, I. 869
Salopek, D. 826
Samkharadze, I. 236
Samotan, V. 1899
Sanchez, I. 1623
Savin, A.N. 95
Schatten, M. 1603
Schatzberger, G. 63
Schindler, F. 1443
Schuller, B. 190
Schulze, J. 5
Schwinger, W. 263
Sedinić, I. 1663
Semenov, A. 293, 1561
Sertić, M. 1174
Sharifi, M. 283
Shilov, I.P. 20
Shimojo, S. 257
Shopov, M.P. 314, 1348
Sidorov, I.A. 242, 247
Sikimić, U. 383
Sillberg, P. 1392, 1397
Simić, I. 383
Sinanović, H. 1914
Singh, R. 1469
Sirotić, I. 152
Skala, K. 175, 298
Skendžić, A. 1852
Slak, J. 450
Slamić, G. 1017

Slavuj, V. 1056
Smrke, A. 1327
Soini, J. 1392, 1397
Soliman, M. 894
Solus, D. 1485
Sovič, M. 962
Sretenović, M. 1741, 1774
Sruk, V. 1409
Stančić, A. 736
Stanič, U. 456
Stanković, I. 569
Stanković, S. 574, 569
Stapić, Z. 772
Stockreiter, C. 110
Stojanović, A. 783
Stojanović, Ž. 1919
Stóka, G. 1125
Stopjaková, V. 57
Stoyanov, I.S. 550, 562
Strahonja, V. 772, 1753
Strmečki, D. 934
Strujić, Dž. 370
Sučić, S. 332
Sukhinskiy, I.V. 1884
Suligoj, T. 34, 40, 46, 51
Suljanović, N. 650
Sun, Z. 1448
Sušanj, D. 253, 380
Svoboda, V. 337, 1084
Sylejmani, K. 1527
Šabec, J. 622
Šakić, K. 1699
Šarić, E. 1914
Šarić-Kekić, N. 1863
Šegmanović, F. 40
Šegvić, M. 1306
Šerifović-Trbalić, A. 462
Ševo, M. 1895
Šikić, M. 472
Šimunić, D. 141, 667, 699
Šipuš, D. 579
Šitin, I. 946
Škoda, P. 353
Škorić, I. 982
Škorjanc, T. 841
Škorput, P. 656
Škrbić, M. 599, 606
Škvorc, D. 1369
Šneler, L. 86
Šojat, Z. 175
Šokac, D. 1196
Šolić, K. 1174
Šoša Anić, M. 791
Šovčík, M. 57
Špeh, I. 1834
Špoljarić, M. 1339
Špoljarić, T. 1339, 1889
Štajcer, M. 590
Štajcer, M. 590
Štefan Trubić, M. 1190
Štimac, M. 826
Štokić, N. 644
Štruc, V. 1433
Šurić, J. 1262
Talebi, M.M. 283
Tan, Z.-H. 1448
Taneski, V. 1680
Tasič, J. 420
Taşpinar, N. 558
Taylor, J. 298
Teder, S. 1115
Telenta, M. 303
Temerinac, M. 923
Tepeš, B. 1039
Thürk, F. 1930
Tijan, E. 1769, 1852
Tikvić, I. 1311
Tokárová, I. 1443
Tomes, L. 929
Tomić, S. 383
Tomljanović, J. 1811
Tonkovikj, P. 976
Topolčić, L. 1222
Tošić, M. 521
Tošić, M. 628
Trbalić, A. 462
Trengoska, J. 707
Tretinjak, M. 1250
Trivodaliev, K. 1137
Trobec, R. 481
Tucaković, Z. 1879
Tudić, V. 24
Turán, J. 1485
Turina, T. 1811
Tusun, S. 152
Tutek, Ž. 1215
Ul'yanov, S.A. 1315
Ungermanns, C. 1121
Upadhyay, V.P. 1508
Urem, F. 1222, 1225, 1254
Uroda, I. 996
Vai, M.I. 289, 432
Varga, M. 1786
Vaser, R. 472
Veiga, L. 634
Veispahić, A. 703
Vejačka, M. 1023
Venuti, A. 1271
Vidas-Bubanja, M. 1780
Villari, M. 426
Vinko, D. 116, 135
Vittori, M. 1479
Vlahović, N. 1572
Vlaović, J. 135
Vojković, G. 1694
Vračić, T. 1786
Vrana, R. 875, 1067
Vranešić, P. 1017, 1033, 1644
Vrbovčan, I. 791
Vresk, T. 1403
Vrhovec, S.L.R. 436, 1364, 1635, 1639
Vučić, M. 100
Vujisić, G. 1311
Vukmirović, S. 1161
Vuković, M. 1706
Vunarić, I. 1230

Vyroubal, V. 736
Wambacq, P. 72
Watashiba, Y. 257
Weber, M. 713
Weiss, B. 110
Werth, W. 1121
Xhafa, V. 185
Yakovleva, G.E. 11
Yamanaka, H. 257
Zaikin, O. 293, 1561
Zdravevski, V. 208
Zekanović-Korona, Lj. 1841
Zekić-Sušac, M. 1747
Zimmermann, H. 68
Zinterhof, P. 320
Zmijanac, M. 1420
Zorić, B. 1380
Žagar, M. 1936
Žeželj, B. 831
Žigulić, R. 841
Žilak, J. 46, 51
Žilić, T. 814
Žitnik, T. 640
Žižić, A. 946
Žlof, V. 821, 847
Žunić, B. 1521
Žunić, E. 1521

FOREWORD

The *39th International ICT Convention MIPRO 2016* was held from 30th of May until 3rd of June 2016 in Opatija, the Adriatic Coast, Croatia. The Convention consisted of nine conferences under the titles: *Microelectronics, Electronics and Electronic Technology (MEET), Distributed Computing, Visualization and Biomedical Engineering (DC VIS), Telecommunications & Information (CTI), Computers in Education (CE), Computers in Technical Systems (CTS), Intelligent Systems (CIS), Information Systems Security (ISS), Business Intelligence Systems (miproBIS), Digital Economy and Government, Local Government, Public Services (DE/GLGPS)*. A special conference was dedicated to the works of students: *MIPRO Junior-Student Papers (SP)*. Along with this, special sessions on *Biometrics & Forensics & De-Identification and Privacy Protection (BiForD)* and *Future Networks and Services (FNS)* were also held as a part of convention MIPRO.

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

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

To all who helped organizing the 39th International ICT Convention *MIPRO 2016* as well as editing of this Book of Proceedings we extend our heartfelt thanks.

Prof. Petar Biljanović, PhD
International Program Committee
General Chair

MEET

International Conference on

MICROELECTRONICS, ELECTRONICS AND ELECTRONIC TECHNOLOGY

Steering Committee

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

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

INVITED PAPER

(Si)GeSn Nanostructures for Optoelectronic Device Applications

I. A. Fischer*, F. Oliveira*, **, A. Benedetti§, S. Chiussi §§ and J. Schulze*
* Institute of Semiconductor Engineering, Pfaffenwaldring 47, 70569 Stuttgart, Germany
** Centre of Physics, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
§ CACTI, Univ. de Vigo, Campus Universitario Lagoas Marcosende 15, Vigo, Spain
§§ Dpto. Fisica Aplicada, Univ. de Vigo, Rua Maxwell s/n, Campus Universitario Lagoas Marcosende, Vigo, Spain
fischer@iht.uni-stuttgart.de

Abstract – We present an overview of recent results on the fabrication of GeSn- and SiGeSn-nanostructures for optoelectronic device application.

I. INTRODUCTION

Recent years have seen numerous experimental efforts directed at integrating photonics with electronics based on Group-IV optoelectronic devices. The use of Si and Ge in optoelectronics has been limited by their indirect bandgap resulting in low efficiency. More recently, significant progress has been made in enhancing optical properties of Group-IV alloys by including Sn. The unstrained binary alloy $Ge_{1-y}Sn_y$ is predicted to become a direct bandgap semiconductor for $y > 0.073$ [1], while $y > 0.17$ is needed to obtain a direct bandgap material for $Ge_{1-y}Sn_y$ grown pseudomorphically on Ge [2]. The growth of $Ge_{1-y}Sn_y$ on Ge is challenging because of the 14.7 % lattice mismatch between α-Sn (with lattice constant a_{Sn} = 6,493 Å) and Ge (a_{Ge} = 5,658 Å). The existence of a direct bandgap material has been confirmed for a partially relaxed $Ge_{0.874}Sn_{0.126}$ layer on Ge [3].

Electrical and optical properties can be tuned further by adding Si (a_{Si} = 5,431 Å) i. e. investigating the ternary alloy. $Si_xGe_{1-x-y}Sn_y$ can be grown without lattice mismatch on Ge for

$$x = \frac{a_{Sn} - a_{Ge}}{a_{Ge} - a_{Si}} y \,. \qquad (1)$$

Compared to $Ge_{1-y}Sn_y$ the material properties such as the composition-dependent bandgap of the ternary alloy $Si_xGe_{1-x-y}Sn_y$ are much less understood and subject of ongoing experimental efforts.

Nanostructures such as quantum wells and islands are well suited for application in optoelectronic devices because of the improved optical properties that originate from carrier confinement in one or more directions. The properties of nanostructures such as quantum wells and quantum dots for application in III-V semiconductor-based optical devices have led to the fabrication and commercialization of devices such as quantum-well and quantum-dot lasers, quantum cascade lasers and quantum-well infrared photodetectors. SiGe nanostructures such as quantum wells [4] and islands [5] have also been intensively investigated for applications in modulators or photodetectors. The addition of Sn opens up exciting new possibilities for the use of Group-IV-nanostructures in optical device applications. There are a number of theoretical proposals for lasers and infrared photodetectors based on (Si)GeSn multi-quantum-well structures [6]–[8]. While nanostructures are interesting for optoelectronic applications in their own right, fabricating and functionalizing nanostructures in the context of group-IV-optoelectronics could be a possible route towards obtaining high Sn content in such structures, with the concomitant advantages for optical efficiency.

Group-IV nanostructures containing Sn have previously been fabricated using different techniques. Sn dots were grown on Ge [9], $Ge_{1-y}Sn_y$ dots were grown on thin SiO_2 layers on top of Si (111) substrates [10] and Sn nanostructures were embedded into a Si or Ge matrix by annealing of SiSn or GeSn films [11]. Here, we present a brief overview of our recent progress in fabricating (Si)GeSn-nanostructures that can be fully embedded in Group-IV-based devices.

II. NANOSTRUCTURE FABRICATION

A. Molecular Beam Epitaxy

Molecular Beam Epitaxy (MBE) is a growth method that is ideally suited to material deposition with monolayer precision. The growth of GeSn and SiGeSn alloys is usually performed at very low substrate temperatures in order to prevent precipitation and segregation of Sn. Deposition temperatures of 100 °C – 160 °C are often selected for GeSn, while SiGeSn has also been grown at higher temperatures [12]. In bulk materials, the substrate temperature has a strong impact on layer quality and, thus, on device quality. For nanostructures, material diffusion and segregation is an additional concern that needs to be addressed when selecting growth temperatures [12]. While a higher growth temperature could also yield better layer quality, it can be expected to negatively affect the apruptness of heterotransitions.

All structures discussed here were grown using solid source MBE at a base pressure lower than 10^{-10} mbar. An electron beam evaporator was used for Si evaporation,

This work was partly supported by the Portuguese Foundation for Science and Technology (FCT) through Strategic Project PEst-C/FIS/UI0607/2013 and PhD Fellowship (F. Oliveira).

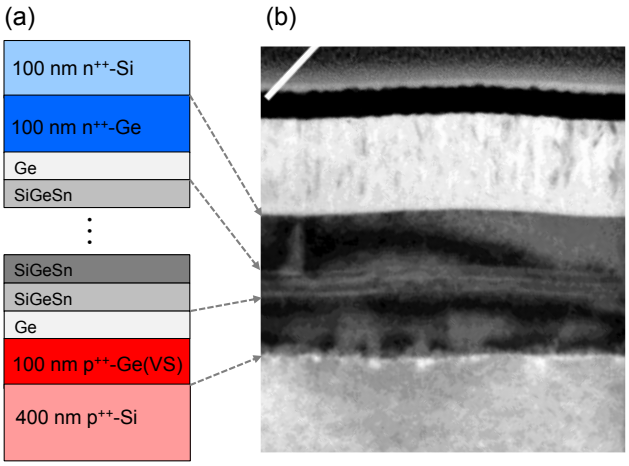


Figure 1. (a) Schematic layer structure and (b) TEM image of SiGeSn multi-quantum-well structures. The contrast indicates material transitions.

while Knu
The Si as
quadruple
flux stabil
temperature. The Si flux was calibrated by growing Si films on Si (100) wafers with a growth rate of 1 Å/s and measuring film thickness with a profilometer. The Ge flux was calibrated by growing a relaxed epitaxial Ge film on a Si (100) wafer with a slow flux of ≈ 0.1 Å/s and measuring film thickness by ellipsometric spectroscopy. The flux of Sn was calibrated by growing thin epitaxial films of $Ge_{1-x}Sn_x$ on Ge buffer layers on Si (100) wafers. The absolute concentration of Sn was subsequently measured using Rutherford backscattering spectroscopy.

All samples discussed in the following subsections were fabricated on 4" Si (100) wafers. After placing the wafers into the MBE chamber they were first subjected to a thermal desorption step at 900 ºC for 5 minutes to remove the surface SiO_2 layer. This was followed by the growth of 50 nm of Si to cover remaining surface contaminants and obtain a smooth surface for the subsequent growth steps. For (Si)GeSn layers grown on Ge it was necessary to form a Ge virtual substrate (VS) on the Si wafer to accommodate the lattice difference between Si and Ge and enable the growth of high quality layers on Ge. This virtual substrate was formed in two steps. A Ge buffer layer was grown at 1 Å/s and 330 ºC, ng treatment at 850 ºC to reduce the lensity and form a virtual substrate. itaxial Ge was then deposited to ace onto which high-quality layers

ntum Well Structures

Sn multi-quantum-well (MQW) the advantage that while both well be strained with respect to Ge, the r and well layers can be chosen in strain of the MQW structure on Ge o. Such structures can, thus, make l Ge VS technology for integration re, investigated SiGeSn MQW two and four wells grown on Ge strate temperature of 160 °C [13]. of the well and barrier layers can clearly be images (Fig. 1 (b)). Both barrier and well 10 nm thick, are composed of $Si_xGe_{1-x-y}Sn_y$ t fractions of the semiconductor materials $_{0.07}$ for the barrier and $Si_{0.25}Ge_{0.63}Sn_{0.12}$ for the well layers). The material compositions were chosen for their bandgaps and such that the lattice constant of the unstrained well (barrier) ternary alloy is larger (smaller) than that of Ge in order to obtain a MQW layer stack with small residual strain on the Ge VS.

Optoelectronic device functionality has been demonstrated for these structures by placing them in the intrinsic layer of a PIN-photodiode layer stack. The presence of the SiGeSn-MQW structure can be seen to influence both photocurrent and electroluminescence measurements [13].

C. *Sn-rich GeSn Multi-Quantum-Well-Structures in Ge*

Similarly to the growth of SiGeSn MQW structures, the fabrication of Ge/GeSn MQW structures can be achieved by sandwiching GeSn layers with thicknesses of a few nanometers between Ge spacer layers. Light-emitting diodes (LEDs) containing $Ge/Ge_{0.93}Sn_{0.07}$ MQW layers with varying thicknesses have been shown to emit light with much higher intensity than a reference Ge LED

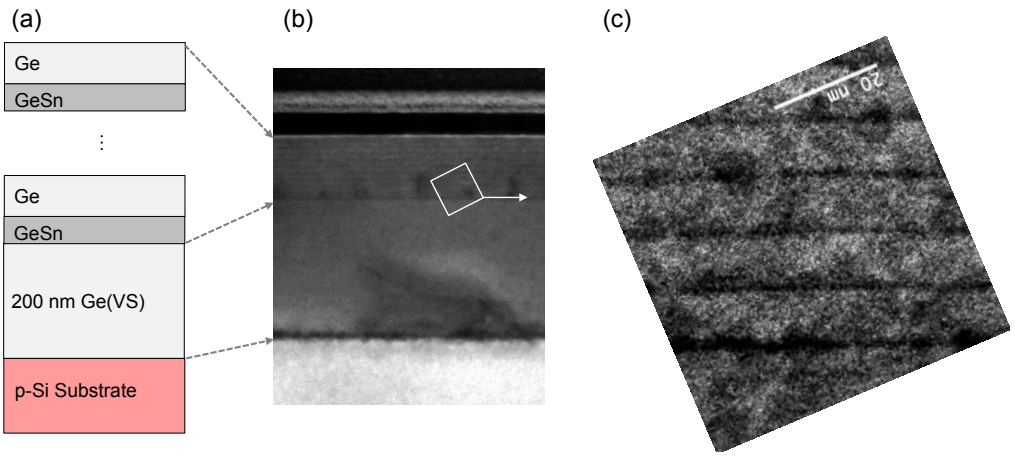


Figure 2. (a) Schematic layer structure and (b) TEM BF as well as (c) HR-TEM images of GeSn multi-quantum-well structures. The contrast indicates material transitions.

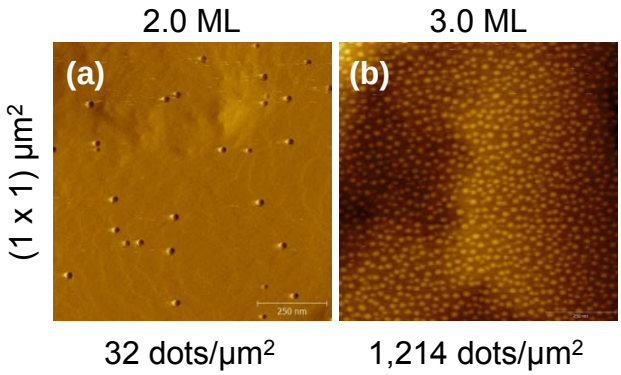


Figure 3. AFM measurements of (a) 2.0 and (b) 3.0 ML of Sn deposited on Ge (100). Large-scale dot formation is observed for 3.0 ML of Sn.

[14].

A particularly high Sn content in Ge/GeSn MQW structures can be achieved by depositing pure Sn layers on Ge with a total layer thickness that is below the critical layer thickness t_c at which the transition from 2D to 3D growth (Stranski-Krastanov growth) sets in. Similarly to the growth of Ge on Si, which has been investigated for nearly three decades, the growth of Sn on Ge proceeds in two stages. In the first stage, Sn forms a fully strained 2D-wetting layer on Ge. At a critical layer thickness, relaxation sets in via formation of local, Sn-rich material accumulations: 3D island growth is obtained. Using Sn layers with thicknesses below t_c and overgrowing them with 5 – 10 nm of Ge, ultra-thin multi-quantum well layers with high Sn-content can be fabricated [15].

Fig. 2 shows cross sectional TEM images of a sample with 10 Sn-rich wells separated by 10 nm Ge spacer layers. The sample was fabricated by repeatedly depositing 2 ML of Sn and overgrowing them with 10 nm Ge at a constant substrate temperature of 100 °C. The resulting position-dependent Sn composition is likely to be influenced by material diffusion and Sn segregation but can be expected to be high.

D. Sn-rich GeSn islands on Ge

When the critical thickness is exceeded during the deposition of pure Sn on Ge, we can observe 3D island growth. We investigated the transition from 2D to 3D growth of pure Sn on Ge by depositing 0 – 5 ML of Sn on Ge at a substrate temperature of 100 °C [15]. AFM images of two selected samples are shown in Figure 3. At these growth temperatures, we determined the critical thickness to be t_c = 2.25 ML. For Sn layers with thicknesses larger than t_c we observe the onset of large-scale dot formation as shown in Fig. 3 (b). Only few dots can be found for Sn layers with thicknesses below t_c as shown in Fig. 3 (a). No facets can be observed in AFM measurements. An interesting question is whether the Sn islands contain α-Sn, whose crystal structure is identical to that of Si and Ge, or β-Sn, which is the thermodynamically stable phase for temperatures above 13 °C. The fact that such layers overgrown with Ge show perfect crystallinity in TEM measurements seems to indicate that the islands observed in AFM measurements are indeed composed of α-Sn [15].

E. GeSn islands in Si

Finally, a strategy of producing GeSn islands on Si that closely mirrors the growth of self-assembled Ge islands on Si consists of growing few MLs of $Ge_{1-y}Sn_y$ on Si. Self-assembled GeSn islands were fabricated by depositing 5.5 ML of $Ge_{0.96}Sn_{0.04}$ on Si at a constant substrate temperature of 350 °C and overgrowing them with 10 nm of Si [16]. This growth sequence could be repeated up to four times until layer quality started to deteriorate [16]. TEM images of a sample with three stacked layers of self-assembled GeSn islands are shown in Fig. 4. At a thickness of 5.5 ML the deposited $Ge_{0.96}Sn_{0.04}$ layers exceed the critical thickness for island formation and relaxation sets in via the local accumulation of GeSn. The resulting structures are clearly visible in the cross sectional TEM images. Local material accumulation can be seen to be accompanied by a thinning of the wetting layer in the vicinity of the islands. The resulting island composition can, thus, be expected to be a result of position-dependent intermixing as is the case with Ge layers deposited on Si. An experimental technique with sub-nanometre precision would be required to investigate the position-dependent composition of the resulting islands in detail.

III. CONLUSION

We have explored several growth strategies for

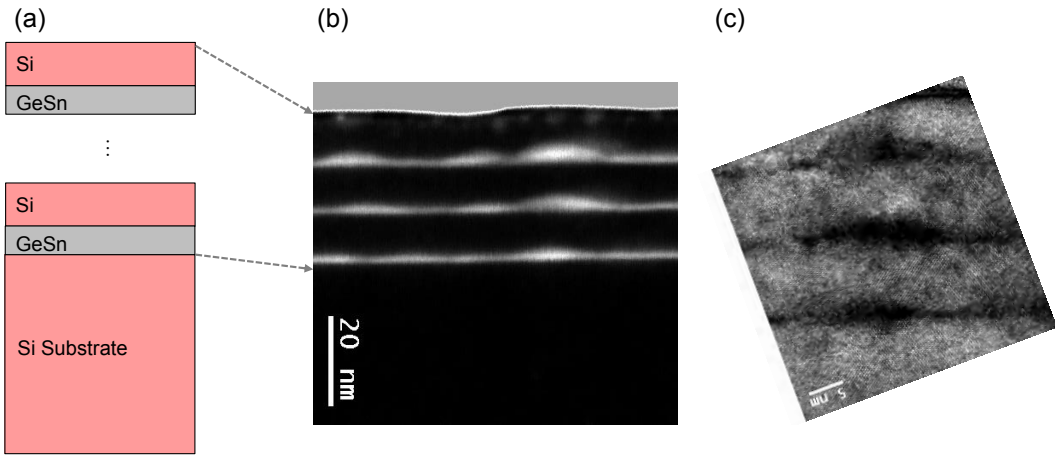


Figure 4. (a) Schematic layer structure and (b) STEM as well as (c) HR-TEM images of GeSn dots grown on Si and capped with Si. Local GeSn-rich islands form as a result of the lattice mismatch between $Ge_{0.96}Sn_{0.04}$ and Si.

growing Sn-rich quantum wells and islands that can be fully embedded either in Si or Ge for future possible applications in electrooptic devices. We are able to fabricate Sn-rich nanostructures, some of which have already been integrated into diodes to demonstrate optoelectronic functionality. Future steps consist of improving layer growth and, most importantly, explore experimental strategies to determine composition and strain on a nanometer and even sub-nanometer scale in order to be able to tailor those nanostructures for application needs.

ACKNOWLEDGMENT

We thank G. Capellini, M. Virgilio, T. Wendav and K. Busch for insightful discussions.

REFERENCES

[1] L. Jiang, J. D. Gallagher, C. L. Senaratne, T. Aoki, J. Mathews, J. Kouvetakis, and J. Menéndez, "Compositional dependence of the direct and indirect band gaps in $Ge_{1-y}Sn_y$ alloys from room temperature photoluminescence," Semicond. Sci. Technol., vol. 29, p. 115028, 2014.

[2] A. A. Tonkikh, C. Eisenschmidt, V. G. Talalaev, N. D. Zakharov, J. Schilling, G. Schmidt, and P. Werner, "Pseudomorphic GeSn/Ge(001) quantum wells: Examining indirect band gap bowing," Appl. Phys. Lett., vol. 103, p. 32106, Jul. 2013.

[3] S. Wirths, R. Geiger, N. von den Driesch, G. Mussler, T. Stoica, S. Mantl, Z. Ikonic, M. Luysberg, S. Chiussi, J. M. Hartmann, H. Sigg, J. Faist, D. Buca, and D. Grützmacher, "Lasing in direct-bandgap GeSn alloy grown on Si," Nat. Photonics, vol. 9, pp. 88–92, Feb. 2015.

[4] Y.-H. Kuo, Y. K. Lee, Y. Ge, S. Ren, J. E. Roth, T. I. Kamins, D. A. B. Miller, and J. S. Harris, "Quantum-Confined Stark Effect in Ge/SiGe Quantum Wells on Si for Optical Modulators," IEEE J. Sel. Top. Quantum Electron., vol. 12, pp. 1503–1513, Dec. 2006.

[5] K. L. Wang, Dongho Cha, Jianlin Liu, and C. Chen, "Ge/Si Self-Assembled Quantum Dots and Their Optoelectronic Device Applications," Proc. IEEE, vol. 95, pp. 1866–1883, Sep. 2007.

[6] G.-E. Chang, S.-W. Chang, and S.-L. Chuang, "Strain-Balanced Multiple-Quantum-Well Lasers," IEEE J. Quantum Electron., vol. 46, pp. 1813–1820, 2010.

[7] G. Sun, R. A. Soref, and H. H. Cheng, "Design of a Si-based lattice-matched room-temperature GeSn/GeSiSn multi-quantum-well mid-infrared laser diode," Opt. Express, vol. 18, pp. 19957–19965, Sep. 2010.

[8] G. Sun, R. A. Soref, and H. H. Cheng, "Design of an electrically pumped SiGeSn/GeSn/SiGeSn double-heterostructure midinfrared laser," J. Appl. Phys., vol. 108, p. 33107, Aug. 2010.

[9] W. Dondl, P. Schittenhelm, and G. Abstreiter, "Self-assembled growth of Sn on Ge (001)," Thin Solid Films, vol. 294, pp. 308–310, Feb. 1997.

[10] Y. Nakamura, A. Masada, S.-P. Cho, N. Tanaka, and M. Ichikawa, "Epitaxial growth of ultrahigh density $Ge_{1-x}Sn_x$ quantum dots on Si (111) substrates by codeposition of Ge and Sn on ultrathin SiO2 films," J. Appl. Phys., vol. 102, pp. 124302-124302–6, Dec. 2007.

[11] R. Ragan, K. S. Min, and H. A. Atwater, "Direct energy gap group IV semiconductor alloys and quantum dot arrays in Sn_xGe_{1-x}/Ge and Sn_xSi_{1-x} /Si alloy systems," Mater. Sci. Eng. B, vol. 87, pp. 204–213, Dec. 2001.

[12] N. Taoka, T. Asano, T. Yamaha, T. Terashima, O. Nakatsuka, I. Costina, P. Zaumseil, G. Capellini, S. Zaima, and T. Schroeder, "Non-uniform depth distributions of Sn concentration induced by Sn migration and desorption during GeSnSi layer formation," Appl. Phys. Lett., vol. 106, p. 61107, Feb. 2015.

[13] I. A. Fischer, T. Wendav, L. Augel, S. Jitpakdeebodin, F. Oliveira, A. Benedetti, S. Stefanov, S. Chiussi, G. Capellini, K. Busch, and J. Schulze, "Growth and characterization of SiGeSn quantum well photodiodes," Opt. Express, vol. 23, p. 25048, Sep. 2015.

[14] B. Schwartz, M. Oehme, K. Kostecki, D. Widmann, M. Gollhofer, R. Koerner, S. Bechler, I. A. Fischer, T. Wendav, E. Kasper, J. Schulze, and M. Kittler, "Electroluminescence of GeSn/Ge MQW LEDs on Si substrate," Opt. Lett., vol. 40, p. 3209, Jul. 2015.

[15] F. Oliveira, I. A. Fischer, A. Benedetti, P. Zaumseil, M. F. Cerqueira, M. I. Vasilevskiy, S. Stefanov, S. Chiussi, and J. Schulze, "Fabrication of GeSn-multiple quantum wells by overgrowth of Sn on Ge by using molecular beam epitaxy," Appl. Phys. Lett., vol. 107, p. 262102, Dec. 2015.

[16] F. Oliveira, I. A. Fischer, A. Benedetti, M. F. Cerqueira, M. I. Vasilevskiy, S. Stefanov, S. Chiussi, and J. Schulze, "Multi-stacks of epitaxial GeSn self-assembled dots in Si: Structural analysis," J. Appl. Phys., vol. 117, p. 125706, Mar. 2015.

PAPERS

Thermoelectric Properties of Polycrystalline WS_2 and Solid Solutions of $WS_{2-y}Se_y$ Types

G.E. Yakovleva*, A.I. Romanenko*, A.S. Berdinsky **, A.Yu. Ledneva*, V.A. Kuznetsov**, M.K. Han***, S.J. Kim*** and V.E. Fedorov*

*Nikolaev Institute of Inorganic Chemistry, Russian Academy of Sciences, Novosibirsk, Russia
** Novosibirsk State Technical University/Semiconductor Devices & Microelectronics, Novosibirsk, Russia
***Ewha Womans University/Dept. Chemistry and Nano Science, Seoul, Korea
e-mail: fed@niic.nsc.ru

Transition metal chalcogenides are perspective thermoelectric materials which have a great interest for application. In this work, the polycrystalline bulk WS_2 and solid solutions $WS_{2-y}Se_y$ types have been studied. In contrast to the literature data obtained at higher temperatures, we have investigated the thermoelectric properties of these materials at low and middle temperatures (77-450K). The temperature dependences of electrical conductivity and Seebeck coefficient were received from experimental data. The Seebeck coefficients of these materials have a high values, the maximum value up to 2000μV/K has been obtained.

I. INTRODUCTION

At our days, the main ecological problem is an environmental pollution. Over 60% of energy is wasting worldwide, mostly in the form of waste heat [1]. Therefore, thermoelectric power sources are one of the perspective fields of study. Thermoelectric materials have an ability to convert heat into electrical energy. The efficiency of thermoelectric materials is characterized by dimensionless thermoelectric quality factor ZT. This parameter depends on electrical conductivity (σ), Seebeck coefficient (S) and thermal conductivity (λ).

The layered transition metals chalcogenides are typical 2-D solid materials in a bulk form. Such materials have been used as solid lubricant, photovoltaic and photocatalytic solar energy converters, catalysts in many other industrial applications and others for many years [2]. Nevertheless, thermoelectric properties of layered transition metals chalcogenides are of great interest due to their low thermal conductivity.

One of the brightest representatives of such materials is a tungsten disulfide. Single-layer tungsten disulfide is a two-dimensional quasi-crystal, that consists of close-packed layer of tungsten in between of two close-packed layers of sulfur S-W-S. Layers are held together due to weak Van der Waals forces, whereas atomic structures of layers are tied together by strong covalent forces.

There are some papers devoted to the study of chalcogenides thermoelectric properties of transition metals such as WS_2 and WSe_2 where authors mainly studied single crystals of chalcogenides of transition metals [3]. In paper [4] authors investigated transport properties of ternary mixed $WS_{2-y}Se_y$ single crystals. In our work we have researched thermoelectric properties of polycrystalline solid solutions of $WS_{2-y}Se_y$ and $W_{1-x}Nb_xS_2$ for thermoelectric applications.

The work was supported by Russian Science Foundation, grant 14-13-00674.

II. EXPERIMENTAL RESULTS AND DISCUSION

A. Synthesis and characterization of the compositions

A series of samples of compositions $WS_{2-y}Se_y$ (y = 0.1, 0.2, 0.25) and $W_{1-x}Nb_xS_2$ (x = 0.05 and 0.15) were synthesized by means of ampoule high-temperature method. High purity elements were used for the syntheses. The starting powders of metals were annealed in hydrogen flow at 1000°C for 1 hour in order to remove adsorbed water and traces of oxides. Stoichiometric amounts of metal powder and chalcogenes were placed in quartz ampoules. The ampoules were vacuumed and sealed, heated up to 800°C during 5 hours and kept at 800°C for 4 days, then cooled and opened.

According XRD analysis the samples were single phases corresponded to 2H-WS_2 type (hexagonal, $P6_3/mmc$) with some broadening of reflections in comparison to pure WS_2(Fig.1).

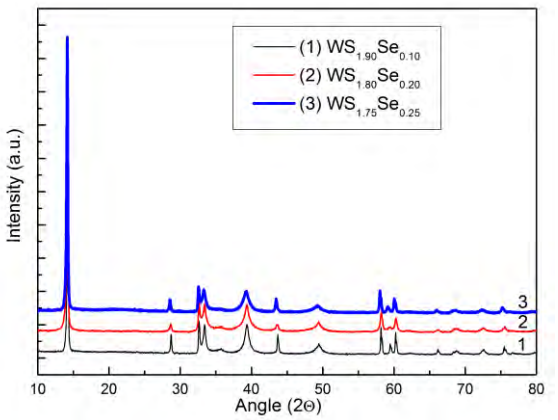


Figure 1. XRD powder patterns of $WS_{1.90}Se_{0.10}$, $WS_{1.80}Se_{0.20}$, $WS_{1.75}Se_{0.25}$

The crystal system is hexagonal, space group is $P6_3/mmc$ (no. 194). In chemistry of dichalcogenides this symmetry also called 2H-type of WQ_2 (or MoQ_2, where Q = S, Se, Te). There are also exist 3R-type corresponding to rhombohedral $R3m$ (no. 160) space group. All obtained compound are 2H-type which means they have $P6_3/mmc$ space group. If our samples were single-crystal, the diffraction reflexes would have been tight. Since the samples are fine powders, the broadening of reflexes occurs. The doping of Nb or Se leads to insignificant changes in unit cell parameters and small shift (about 0.1 degree) of reflexes in comparison with pure WS_2. Combination of these two factors leads to a broadening of the reflections compared to pure WS_2.

B. Preparation of the samples and measurement techique

Before measurements all samples have been kept under vacuum during 2 hours at 300° in order to remove absorbed water and oxygen from air. The XRD powder patterns of annealed samples completely agree with ones before evacuating. Thermal study of the samples shows stability of compounds up to 400°C (673 K).

The powdered materials were pressed to 10 mm in diameter pellets-shape samples. The samples 10 × 2 ×2 mm in size were cut from the pellets. Silver paste was used in order to obtain ohmic contact with samples.

The measurements of the electrical conductivity temperature dependence were performed by means of four-contact method. The thermopower was measured by means of static dc method. Thermal conductivity was determined by combining the thermal diffusivity D(T), specific heat Cp (T) and sample density ρ (T) according to κ_{tot} (T) = D(T) × Cp(T) × ρ(T). The thermal diffusivity D(T) and specific heat Cp(T) of several specimens were determined by the flash diffusivity-heat capacity method using NETZSCH LFA 457 MicroFlash™ instrument.

The temperature ranges for temperature dependence of Seebeck coefficient and electrical conductivity are different due to technical capability of measuring instruments.

C. Electrical properties

Our bulk samples consist of big quantity of the pressed nanoparticles. The current flow in such bulk sample will perform both inside nanoparticles and through the contacts between them. This leads to the additional variation in the electron transport properties of massive samples that consist of numerous nanoparticles. In addition, a poorly conducting layer is formed on the surface of the most of nanodimensional electroconductive particles.

In many cases, the electrical conductivity in arrays of such nanoparticles is determined mainly by the contact resistance. It is established experimentally that the electron transport properties (electrical conductivity, current–voltage characteristics, magnetoresistance, etc.) of materials substantially vary with their decreasing to nanometer sizes [5-10].

Our investigations of the electrical properties of materials containing nanoparticles with characteristic sizes of the order of several nanometers in different dielectric matrices manifested the variation of not only electrical conductivity but electron transport mechanisms as well [11-14]. It was established by the example of numerous systems that conductivity in polycrystalline materials with a high contact resistance is performed by tunneling charge carriers between the crystallites separated by conducting barriers (contacts between the crystallites) [15, 16].

If the sizes of crystalline islands are large sufficiently, the temperature dependence of conductivity σ(T) is described by the fluctuation model of tunneling – the fluctuation induced tunneling conduction (FIT) [17]:

$$\sigma(T) = \sigma_1 \cdot \exp[-T_t/(T+T_s)] \qquad (1)$$

where temperature T_t corresponds to the energy necessary for the electron transition between the crystallites (in fact, this transition is associated with overcoming energy gap $Eg \sim k_B \cdot T_t$); σ_1 is the intrinsic conductivity and T_s is the temperature, below which the conductivity reaches the saturation [17].

Electrical conductivity of the samples was measured in helium atmosphere in a temperature range from 77 K to 450 K. The measured results are displayed in Fig.2. The straight lines are approximations of the low temperature data by "(1)" with T_s=5K for all samples.

The results of temperature dependences of conductivity demonstrate that the FIT mechanism makes the dominant contribution to the low temperature dependence of conductivity for all samples. This fact leads us to conclusion that contact resistance between particles makes the main contribution to resistivity of all samples. On the figure 3 the energy gap ($Eg \sim k_B \cdot T_t$) of the $WS_{2-y}Se_y$ are presented.

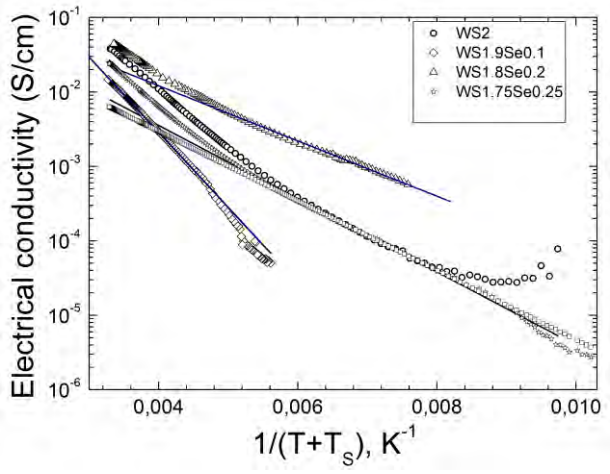


Figure 2. Temperature dependence of conductivity of the $WS_{2-y}Se_y$ samples in coordinates of dependence "(1)".

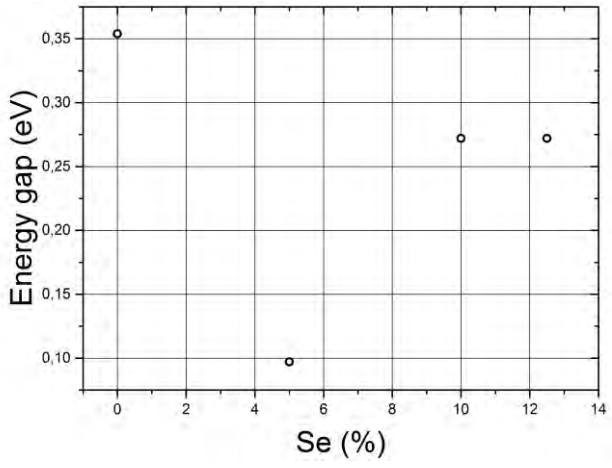


Figure 3. Material composition dependence of the $WS_{2-y}Se_y$ energy gap

D. *Thermoelectric properties*

Seebeck coefficient of the samples was measured at the temperature range of 190 to 450 K. The results on Seebeck coefficient are presented in Fig.4.

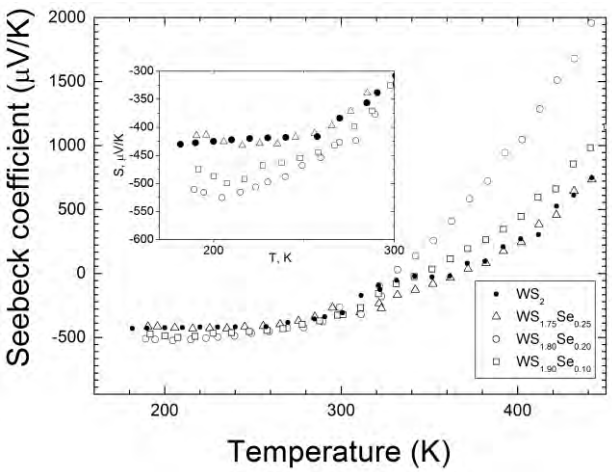


Figure 4. Temperature dependence of Seebeck coefficient of the $WS_{2-y}Se_y$ samples. The inset shows the results of measurements at low temperatures.

According to obtained data the maximum value of Seebeck coefficient S = 2000 μV/K has material with composition $WS_{1.8}Se_{0.2}$. And one has maximum value at low (S=-525μV/K at T=205 K) and middle temperatures (S = 1950 μV/K at T=445 K).

Such replacement of chalogen atom leads to increase in Seebeck coefficient. According to our preliminary measurements, the thermal conductivity of $WS_{2-y}Se_y$ is about 1-1.2 W/m*K. Such materials have a large value of the Seebeck coefficient on the one hand, and very low value of the electrical conductivity on the other hand. Therefore, in order to increase the electrical conductivity we have replaced transition metal atoms W partially by atoms Nb.

As we have shown in paper [18] such replacements significantly increased the electrical conductivity. The results of our work on measurements of replaced transition metal atoms in WS_2 by Nb are presented in Fig.5.

Such partial substitution of metal atoms in WS_2 leads to decrease in its energy gap and change its conductivity behavior to metallic one.

As seen from Figure 5, the material with composition $W_{0.85}Nb_{0.15}S_2$ has value of electrical conductivity greater than the one of the material with replacement $WS_{2-y}Se_y$ approximately by 10^3 times at T=320K. The materials $W_{0.85}Nb_{0.15}S_2$ and $W_{0.95}Nb_{0.05}S_2$ have energy gap value of 0.0001 eV and 0.02 eV correspondingly.

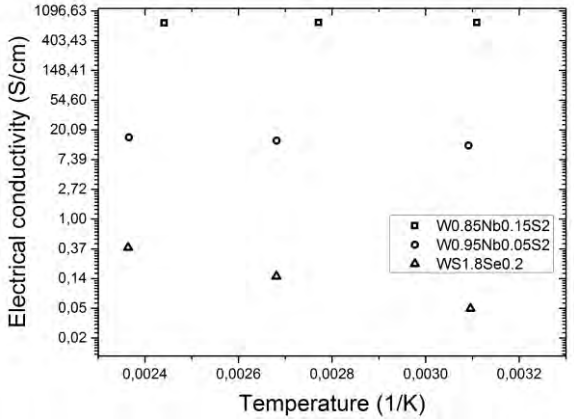


Figure 5. Temperature dependence of electrical conductivity of the $W_{1-x}Nb_xS_2$ samples.

E. *Power factor*

In present work we researched temperature dependence of Seebeck coefficient and electrical conductivity and we used power factor for the characterization of thermoelectric efficiency.

To evaluate the effectiveness of thermoelectric materials power factor was calculated by the formula "(2)".

$$P = S^2\sigma \qquad (2)$$

where S – Seebeck coefficient (μV/K), σ – electrical conductivity (S/cm).

The results of the thermoelectric power factor calculation are presented in Fig. 6.

One can see that these materials have not very high value of power factor in comparison with modern thermoelectric materials due to low electrical conductivity. But polycrystalline solid solutions of $WS_{2-y}Se_y$ have a very high Seebeck coefficient. These parameters are interdependent quantities [19] so a compromise between electrical conductivity and Seebeck coefficient needs to be found. We believe that replacement of metal atoms W in WS_2 partially by atoms Nb should be an efficient way of tuning thermoelectric properties towards the system's optimum.

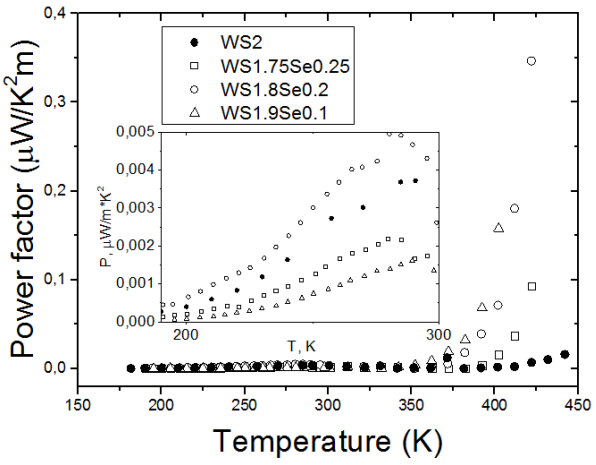


Figure 6. Temperature dependence of power factor of the $WS_{2-y}Se_y$ samples. The inset shows the results of measurements at low temperatures.

III. CONLUSION

A series of samples of compositions $WS_{2-y}Se_y$ (y = 0.1, 0.2, 0.25) and $W_{1-x}Nb_xS_2$ (x = 0.05 and 0.15) were synthesized by means of ampoule high-temperature method. Electron transport properties of obtained nanocomposite bulk materials WS_2 and solid solutions $W_{1-x}Nb_xS_2$, $WS_{2-y}Se_y$ at temperature range from 77 K to 450K were investigated.

We have found that Seebeck coefficients of these materials have a high value. The maximum value of the Seebeck coefficient up to 2000μV/K has been obtained for composition $WS_{1.8}Se_{0.2}$. However, the power factor P was found to be low. We have found that in nanocomposite bulk materials $W_{0.85}Nb_{0.15}S_2$ electrical conductivity increased by 10^3 times at room temperature, but the Seebeck coefficient was lower than in any of $WS_{2-y}Se_y$.

It was shown that it is possible to change thermoelectric and transport properties by controlling of Se and Nb contents in composites $WS_{2-y}Se_y$ and $W_{1-x}Nb_xS_2$.

In the scope of present work we succeed in increasing of the Seebeck coefficient by replacing chalcogen atoms in WS_2 by the factor of 2. But despite such a significant change of the Seebeck coefficient, these bulk polycrystalline composite materials still cannot compete with modern thermoelectric materials such as Bi_2Te_3 thin films with P= 2000 μW/m*K at room temperature [20].

ACKNOWLEDGMENT

The work was supported by Russian Science Foundation, grant 14-13-00674.

References

[1] M. G. Kanatzidis, “Nanostructured thermoelectrics: the new paradigm?,” Chemistry of materials, vol. 22(3), pp. 648–659, 2010.

[2] Haotian Wang, Hongtao Yuan, Seung Sae Hong, Yanbin Li and Yi Cui, “Physical and chemical tuning of two-dimensional transition metal dichalcogenides,” Royal society of chemistry, vol. 44, pp. 2664–2680, 2015.

[3] Jong-Young Kim, Soon Mok Choi, Won-Seon Seo and Woo-Seok Cho, “ Thermal and electronic properties of exfoliated metal chalcogenides,” Korean Chem.Soc, vol.3, pp.3225-3227, 2010.

[4] G. K. Solanki, D.N. Gujarathi, M. P. Lakshminarayana and M. K. Agarwal, “Transport property measurement in tungsten sulphoselenide single crystals grown by a CVT technique,” Crystal research and technology, vol. 43, pp. 179–185, 2008.

[5] Zaitsev-Zotov S.V., Pokrovskii V. Y., Monso P. “ Transition to 1D conduction with decreasing thickness of the crystals of TaS3 and NbSe3 quasi-1D conductors,” JETP Lett., vol. 73, pp. 29-32, 2001.

[6] Hor Y. S., Xiao Z. L., Welp U., Ito Y., Mitchell J. F., Cook R. E., et al. “Nanowires and Nanoribbons of Charge-Density-Wave Conductor NbSe3,” Nano Letters, vol.5,pp. 397-401, 2005.

[7] Stabile A. A., Whittaker L, Wu T. L., Marley P. M., Banerjee S, Sambandamurthy G. “Synthesis, characterization, and finite size effects on electrical transport of nanoribbons of the charge density wave conductor NbSe3,”Nanotechnology, vol.22, pp. 1 -12, 2011.

[8] Monceau P. “Electronic crystals: an experimental overview,” Advances in Physics, vol.61, pp. 325-581, 2012.

[9] Romanenko A.I., Anikeeva O.B., Kuznetsov V.L., Obrastsov A.N., Volkov A. P., Garshev A. V. “Quasi-two-dimensional conductivity and magnetoconductivity of graphite-like nanosize crystallites,” Solid State Communications, vol.137, pp. 625-629, 2006.

[10] Romanenko A. I., Anikeeva O. B., Buryakov T. I., Tkachev E. N., Zhdanov K. R., Kuznetsov V. L., et al. “Electrophysical properties of multiwalled carbon nanotubes with various diameters,” Phys Status Solidi B,vol.246, pp. 2641-2644, 2009.

[11] Chen J, Zhang G, Li B. “Impacts of Atomistic Coating on Thermal Conductivity of Germanium Nanowires,” Nano Lett, vol. 12, pp. 2826-2832, 2012.

[12] Romanenko A. I., Anikeeva O. B., Buryakov T. I., Tkachev E. N., Zhdanov K. R., Kuznetsov V. L., et al. “Influence of surface layer conditions of multiwall carbon nanotubes on their electrophysical properties,” Diamond & Related Materials, vol.19, pp. 964-967, 2010.

[13] Mazov I. N., Kuznetsov V. L., Moseenkov S. I., Ishchenko A. V., Rudina N. A., Romanenko A. I., et al. “Structure and Electrophysical Properties of Multiwalled Carbon Nanotube/Polymethylmethacrylate Composites Prepared via Coagulation Technique,” Nanoscience and Nanotechnology Letters, vol.3, pp. 18-23, 2011.

[14] Romanenko A. I., Fedorov V. E., Artemkina S. B., Anikeeva O. B., Poltarak P. A. “Temperature Dependences of Transport Properties of Films, Bulk Samples of Nanocrystals, and Single Crystals of Niobium Triselenide,” Physics of the Solid State. vol.57, pp. 1850-1854, 2015.

[15] Zhao Y, Li W. “Fluctuation-induced tunneling dominated electrical transport in multi-layered single-walled carbon nanotube films,” Thin Solid Films, vol.519, pp. 7987-7991, 2011.

[16] Romanenko A. I., Dybtsev D. N., Fedin V. P., Aliev S. B., Limaev K. M. “Electric-Field-Induced Metastable State of Electrical Conductivity in Polyaniline Nanoparticles Polymerized in Nanopores of a MIL-101 Dielectric Matrix,” JETP Letters, vol. 101, pp. 59-63, 2015.

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

[18] V.E. Fedorov, N.G. Naumov, A.N. Lavrov, M.S. Tarasenko, S.B. Artemkina, A.I. Romanenko and M.V. Medvedev, “Tuning Electronic Properties of Molybdenum Disulfide by a Substitution in Metal Sublattice,” 6th International Convention on Information & Communication Technology Electronics & Microelectronics,pp.11-14, 2013.

[19] Miroslav Ocko, Sanja Zonja and Mile Ivanda “Thermoelectric materials: problems and perspectives,”MIPRO 2010, pp.16-21, 2010.

[20] Jae-Hwan Kim, Jung-Yeol Choi, Jae-Man Bae, Min-Young Kim and Tae-Sung Oh “Thermoelectric characteristics of n-type Bi_2Te_3 and p-type Sb_2Te_3 thin films prepared by co-evaporation and annealing for thermopile sensor application,” Materials Transactions,vol.54, pp. 618-625, 2013

Piezoresistive Effect in Polycrystalline Bulk and Film Layered Sulphide $W_{0.95}Re_{0.05}S_2$

V. A. Kuznetsov[*, **, a], A. I. Romanenko[*], A. S. Berdinsky[**],
A. Yu. Ledneva[**], S. B. Artemkina[**], V. E. Fedorov[**, ***, b]
[*] Novosibirsk State Technical University / Semiconductor Devices and Microelectronics, Novosibirsk, Russia
[**] Nikolaev Institute of Inorganic Chemistry, Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia
[***] Novosibirsk State University, Novosibirsk, Russia
[a] e-mail address: vitalii.a.kuznetsov@gmail.com
[b] e-mail address: fed@niic.nsc.ru

Abstract - The paper reports the results of an experimental investigation of electron transport properties and piezoresistive effect of polycrystalline film and bulk samples of tungsten-rhenium disulphide $W_{0.95}Re_{0.05}S_2$. Polycrystalline powder of the composition was synthesized by direct high temperature reaction of elements with stoichiometric ration. The film samples were prepared by ultrasonication of the powder in 35% ethanol with subsequent spraying of the colloidal dispersion onto preheated substrates. The bulk samples were formed by conventional compress technology at a pressure of 1.25 GPa. The strain gauge factor is equal to 13 and 19 for film and bulk samples, respectively. The band gaps were estimated from temperature dependences of conductivity to be about 360 and 470 meV, respectively.

I. INTRODUCTION

New functional materials are very prospective for microelectronic application. Sensor electronics is one of the most interesting branches to use unusual properties of functional materials. Traditionally metal and semiconductor strain sensors are used to measure mechanical quantities. Strain gauge factor (SGF) is one of the main parameters to evaluate efficiency of strain sensors. The SGF of semiconductor sensors is one-two orders more than of the metallic ones, as well as SGF of monocrystalline semiconductor samples is several times more than of polycrystalline ones [1, 2].

One of the famous layered materials is graphene and graphene-based compounds. Electron transport and strain sensing properties of such materials were investigated quite well [3-10]. Similar to graphene transition metal dichalcogenides (TMDC) have layered structure, but their electron transport and strain sensing properties are not well known.

Electron transport properties of the thin polycrystalline films are different in comparison with bulk samples, because when passing from separate micro- or nanosized monocrystalline particles to polycrystalline arrays of them, the current flow in a bulk sample will perform both inside the particles and through the contacts between them. That leads to the additional variation in the electron transport properties of massive samples that consist of numerous nano- and microcrystallites. In addition, a low conducting layer is formed on the surface of the most of nanodimensional electroconductive particles. In many cases, the electrical conductivity in arrays of such particles is determined mainly by the contact resistance. It is established experimentally that the electron transport properties (electrical conductivity, current–voltage characteristics, magnetoresistance, etc.) of materials substantially vary upon their decreasing to nanometre sizes [11-15].

It was established by the example of numerous systems that conductivity in polycrystalline materials with a high contact resistance was performed by tunnelling charge carriers between the crystallites separated by conducting barriers (contacts between the crystallites) [16, 17]. Such electron transport is described by fluctuation induced tunnelling conduction (FITC) model with several parameters to be estimated [18]:

$$\sigma(T) = \sigma_1 \cdot \exp[-T_t / (T + T_s)], \quad (1)$$

where T_t is the temperature which corresponds to the energy necessary for the electron transition between the crystallites and T_s is the temperature, below which the conductivity reaches the saturation.

Recently we investigated electron transport properties of bulk polycrystalline samples of transition metal dichalcogenides (TMDC) and strain sensing properties of film samples of molybdenum-rhenium disulphide [19, 20]. But we have not investigated bulk samples for strain sensing properties. Tungsten disulphide WS_2 is one of bright representatives from layered materials family. 2H-WS_2 polymeric three-atom thick (S-W-S) layers are bound to the neighbouring layers via van der Waals S…S bonding. Due to the presence of weak van der Waals interaction MoS_2 and WS_2 can be dispersed under ultrasonic treatment in liquid medium forming colloidal dispersions with nanosized sheets. That exfoliation method allows producing of stable colloidal dispersions with the particles a few layers thick and tens to hundreds nanometre in lateral size [21], and such dispersions can be used to form thin films, which consist of the nanoparticles.

The study was supported by the Russian Science Foundation (Grant no. 14-13-00674).

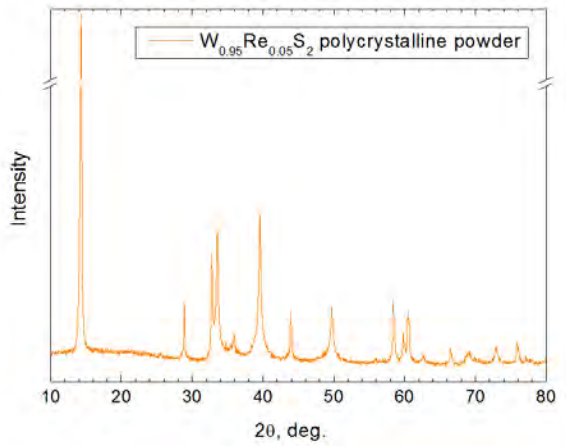


Figure 1. X-ray powder diffraction pattern of $W_{0.95}Re_{0.05}S_2$ synthesized in high-temperature ampoule approach

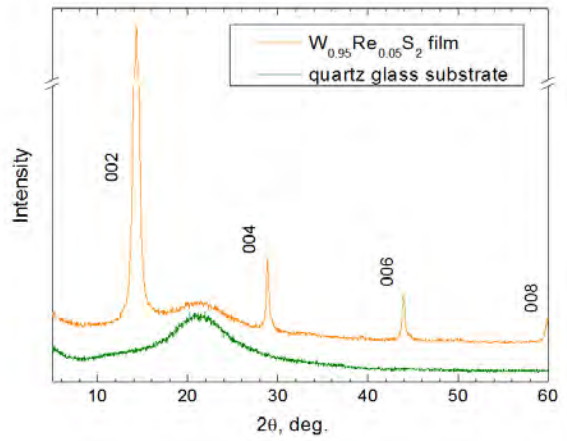


Figure 2. X-ray powder diffraction pattern of the $W_{0.95}Re_{0.05}S_2$ film sprayed onto preheated amorphous quartz glass substrate

The goal of this work is to investigate strain gauge factor and electron transport properties of film and bulk samples of tungsten disulphide doped with rhenium.

II. EXPERIMENTAL

A. *Preparation of $W_{0.95}Re_{0.05}S_2$ film and bulk samples*

We used the same methods and apparatuses as we did in [22]. $W_{0.95}Re_{0.05}S_2$ was synthesized via ampoule high-temperature method from stoichiometric mixture of the elements. The synthesized $W_{0.95}Re_{0.05}S_2$ was analysed by X-ray powder diffraction method. XRD-analysis has shown the single phase with some broadening of reflections (see Fig. 1). To prepare film samples 1.0 g of the powder was placed into a glass flask with 250 ml of ethanol-water solution (35 / 65% in volume) and ultrasonicated for 48 h. The resulting mixture was centrifuged at 1600 rpm for 25 min. The colloidal dispersion was sprayed onto a preheated to 180°C steel beam covered with polymer glue VL-931 like in [22]. Thickness of the samples resulted was estimated by weighting to be about 1 μm. For studying electron transport properties the dispersion was sprayed the same way onto Al_2O_3 polished substrates. The XRD pattern of thin film sample sprayed onto quartz glass polished is shown in Fig. 2. As one can see from Fig. 2 only *00l* peaks are visible and it is indirect proof that the particles in films sprayed are oriented generally in parallel to the substrates.

Bulk samples were formed of the powder with a laboratory hydraulic press at a pressure of 1.25 GPa. Thickness of the tablets was about 0.25 mm. To investigate the properties the tablets were cut to strips 2 mm in width. Samples for studying strain sensing properties were glued to a beam of uniform strength (in bending) like in [22]. The glue was polymerized according to its bonding technology with the maximum temperature of 180°C. Before heating samples were covered additionally with the glue to reduce an environmental influence.

TABLE I. SAMPLES' PARAMETERS

Samples	Sprayed	Pressed
Thickness, *d* (μm)	1	250
Resistivity, ρ (Ohm·cm)	11000	166
Band gap, Δ (meV)	470	360
Strain gauge factor, *K*	13.0	19.5

B. *Electron transport properties*

The temperature dependences of resistivity were measured using four-point probe method for bulk samples and two-point probe method for film ones. Contacts to the samples were made of the silver paste and thin gold wires. Two-point probe method was used for films instead of four-probe one because of high resistance of the films, but it should be noted that the contacts were ohmic. The dependences measured are shown in Fig. 3 and Fig. 4 in different axes. As one can see from Fig. 4 the experimental curves are well fitted by the equation (1). That exponential dependence is typical for fluctuation induced tunnelling conduction with the temperature T_t corresponding to the energy necessary for the electron transition between the crystallites. In fact that transition is associated with overcoming energy gap $\Delta \sim k_B \cdot T_t$, the energy gap an electron has to tunnel through from one crystallite to another. The energy gaps Δ were estimated from the slopes of the fitting straight lines (see Fig. 4). The results are shown in Table 1. As intercrystalline boundaries are most likely to be smaller in tablets in contrast to films due to pressing, it seems reasonable to say that the band gaps difference is concerned with the difference of the contacts.

C. *Strain sensing properties*

The device for measuring SGF was the same as we used in [20, 22]. Electrical contacts to the samples were made of silver paste "Dotite D-500" and thin copper wires. The contacts were ohmic. Every compression-tension cycle was lasted for 10 minutes with loading and unloading period being the same. The dependences of resistance on strain are shown in Fig. 5 for film samples in comparison to bulk ones. SGF was defined as slope of the line connecting two extreme values. The results are shown in Table 1. One can see that SGF of bulk samples is more

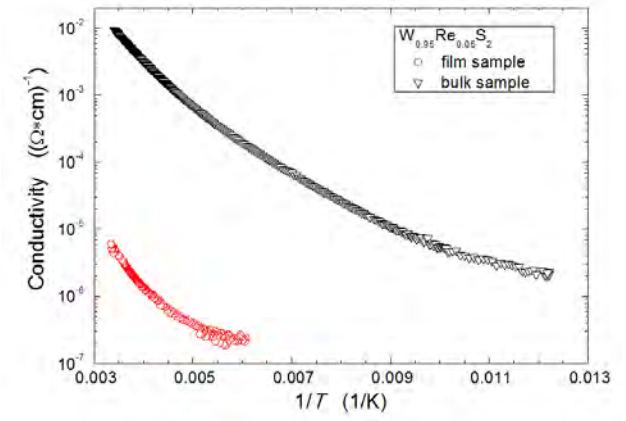


Figure 3. Temperature dependences of conductivity of the film and bulk samples

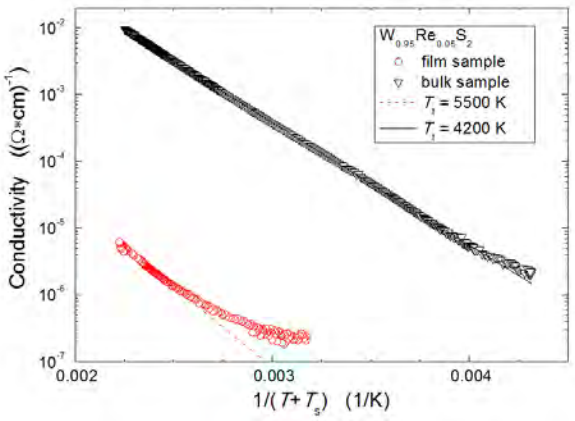


Figure 4. Temperature dependences of conductivity with the fitting lines for energy gaps estimation by equation (1)

than SGF of thin films. It was shown above that the intercrystalline contacts were different in bulk and film samples, and most likely the reason of the difference in the SGF is different contribution of the crystallites themselves and the contacts between them to changing of electrical resistance of the bulk and film samples when mechanical strain is applied. Especially as particles of layered compounds a few layer thick like ones the films involved consist of have great Young's modulus [23]. Anyway the difference should be estimated with a conventional beam because of small thickness of the beams used and relatively large thickness of the bulk samples.

III. CONLUSION

We have studied piezoresistive effect and electron transport properties of film and bulk polycrystalline tungsten-rhenium disulphide $W_{0.95}Re_{0.05}S_2$ samples. Temperature dependences of conductivity are well described by fluctuation induced tunnelling conduction model. The band gaps have been estimated to be about 470 and 360 meV for film and bulk samples, respectively. The strain gauge factors of film and bulk samples have been estimated from experimental curves to be about 13 and 19, respectively. The difference in energy gaps can be explained by lesser contribution of contact resistance in electrical conductivity of bulk samples in contrast to films as a result of better densification of the crystallites due to pressing. The SGF of bulk samples is higher than the factor of films. It can be circumstantial evidence that intercrystalline boundaries make great contribution in piezoresistive effect.

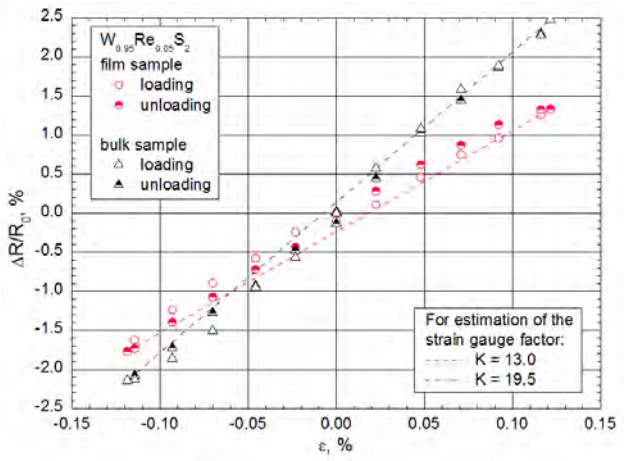


Figure 5. Resistance-strain characteristic of $W_{0.95}Re_{0.05}S_2$ film and bulk samples at room temperature in ambient

REFERENCES

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

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

[3] S. D. Sarma, A. K. Geim, P. Kim, and A. H. MacDonald, "Special edition: Exploring graphene - Recent research advances," Solid State Communications, vol. 143, pp. 1-126, 2007.

[4] B. Partoens and F. M. Peeters, "From graphene to graphite: Electronic structure around the K point," Physical Review B, vol. 74, pp. 075404-1-10, 2006.

[5] C. Gomez-Navarro, R. T. Weitz, A. M. Bittner, M. Scolari, A. Mews, M. Burghard, et al., "Electronic transport properties of individual chemically reduced graphene oxide sheets," Nano Letters, vol. 7, pp. 3499-3503, 2007.

[6] X. Du, I. Skachko, A. Barker, and E. Y. Andrei, "Approaching ballistic transport in suspended graphene," Nature Nanotechnology, vol. 3, pp. 491-495, 2008.

[7] G. Eda, G. Fanchini, and M. Chhowalla, "Large-area ultrathin films of reduced graphene oxide as a transparent and flexible electronic material," Nature Nanotechnology, vol. 3, pp. 270-274, 2008.

[8] G. T. Pham, Y.-B. Park, Z. Liang, C. Zhang, and B. Wang, "Processing and modeling of conductive thermoplastic/carbon nanotube films for strain sensing," Composites Part B-Engineering, vol. 39, pp. 209-216, 2008.

[9] N. Hu, Y. Karube, M. Arai, T. Watanabe, C. Yan, Y. Li, et al., "Investigation on sensitivity of a polymer/carbon nanotube composite strain sensor," Carbon, vol. 48, pp. 680-687, 2010.

[10] A. Bessonov, M. Kirikova, S. Haque, I. Gartseev, and M. J. A. Bailey, "Highly reproducible printable graphite strain gauges for flexible devices," Sensors and Actuators A-Physical, vol. 206, pp. 75-80, 2014.

[11] P. Monceau, "Electronic crystals: an experimental overview," Advances in Physics, vol. 61, pp. 325-581, 2012.

[12] A. A. Stabile, L. Whittaker, T. L. Wu, P. M. Marley, S. Banerjee, and G. Sambandamurthy, "Synthesis, characterization, and finite size effects on electrical transport of nanoribbons of the charge density wave conductor $NbSe_3$," Nanotechnology, vol. 22, p. 485201 (6pp), 2011.

[13] A. I. Romanenko, O. B. Anikeeva, V. L. Kuznetsov, A. N. Obrastsov, A. P. Volkov, and A. V. Garshev, "Quasi-two-dimensional conductivity and magnetoconductivity of graphite-like nanosize crystallites," Solid State Communications, vol. 137, pp. 625-629, 2006.

[14] Y. S. Hor, Z. L. Xiao, U. Welp, Y. Ito, J. F. Mitchell, R. E. Cook, et al., "Nanowires and Nanoribbons of Charge-Density-Wave Conductor $NbSe_3$," Nano Letters, vol. 5, pp. 397-401, 2005.

[15] S. V. Zaitsev-Zotov, V. Y. Pokrovskii, and P. Monceau, "Transition to 1D Conduction with Decreasing Thickness of the Crystals of TaS_3 and $NbSe_3$ Quasi-1D Conductors," JETP Lett., vol. 73, pp. 29-32, 2001.

[16] A. I. Romanenko, D. N. Dybtsev, V. P. Fedin, S. B. Aliev, and K. M. Limaev, "Electric-Field-Induced Metastable State of Electrical Conductivity in Polyaniline Nanoparticles Polymerized in Nanopores of a MIL-101 Dielectric Matrix," JETP Letters, vol. 101, pp. 59-63, 2015.

[17] Y. Zhao and W. Li, "Fluctuation-induced tunneling dominated electrical transport in multi-layered single-walled carbon nanotube films," Thin Solid Films, vol. 519, pp. 7987-7991, 2011.

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

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

[20] V. A. Kuznetsov, A. S. Berdinsky, A. Y. Ledneva, S. B. Artemkina, M. S. Tarasenko, and V. E. Fedorov, "Strain-sensing element based on layered sulfide $Mo_{0.95}Re_{0.05}S_2$," in 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2015, pp. 15-18.

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

[22] V. A. Kuznetsov, A. S. Berdinsky, A. Y. Ledneva, S. B. Artemkina, M. S. Tarasenko, and V. E. Fedorov, "Film $Mo_{0.95}Re_{0.05}S_2$ as a strain-sensing element," Sensors and Actuators A: Physical, vol. 226, pp. 5-10, 2015.

[23] A. Castellanos-Gomez, M. Poot, G. A. Steele, H. S. J. van der Zant, N. Agrait, and G. Rubio-Bollinger, "Elastic Properties of Freely Suspended MoS_2 Nanosheets," Advanced Materials, vol. 24, pp. 772-775, 2012.

Luminescent diagnostics in the NIR-region on a base of Yb-porphyrin complexes

V.D. Rumyantseva[1,2], I.P. Shilov[2], Yu.V. Alekseev[3], A.S. Gorshkova[1]

1. Moscow Technological University, 119454 Moscow, Russia. E-mail: vdrum@mail.ru
2. Kotel'nikov Institute of Radioengineering and Electronics RAS, 141190 Fryazino, Moscow region, Russia. E-mail: ipshilov@ms.ire.rssi.ru
3. State Scientific Center of Laser Medicine, Moscow, Russia. E-mail: ural377@mail.ru

The problem of early diagnostics, necessary for successful therapy of oncological diseases, is well known. In this field an optical diagnostic methods are the most effective. The optical response of a luminescent label can indicate the condition of biological tissues and biochemical processes occurring in them in real time. The most promising luminescent labels for non-invasive diagnostics are those that absorb and emit light in the NIR-region (~700-1100 nm), where the biological tissues absorption and auto-fluorescence are minimal. One of the most promising studied by us label is Yb-complex of 2,4-dimethoxyhematoporhyrin IX, it possess the optimum chemical and photophysical properties such as a tumorotropic, a high molar extinction coefficient, a large Stokes shift, an effective luminescence, chemical and light stability and an ability to be used in aqueous media. Therefore, a special attention is given to luminescent diagnostics method development for endoscopic and visually available cancer forms on a base of Yb-complex of 2,4-dimethoxyhematoporhyrin IX and laser-fiber NIR-range fluorimeter.

I. Introduction

Various porphyrin compounds perform important functions, which are indispensable for the existence and development of flora and fauna on Earth. At the end of the XX century porphyrins have been used in photodynamic therapy and diagnostics of malignant tumors due to their ability to accumulate in various types of cancer cells and tumor microvessels. For today a whole series of photosensitizers effectively generating a singlet oxygen, were synthesized: Photofrin II, Photogem, Foscan, Fotoditazin, Photolon, Radachlorin, Photosens, Alasens, Tookad and others.

However, the free bases of these macroheterocycles have a side effect - phototoxicity, which unfavorable influence during diagnostics procedures. It is possible to overcome that drawback by use of diagnostics photosensitizers which are practically do not regenerate a singlet oxygen, while maintaining a high affinity to malignant tumors. Ytterbium complexes of natural and synthetic porphyrins are such compounds [1, 2].

II. Results and discussion

Lanthanide luminescence weak enough by itself is significantly enhanced in a porphyrins metallocomplexes. It is connected to the transfer of macrocycle excitation energy to the La^{3+} ion [3]. Among lanthanides the erbium, neodymium and ytterbium complexes whose La^{3+} luminescent level lies below T_1-porphyrin level, are of the most interest.

Porphyrins complexes with Er, Nd and Yb possess 4f-luminescence in the near-infrared region (NIR-region) of spectrum, which is become possible because of an intramolecular energy transfer from the triplet state of a porphyrin (located in the range of 12 500-13 500 cm^{-1}) to the lower resonance levels of Er^{3+}, Nd^{3+} and Yb^{3+} (6 450, 11 500 and 10 200 cm^{-1} respectively) (fig. 1) [4].

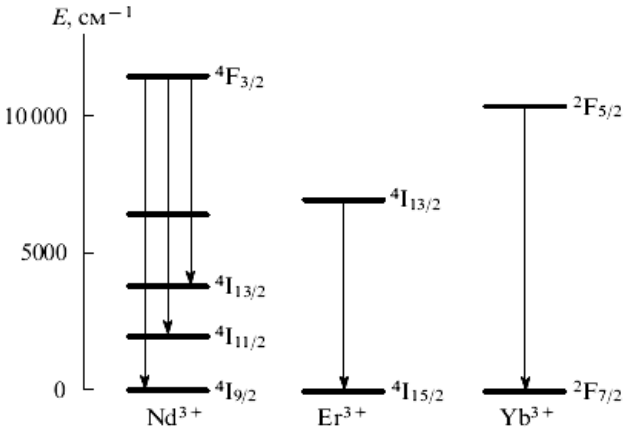


Figure 1. Diagram of energy levels and luminescence spectra of Er^{3+}, Nd^{3+} и Yb^{3+} ions.

Ytterbium porphyrins complexes have been chosen as research objects due to the fact that under excitation of the π-electron part of molecule the luminescence which occur from transitions $^2F_{5/2} \rightarrow {}^2F_{7/2}$ of the Yb^{3+} 4f-electron level ($^2F_{5/2}$ - excited state, $^2F_{7/2}$ - the ground state), is observed. Introduction of ytterbium ion in a porphyrin leads to reduce of photochemical activity, but the selectivity of accumulation in tumors, characteristic to the most of porphyrins, still remains. The reduction of a singlet oxygen quantum yield is causes by that the luminescent level of Yb^{3+} ion lies rather below the triplet state of the molecule organic part, but higher than that of a singlet oxygen. As a result, the porphyrin matrix excitation under the influence of external light radiation is not transferred to oxygen, but intercepted by Yb^{3+} ion, thereby strongly reducing the sensitized by porphyrin singlet oxygen generation [5]. These transformations are shown in fig. 2.

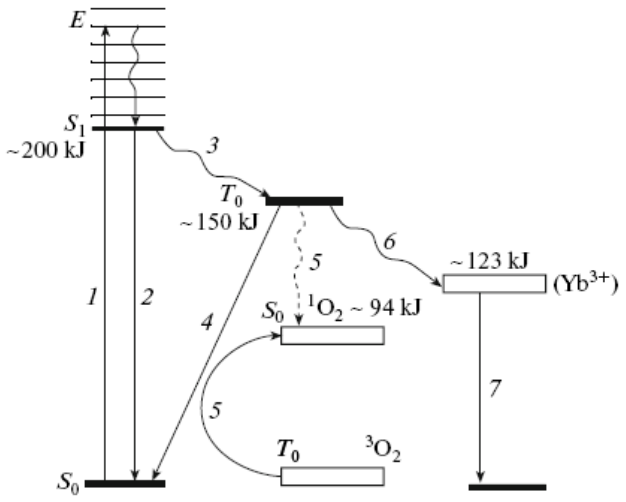


Figure 2. Scheme of electronic transitions of porphyrin sensitizers and singlet oxygen generation: *(1)* absorption, (2) fluorescence, (3) intercombination conversion, (4) phosphorescence, (5) excitation transfer to oxygen and the transition of triplet oxygen $^3O_2$ to singlet oxygen, (6) excitation transfer to the Yb^{3+} ion, and (7) luminescence of the Yb^{3+} ion.

Porphyrin molecules form stable complexes with ytterbium ions which have intense absorption in the near infrared region (NIR-region) of spectrum [6]. The extinction coefficient (ε) for Yb-complexes is 10^4-10^5 $M^{-1}cm^{-1}$ that is almost 4 orders higher than ε value under direct excitation of Yb^{3+} ion itself, so one can assume that Yb^{3+} ion excitation through porphyrin matrix provides the most effective way for strong 4f-luminescence than direct excitation of Yb^{3+}. At the same time, introduction of various substituents in *meso* and/or β-positions of macrocycle allows to drastically modify the physicochemical properties of lanthanide porphyrins complexes, that plays the important role in medicine and photochemistry use [7].

The first studies of ytterbium porphyrins complexes as luminescent markers on animals with malignant tumors were carried out on liposomal forms of coproporphyrin III, protoporphyrin IX, hematoporphyrin IX as their methyl ethers [8], as well as water-soluble synthetic derivatives of tetraphenylporphyrin [9].

Continuing these studies more than two dozen ytterbium complexes of natural and synthetic porphyrins were synthesized [10]. The analysis of physical, chemical and luminescence characteristics and the results of biological tests revealed that one of the most promising compounds for diagnostic purposes is the dipotassium salt of Yb-2,4-dimethoxyhematoporphyrin IX complex [11] (fig. 3).

This complex is similar in structure to natural protoporphyrin IX, which iron complex is the hemoglobin prosthetic group. The substance is low toxic, well soluble in water, its synthesis based on blood hemin, is simple and cheap [12].

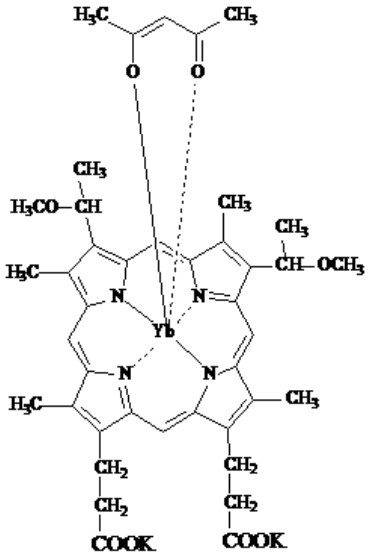


Figure 3. Structural formula of dipotassium salt of Yb-2,4-dimethoxyhematoporphyrin IX complex.

In studies of ytterbium ions infrared luminescence in complexes solutions with organic reagents the main purpose is to minimize a non-radiative loss of excitation energy. In macrocyclic ligands a complexing ion is effectively protected from effects of high frequency O-H and C-H oscillations of solvent molecules, which play an important role in a non-radiative degradation of electronic excitation energy.

Ytterbium porphyrins complexes have the characteristic for rare-earth ions narrow and rather bright luminescence line, which for Yb^{3+} ion locates in the infrared range of 975-985 nm in a "therapeutic window of biological tissues transparency", where their own luminescence is practically absent. The lifetime (τ) for the Yb-2,4-dimethoxyhematoporphyrin IX complex was 11 μs [13], the luminescence decline have a non-exponential pattern that is due to a strong luminescence quenching by a fluctuations of OH-groups from an inner environment of the ytterbium ion. A significant difference in lifetimes of the excited state of ytterbium complexes of porphyrins hydrophobic ethers and their acids is caused by a presence of intermolecular hydrogen bonds and luminescence quenching by water [14].

To increase diagnostic potential of ytterbium porphyrin complexes it is necessary to isolate them from the quenching effect of a water environment whenever possible. A preferred solvent for such compounds may be DMSO, which has unique biomedical and pharmacological properties: it penetrates through biological membranes, improves transport properties of drugs, stimulates immune system.

The emission luminescence spectra of the Yb-2,4-dimethoxyhematoporphyrin IX complex in aqueous solutions with different concentrations of DMSO are shown on fig. 4.

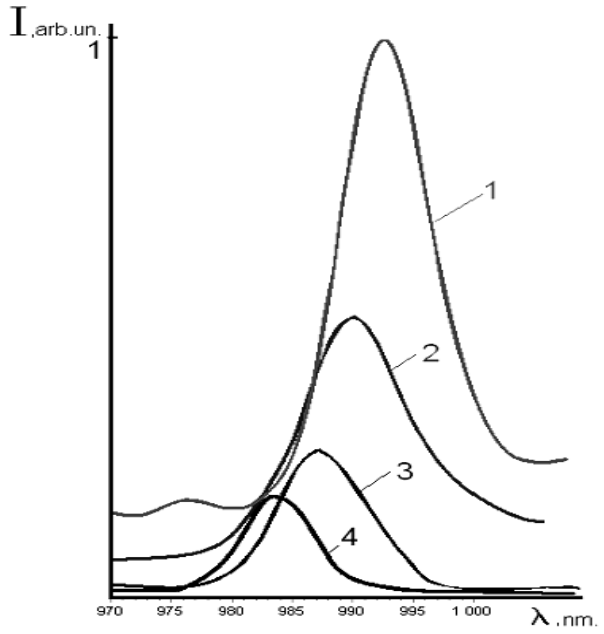


Figure 4. Emission luminescence spectra of the Yb-2,4-dimethoxy-hematoporphyrin IX in aqueous solutions with different DMSO concentrations: 1 - 100% DMSO, 2 - 50% DMSO, 3 - 20% DMSO, 4 - 100% H_2O.

Under conditions of lower polarity (solutions with growing concentration of DMSO) emission maxima are shifted toward long-wave spectrum region (solvatochromism phenomenon) [15]. From fig. 4 one can see that the luminescence intensity increases significantly with increasing of DMSO concentration (> 10 times in transition from Yb-complex aqueous solution to 100% DMSO solution), and the emission spectrum maximum shifts to almost 10 nm at the same time. The lifetime in 100% DMSO solution was ~ 22 μs. A 20-30% aqueous DMSO solutions are allowed in medicine and are of practical interest in use them for intravenous injections. For them $\tau \sim 5 \div 10$ μs.

Study of a photosensitized luminescence kinetic signals of singlet oxygen in aqueous solutions showed that the quantum yield of a singlet oxygen generation for the Yb-2,4-dimethoxyhematoporphyrin IX complex reduces almost 4 times (to 11%) from 40% for a free base porphyrin, that experimentally confirms its low phototoxicity.

Previously in *in vivo* experiments on mice with subcutaneously grafted sarcoma 5-37 it was found by luminescence in the NIR-region the preferential Yb-complex accumulation in tumor tissue compared with normal one [16]. The injection of the dipotassium salt of Yb-complex was carried out intravenously, the luminescent accumulation contrast was determined after 48 hours.

Continuing the studies we designed the amphiphilic pharmaceutical compositions in the form of gels for epikutan use as well as for application to mucous membranes. The optimum concentration of the Yb-2,4-dimethoxyhematoporphyrin IX complex (~ 0.05%, w/w) was received, gels Tisolum, Kalgel, Cremophor and their mixtures in various proportions were used. The Yb-complex pharmaceutical composition was found to accumulate fast enough (less than 1 hour) in places with pathological changed skin and mucous membranes. Herewith the clear luminescence intensity difference was established compared to healthy tissue (900-1100 nm range). The luminescence parameters change depends on a measurement time and a pathological process character [17].

This method can be successfully applied in dermatology, dentistry, gynecology, veterinary and other fields of medicine; it is characterized by simplicity of performance, availability, informativity and low toxicity. The best results of accumulation in pathological changed skin were obtained for the pharmaceutical composition based on Tisolum. The contrast index value was in range 3.0-15.0.

III. Conlusion

Thus, ytterbium porphyrins complexes are promising diagnostic markers of malignant tumors and pathological changes of skin and mucous membranes in the NIR-region of spectrum and possess a low phototoxicity.

Acknowledgment

This work was supported by the Russian Federation Ministry of Education and Science, the project № 4.128.2014/K.

References

[1] M.I. Gaiduk, V.V. Grigoryants, A.F. Mironov, L.D. Roytman, V.I. Chissov, V.D. Rumyantseva, G.M. Sukhin. Dokl. AN SSSR. Seriya Biofizika. 1989, vol. 309, N 4, pp. 980-983 (in russ).

[2] M.I. Gaiduk, V.V. Grigoryants, A.F. Mironov, V.D. Rumyantseva, V.I. Chissov, G.M. Sukhin. J. Photochem. Photobiol., B.: Biology. 1990. vol. 7, pp. 15-20.

[3] A.F. Mironov. Uspekhi Khimii. 2013, vol. 82, N 4, pp. 333-351 (in russ).

[4] I.P. Shilov, A.V. Ivanov, V.D. Rumyantseva, A.F. Mironov. In book: Fundamental scienses for medicine. Biophysical medical technologies (Ed. A.I. Grigorev and Yu.A. Vladimirov). M.: Maks-Press. Vol. 2, 2015, pp. 110-144 (in russ).

[5] A.V. Ivanov, V.D. Rumyantseva, K.S. Shchamkhalov, I.P. Shilov. Laser Physics, 2010, vol. 20, N 12, pp. 2056-2065.

[6] M. Gouterman. The Porphyrins (Ed. D. Dolphin), N 4, San Francisco, London: Academic Press. 1978, vol. 3, pp. 1-156.

[7] M.I. Gaiduk, V.V. Grigoryants, A.F. Mironov, V.D. Rumyantseva. Proc. Estonian Acad. Sci. Phys. Math., 1991, vol. 40, N 3, pp. 198-204.

[8] RF Patent 1340087. Byull. Izobret. N 18 filed 27.06.1995. P. 273.

[9] M.I. Gaiduk, V.V. Grigoryants, V.D. Menenkov, A.F. Mironov, V.D. Rumyantseva. Izv. AN SSSR. Seriya Fizika. 1990, vol. 54, N 10, pp. 1904-1908 (in russ).

[10] V.D. Rumyantseva, A.S. Gorshkova, A.F. Mironov. Fine Chem. Tech., 2014, vol. 9, N 1, pp. 3-17.

[11] RF Patent 2411243. Byull. Izobret. N 4 filed 10.02.2011. P. 731.

[12] V.D. Rumyantseva, A.F. Mironov, I.P. Shilov, K.S. Shchamkhalov, A.S. Ryabov, A.V. Ivanov. Abstracts of the XIII Int. Sci-Techn. Conf. «High-Tech in Chemical Engineering-2010». Russia, Suzdal. 29.06–02.07.2010. P. 192 (in russ).

[13] V.D. Rumyantseva, A.F. Mironov, K.S. Shchamkhalov, G.M. Sukhin, I.P. Shilov, V.M. Markushev, Z.V. Kuzmina, N.I. Polyanskaya, A.V. Ivanov. Lazernaya Medicina, 2010, vol. 14, N 1, pp. 20-25 (in russ).

[14] A.S. Stachevski, V.N. Knyukshto, A.V. Ivanov, V.D. Rumyantseva, I.P. Shilov, V.A. Galievsky, B.M. Dzhagarov. Abstracs of the Int. Conf. «Molecular and cellular basics of biosistems functioning». 17.06–20.06.2014. Belarus, Minsk: Book of Abstracs in 2 ch. Ch. 1, 2014, pp. 128-130 (in russ).

[15] A.S. Stachevski, V.N. Knyukshto, A.V. Ivanov, V.D. Rumyantseva, I.P. Shilov, V.A. Galievsky, B.M. Dzhagarov. J. Appl. Spectr., 2014, vol. 81, N 6, pp. 938-942.

[16] V.D. Rumyantseva, K.S. Shchamkhalov, I.P. Shilov, L.U. Kochmarev, V.M. Markushev, Z.V. Kuzmina, N.I. Polyanskaya, A.S. Ryabov, A.V. Ivanov. Medicinskaya fizika, 2011, N 2, pp. 67-73 (in russ).

[17] Yu.V. Alekseev, A.V. Ivanov, A.S. Ryabov, N.M. Shumilova, I.P. Shilov. Ros. Bioterapevticheskiy Zhurnal, 2015, vol. 14, N 1, P. 59 (in russ).

Simulation Study of the Composite Silicon Solar Cell Efficiency Sensitivity to the Absorption Coefficients and the Thickness of intrinsic Absorber Layer

V. Tudić* N. Posavec*
* Karlovac University of Applied Sciences/Department of Mechanical Engineering, Karlovac, Croatia
vladimir.tudic@vuka.hr
nikola.posavec.1@gmail.com

ABSTRACT - In this paper, two silicon solar cells p^+-i_i-n^+ with homogenous and heterogeneous intrinsic absorber layers based on hydrogenated amorphous-nanocrystalline-microcrystalline silicon (a-Si:H/nc-Si:H/μc-Si:H) have been studied by computer modeling and simulation program (AMPS-1D - Analysis of Microelectronic and Photonic Structures). Various factors that affect cell efficiency performance have been studied such as layers absorption coefficients, band gap and layer thickness up to 1200nm. It was found that in the case of standard solar cell conditions a layers absorption coefficient has a major contribution to solar cell performance according to measurement on the actual solar cell samples. It is demonstrated that, for homogenous a-Si:H/nc-Si:H intrinsic absorber layer with constant crystal fraction of Xc=30% cell efficiency is higher than in case of heterogeneous intrinsic absorber layer which contains various crystal fractions depending of absorber layer thickness. Second case scenario of silicon thin film composite structure is more common in solar cells production by using PECVD and HWCVD deposition techniques which is proven by X-ray diffraction and high resolution electron microscopy measurements.

I. INTRODUCTION

In previous work a semi conducting silicon properties, hetero-junctions and photo effect in intrinsic silicon thin films have been studied, and optical generation and recombination of free carriers have been investigated [1, 2]. Also, bases physical principles of simple silicon solar cells with accent to amorphous-nanocrystalline layers as promissing composite material for high efficiency solar cells called third generation have been studied. In present work basic principles of *pin* structure solar cells have been carried out with one dimensional computer modelling programme AMPS-1D. Suggested solar cell model are based on simple *pin* structure in order to compare simulation results with others in references. Computer modelling programmed allows solar cell parameter calculations and structure design simulations. By varying characteristic set of solar cells parameters such as: illumination spectra, photon flux, absorption coefficient, boundary conditions, front and back contact parameter, general silicon layer parameters as doping and free carrier concentrations, mobility, gap state defect distribution, I-V characteristic, fill factor (FF), and efficiency (η) of solar cell can be determinate. In this work a few simulation groups are carried out with homogenous distribution of Si nc in layer and with layers where crystallinity and crystal sizes change across the layer. Calculated solar cells efficiency of modelled single and multilayer absorber structures were graphically presented and discussed.

A goal of this work is to suggest possible application of a-Si:H/nc-Si:H/μc-Si:H film as active part in a typical *pin* solar cell in respect to overall performance. A detailed comparison of calculated data extracted from defined model suggest optimal absorber layer thickness in composite silicon solar cell and leads to a better understanding of effective solar cell thickness.

II. INTRINSIC LAYER MICROSTRUCTURE

Amorphous-nanocrystalline silicon (a-nc-Si) films of few hundred nanometers in thickness consist of a matrix of amorphous silicon with embedded silicon crystals of nanometric dimensions [3, 4]. This material has improved properties with respect to pure amorphous Si (a-Si), micro-crystalline Si (μc-Si) and bulk crystalline Si (c-Si) owning quantum confinement effect [5]. When compared to amorphous silicon, nc-Si has better electrical transport characteristic [6], possibility to tailor the optical band gap [7] and resistance to light induced degradation [8].

Micro-crystalline silicon (μc-Si) films of 1-2 micrometers in thickness consist of small crystal aggregates in deeper layers and large grains/columns up to film surface [9, 10]. On the other hand micro-crystalline Si (μc-Si) and bulk crystalline Si (c-Si) compared to nc-Si has better conductivity along crystal grains according to higher free carrier mobility and less numbers of grain boundaries. Therefore μc-Si layers has better electrical transport characteristic compared to a-nc-Si [11].

The main advantage of nc-Si in comparison with crystalline silicon is its higher absorption, which allows efficient solar cells in thin film designed device. This

material could be fraction of "small crystals", expected to show effect of increase in optical gap due to quantum confinement predictable according to effective mass theory or quantum dots.

As a result of PECVD and HWCVD deposition techniques micro-crystalline silicon thin films typically forms a microstructure presented in Figure 1. It has a complicated microstructure, mixture of crystalline silicon (c-Si) grains, grain boundaries and/or amorphous-nano-crystalline hydrogenated silicon (a-nc-Si:H) often called "tissue". Approach followed in this paper is to compare on as wide as possible range of a-nc-µc-Si:H layer-samples the microstructure (grain size, crystallinity and roughness) with the optical and transport properties and to find overall performance. We have used the fact that a-nc-µc-Si:H microstructure changes with the thickness of sample [11].

AFM and SEM micrographs of some authors [10, 12] reveals surface morphology of micro-crystalline silicon films thickness as sample of 1.4 micrometers published in [12].

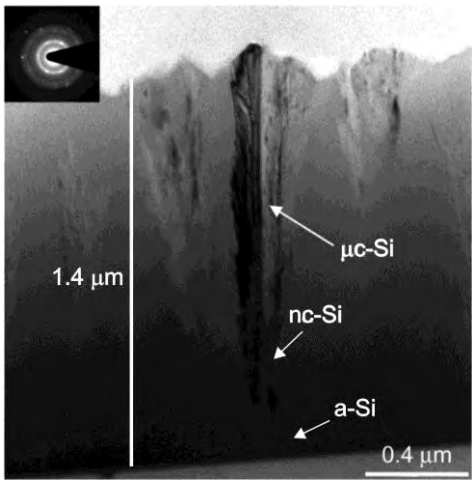


Figure 1. Typical SEM micrograph of composite silicon (a-nc-µc-Si:H) thin film 1.4 µm of thickness produced with Cat-CVD (Hot Wire CVD) technique, published in [12].

Large complexity of microstructure in hydrogenated microcrystalline silicon and existence of tissue with at least two different sizes of crystallites determine the optical properties and therefore modeled complex free carrier mobility based on mechanism of transport.

III. OPTICAL PROPERTIES

The optical properties of composite silicon thin films strongly depend on the production conditions and determination of the structural properties of the thin film. The spectra of absorption coefficient, $\alpha(E)$, of typical nc-Si thin film calculated from transmittance, Fourier transform photocurrent spectroscopy (FTPS) and photo deflection spectroscopy (PDS) are shown in Figure 2. Absorption coefficient data used in our calculation and device simulation was published in literature [3, 4]. Amongst others, sample number two (K2) is chosen in simulation modeling because of convenient value of its band gap energy (E_g=1.82eV) caused by lower crystalline fraction (vol. 30%).

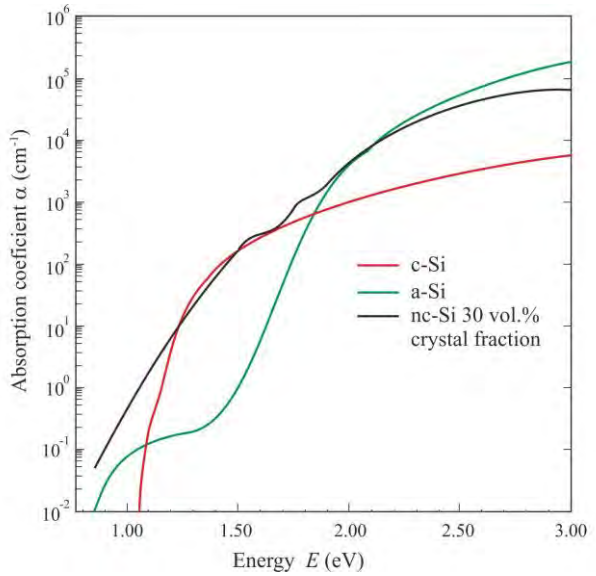


Figure 2. Distribution of absorption coefficient of nc-Si (black line), a-Si (green line) and c-Si (red line) calculated from FTPS and PDS [4].

IV. SIMULATION MODEL

Computer modeling software AMPS-1D can simulate all modeled semiconductor and photovoltaic device structures. In the present version of AMPS-1D user can chose one of two different calculation models: Density of States (DOS approach) and Carrier Lifetime Model (CLM). DOS approach is more suited for silicon amorphous and nano-crystalline thin films layers due to large defect densities in midgap states [13, 14].

Model of photovoltaic device can be simulated through optical and material parameters on device designed structure. Material and optical properties affect electrical parameters in tree differential equations in correlation: the Poisson's equation, electron continuity equation and hole continuity equation. Those tree equations are solved simultaneously under non-equilibrium steady-state conditions (i.e., under the effect of light, voltage bias or both) by using method of finite differences and Newton-Raphson technique. The used equations are:

Poisson's equation,

$$\frac{\rho(x)}{\varepsilon} = \frac{\partial \Psi(x)}{\partial x} \tag{1}$$

Electron continuity equation,

$$\frac{1}{q}\frac{\partial J_n(x)}{\partial x} = R(p(x), n(x)) - G_{op}(x) \tag{2}$$

And hole continuity equation,

$$\frac{1}{q}\frac{\partial J_p(x)}{\partial x} = G_{op}(x) - R(p(x), n(x)) \tag{3}$$

All carriers in semiconductor layer can be described by net charge density value $\rho(x)$, expressed like

$$\rho(x) = q[p(x) - n(x) + p_T(x) - n_T(x) + N_A - N_D] \quad (4)$$

And the present electrostatic field E is defined as

$$E = \frac{\partial \Psi(x)}{\partial x}. \quad (5)$$

In equation ε is the dielectric constant, E is electrostatic field, $\Psi(x)$ represents the position of energy in the local vacuum level, x is position in the device, n and p the extended states density in conduction and valence band, respectively, p_T and n_T the trapped hole and electron population density, N_A acceptor doping density, N_D the donor doping density, if exists, the q the electron charge, $R(x)$ the recombination rate, $G_{op}(x)$ the optical generation rate of free electron-hole pair, J_n and J_p the electron and hole current density, respectively. The term $R(x)$ is the net recombination rate resulting from band-to-band (BTB) direct recombination and Shockley-Read-Hall (SRH) indirect recombination traffic through gap states. The model used in AMPS for indirect recombination assumes that the traffic back and forth between the delocalized bands and the various types of delocalized gap states is controlled by SRH, capture and emission mechanisms. Since AMPS has the flexibility to analyze device structures which are under light bias (illumination) as well as voltage bias, the continuity equations include the term $G_{op}(x)$ which is the optical generation rate as a function of x due to externally imposed illumination.

Generally, the three state variables completely define the state of a device: local vacuum energy level Ψ and the quazi-Fermi levels E_{Fp} and E_{Fn}. Once those tree dependent variables are calculated as a function of position in device, all other parameter can be determinate as its function. In thermodynamic equilibrium, the Fermi level is a constant as a function of position and hence the tree equations (1-3) essentially reduce to Poisson's equation. Therefore, local vacuum energy level Ψ is the only variable to solve in thermodynamic equilibrium.

Otherwise, in non-thermodynamic equilibrium steady-state, a system of three correlated non-linear second order differential equations in the tree unknowns (Ψ, E_{Fp}, E_{Fn}) is obtained. Further calculation need six boundary conditions, two for each dependent variable. The first two boundary conditions are modified versions of the ones used for solving Poisson's equation in thermodynamic equilibrium:

$$\Psi(0) = 0 - \chi(L) - \varphi_{bL} + \chi(0) - V \quad (6)$$

And

$$\Psi(L) = 0 \quad (7)$$

Where L is the total length of the modeled device, $\chi(0)$, $\chi(L)$ are the electron affinities at x=0 and x=L, respectively and V is applied voltage. Zero value of local vacuum energy level Ψ is at boundary point x=L. The four other boundary conditions are obtained from imposing constraints on the currents at the boundaries at x=0 and x=L. These constraints force the mathematics to acknowledge the fact that the currents must cross at x=0 and x=L (contact position) by either thermionic emission or interface recombination. Mathematically expressed current values at boundaries:

$$J_n(0) = qS_{n0}[n(0) - n_0(0)] \quad (8)$$

$$J_p(0) = qS_{p0}[p(0) - p_0(0)] \quad (9)$$

$$J_n(L) = qS_{nL}[n(L) - n_0(L)] \quad (10)$$

$$J_p(0) = qS_{pL}[p(L) - p_0(L)] \quad (11)$$

Where S_{n0}, S_{p0} are surface recombination velocities for electrons and holes respectively at the x=0 interface and the quantities are the corresponding velocities at the x=L interface. Strongly limited by thermionic emission largest value of recombination velocities cannot over cross 10^7cms^{-1}. In equations (8-9) n(0) and p(0) are the electron and hole density at x=0, n(L) and p(L) are the same values at x=L. Analogy, $n_0(0)$ and $p_0(0)$, $n_0(L)$ and $p_0(L)$ are the electron and hole density at the thermodynamic equilibrium at boundaries x=0 and x=L, respectively. Now, when all conditions are defined, simultaneously calculation of Ψ, E_{Fp} and E_{Fn} can be obtained.

The model used in our simulation for gap states consist of two exponential tail states distributions and two Gaussian distributions of deep defects states [13].

V. SOLAR CELL PARAMETERS

In this paper we have simulated two types of single junction solar cells both with absorber layer thickness of 1200 nm. The first type have the standard *pin* structure with one homogenous absorber: (p-) a-Si:H/ (i-) a-Si:H/ (n-) a-Si:H and the second type with inhomogeneous absorber has a multi-layer structure with 9 intrinsic absorbers with various structural and therefore optical properties (Fig.2). Thickness of layers for standard *pin* structure is as follows: (p-) 8nm/ (i-) 1400nm/ (n-) 15nm.

We propose an implementation of chemically textured zinc-oxide ZnO:Al film or SnO_2 as a front TCO in our p-i-n solar cells, and in combination with Ag as a textured back reflector-enhancing dielectric layer for additional efficiency. Such modeled cells exhibit excellent optical and light-trapping properties demonstrated by high short-circuit current densities.

A. Boundary conditions

The tree governing equations (1), (2) and (3) must hold every position in the device and the solution to those equations involves determining the state variables $\Psi(x)$ $E_{Fp}(x)$ and $E_{Fn}(x)$. Non-linear and coupled equations cannot be solved analytically, numerical methods must be utilized. Boundary conditions must be imposed on the sets of equations. These are expressed in terms of conditions on the local vacuum level and the currents at the contacts. To be specific the solution to equations (1), (2) and (3) must satisfy the boundary conditions (6-11).

In computer modeling program AMPS-1D Ψ(0) is PHIB0, Ψ(L) is PHIBL, S is recombination speed for holes and electrons depending on carrier position. Parameter RF is reflection coefficient at x=0, RB is reflection coefficient at x=L. In actual situation RF is parameter of lost photon flux transmitting throe glass substrate and TCO layer. Measurements of solar cells internal optical loses (in visible spectrum) showing wavelength (λ) or energy dependence [5, 6, 15]. Reflection coefficient RB depends of optical characteristics of back electrode which acts as an optical mirror and efficient light trapping. The properties of the front contact and back contact used as the model parameters are shown in Table I.

Table I. Boundary conditions in AMPS-1D: PHIB0 is Ψ(0) at x=0; PHIBL is Ψ(L) at x=L (total device length); SN0-SNL-SP0-SPL are surface recombination speeds at x=0 and x=L (N=electron, P=hole); RF is reflection coefficient at x=0; RB is reflection coefficient at x=L.

Front Contact	Back Contact
PHIB0 = 1.730 eV	PHIBL = 0.120 eV
SN0 = $1x10^7$ cm/s	SNL = $1x10^7$ cm/s
SP0 = $1x10^7$ cm/s	SPL = $1x10^7$ cm/s
RF = 0.250	RB = 0.600

B. *Solar cells design*

Second type model of solar cell with multilayer absorbers will be detailed explained in this part. Solar cell design in fact represents a-Si:H/nc-Si:H/μc-Si:H inhomogeneous intrinsic absorber of *pin* silicon solar cell with thickness of 1200 nm (Fig.3).

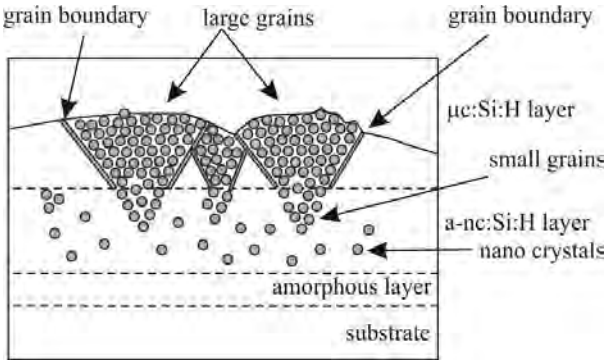


Figure 3. Drawing of composite silicon (a-nc-μc-Si:H) thin film cross section of 1.2 μm thickness; it consists of multi layers with different structure and absorption properties.

Design complies of multi-layer structure with 9 individual intrinsic absorbers arbitrary modeled thickness with various structural and optical properties (Fig.4). Solar cell model structure is as follows: Transparent Conductive Oxide (TCO) is solar cell front contact for collecting holes (0.3 μm), a-Si:H (p-1) window layer (8 nm), a-nc-Si:H absorbers i_1-2/i_4-5 (100nm), nc-Si:H absorber i_5-6/i_6-7 (100 nm), nc-μcSi:H absorber i_7-8/i_8-9 (200 nm), μcSi:H absorber i_9-10 (400 nm), a-Si:H (n-11) back layer (15 nm), Aluminum-doped Zinc Oxide (AZO) is reflection layer (0.1 μm), Ag as a textured back reflector (0.5-2 μm), Aluminum back contact electrode (2-3 mm).

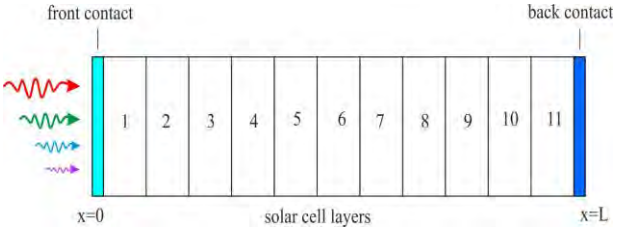


Figure 4. Model of solar cell with multi layer absorbers: (p-) is layer 1, (n-) is layer 11 and layers 2-10 are 9 individual intrinsic absorbers with various thicknesses, structural and optical properties [18].

The input parameters of all modeled solar cell layers in order to simulate efficiency properties of actual solar cell are not given in complete because of complexity. Published selected mobility parameters are taken from references [16, 17] and absorption coefficients from a-nc-Si:H and c-Si:H samples. We used standard boundary conditions and standard global illumination conditions (Table I), air-mass 1/cosΘ (AM1.5 spectrum), 1000W/m^2 at 300 K temperature reference.

C. *Absorption coefficients*

One set of absorption coefficient data we used in our calculation and device simulation was measured on samples and published in literature [3, 9]. It is coefficient for a-nc-Si:H sample and c-Si:H silicon layer. Amongst others, a-nc-Si:H sample is chosen in this modeling because of convenient value of its band gap energy (E_g=1.82eV) caused by lower crystalline fraction (vol. 30%), layer thickness (100 nm) and calculated DC conductivities and free carrier mobility [2]. Presented modeled device of solar cell (Fig.4) consists of 9 absorber layers with 2 actual absorption coefficients and 7 approximated absorption coefficients between a-nc-Si:H (front) and c-Si:H (back) silicon layer. Therefore, absorption coefficients of each individual absorber layer are modeled and proposed by linear approximation-superposition according to different microstructure (Fig.3) through solar cell variable length (parameter x). According to layer thickness or calculated length x each individual absorber layer has different superposition ratio of absorption coefficients between a-nc-Si:H and c-Si:H layers as it is in actual solar cell. For example: absorption layer 2 (i_1-) in modeled solar cell is a-nc-Si:H actual sample with measured absorption coefficient, absorption layer 3 (i_2-) is a-nc-Si:H modeled layer with proposed absorption coefficient in ratio a-nc-Si/c-Si (0,25/0,75), absorption layer 3 (i_2-) is a-nc-Si:H modeled layer with proposed absorption coefficient in ratio a-nc-Si/μc-Si (0,35/0,65), etc. Last layers are: absorption layer 9 (i_8-) in modeled solar cell is nc-μc-Si:H layer with proposed absorption coefficient in ratio nc-Si/μc-Si (0,85/0,15) and absorption layer 10 (i_9-) is μc-Si:H actual sample with measured absorption coefficient (Fig. 5).

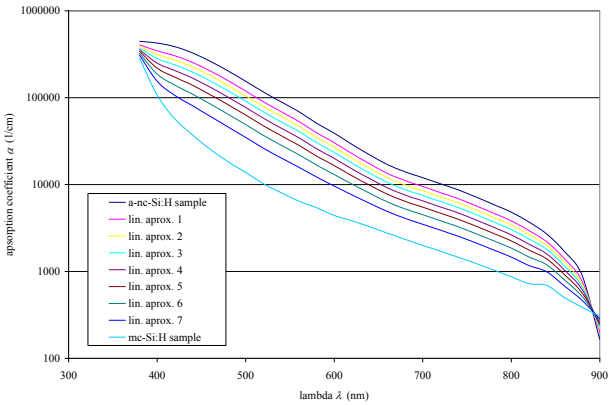


Figure 5. Absorption coefficients $\alpha(\lambda)$ of nine absorber layers in modeled solar cell [18].

D. *Free carriers mobility*

In mixed phases silicon thin film layers transport mechanism strongly depends of carrier mobility μ ($cm^2 V^{-1}s^{-1}$). Electron μ_n and hole μ_p mobility have dependence to crystal lattice temperature also donor-like and acceptor-like doped concentrations [18], defect density [11], DC conductivity [19], suggesting electron mobility at temperature of 300 K maximal values of 1250 $cm^2 V^{-1}s^{-1}$ and hole mobility maximal values of 400 $cm^2 V^{-1}s^{-1}$ in bulk (intrinsic) crystalline silicon. For electron mobility at temperature of 300 K in a-Si:H intrinsic silicon layers in simulation values are modeled as follows: (M_{UN}) 10-20 $cm^2V^{-1}s^{-1}$, (M_{UP}) 2-4 $cm^2V^{-1}s^{-1}$. For a-nc-Si:H thin film layers values are: (M_{UN}) 100-250 $cm^2V^{-1}s^{-1}$, (M_{UP}) 8-60 $cm^2V^{-1}s^{-1}$; for nc-Si:H layers (M_{UN}) 400-650 $cm^2V^{-1}s^{-1}$, (M_{UP}) 100-180 $cm^2V^{-1}s^{-1}$, for nc-Si:H/μc-Si:H layers (M_{UN}) 800-1000 $cm^2V^{-1}s^{-1}$, (M_{UP}) 200-300 $cm^2V^{-1}s^{-1}$; and for μc-Si:H layers (M_{UN}) 1200-1250 $cm^2V^{-1}s^{-1}$, (M_{UP}) 300-400 $cm^2V^{-1}s^{-1}$.

VI SIMULATION RESULTS AND DISCUSION

The performance of modeled solar cells was analyzed in respect to the current density (J_{SC}) and efficiency (η) by incorporating the layer parameters into AMPS- 1D. First *pin* structure consists of homogenous intrinsic a-nc-Si:H absorber in thickness of 1200 nm with constant crystal fraction of Xc=30%. In second *pin* structure with the same thickness we modeled an inhomogeneous intrinsic absorber which consists of 9 individual homogenous layers with different optical and electrical properties suggesting experimentally proven structure in-homogeneity. In standard simulation conditions defined earlier for the first structure simulation results shown predictable curves and maximal value of J_{SC}=17.421mA/cm^2 at 1200 nm of absorber thickness and maximal efficiency of η=13.992% at 935nm (Fig.6.). In second modeled structure calculated values are different: J_{SC}=14.781mA/cm^2 at 1200 nm of absorber thickness and maximal solar cell efficiency of η=12.32% at 492 nm (Fig.7.). For both type of modeled devices simulation promotes typically exponentional rise of current density in the first 300 nm of thickness according to excellent photon absorption and collection of photo-generated electron-hole pairs in absorber structure with 10-100 ns free carrier life time in first few hundred nanometers of thickness of a-nc-Si:H tissue. Current densities in our simulation had never reached its calculated maximum in both cases. In first solar cell model simulation improvement in efficiency is supported by at least one order of magnitude better absorption in a-nc-Si:H homogenous absorber, good conductivity and therefore high free carrier mobility. Also, efficiency curve points to saturation at 600 nm and its decrease at thicknesses higher then 900 nm. Calculated solar cell performances in first model simulation are likely expected according to physical nature of photo-generation and recombination of electron-hole pairs in intrinsic silicon with controlled general silicon layer parameters as doping and free carrier concentrations, mobility and gap state defect distribution.

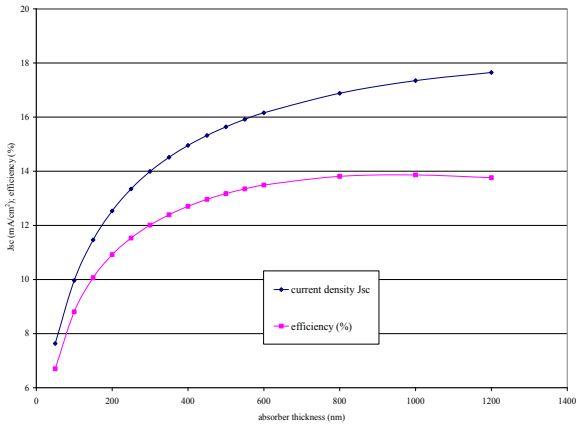


Figure 6. Graphical presentation of calculated solar cell current density J_{SC} (mA/cm^2) and efficiency η (%) in case of homogeneous absorber with crystal fraction of 30%.

Second *pin* model solar cell design implies experimentally proven structure in-homogeneity of silicon CVD thin films. Absorption coefficient decreases drastically through structure instead of increasing of free carrier mobility according to tissue structure changes. As a result of different optical and electrical properties in the structure layers of the in-homogenous solar cell calculated performance is expectable. Optimum efficiency of η=12.32% is reached at 492 nm (Fig.7.) and current density of J_{SC}=12.21 mA/cm^2 exists at the same absorber thickness.

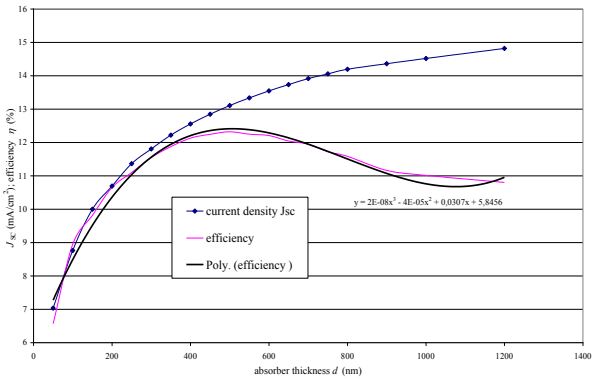


Figure 7. Graphical presentation of calculated solar cell current density J_{SC} (mA/cm^2) and efficiency η (%) in case of inhomogeneous absorber presented with 9 different absorber layers.

In Figure 7 additional curve (black line) represents third degree Polynomial of efficiency curve suggesting efficiency calculation in any absorber thickness (absorber dimension *x*).

VII CONCLUSION

In this study we have simulated two types of single junction solar cells with absorbers of a-nc-Si:H and a-Si:H/nc-Si:H/μc-Si:H tissues. A series of simulations were carried out in order to calculate the efficiency of modeled solar cell by varying the properties each of layers in the range published in the literature. The obtained results show clearly that the cells reach its optimum efficiency at different thicknesses. In case of homogenous silicon a-nc-Si:H tissue expected efficiency is around 14% and this value is in correlation with crystal fraction, absorption coefficient and free carrier mobility, respectively. In case of in-homogenous a-Si:H/nc-Si:H/μc-Si:H absorber expected efficiency is around 12% and lower strongly depending of structure characteristic.

The observed silicon layers specificity in the optical and electrical properties can be explained as a consequence of thin film deposition techniques forming regions of nano and micro crystals with arbitrary concentrations in amorphous matrix to determine free carrier transport model.

However, the goal of this work is quite ambitious in its choice to enable a preventive efficiency calculation of composite silicon solar cells before its production if deposition techniques are known, constant and stabile. Building of data matrix concerning arbitrary concentrations of crystals and thickness of deposition layers it could be possible to predict solar cell efficiency using for example a principal method and third degree Polynomial approximation such as the one derived here.

REFERENCES

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

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

[3] J. Sancho-Parramon, D. Gracin, M. Modreanu, A. Gajovic, "Optical spectroscopy study of nc-Si-based p-i-n solar cells", Solar Energy Materials & Solar Cells 93, 2009, pp. 1768-1772.

[4] D. Gracin, A. Gajović, K. Juraić, J. Sancho-Parramon, M. Čeh: "Correlating Raman-spectroscopy and high-resolution transmission-electron-microscopy studies of amorphous-nanocrystalline multilayered silicon thin films"; Thin Solid Films 517, 2009, vol. 18, pp. 5453-5458.

[5] A. M. Ali, "Origin of photoluminescence in nanocrystalline Si:H films", Journal of Luminescence, 2007, vol. 126, pp. 614-622.

[6] A. V. Shah, J. Meier, E.Vallat-Sauvain, N. Wyrsch. U. Kroll, C. Droz, U. Graf, "Material and solar cell research in microcrystalline silicon", Solar Energy Materials & Solar Cells, 2003, vol. 78, pp. 469-491.

[7] A. M. Ali, "Optical properties of nanocrystalline silicon films deposited by plasma-enhanced chemical vapor deposition", Optical Materials, 2007, vol. 30, pp. 238-243.

[8] S. Hazra, S. Ray, "Nanocrystalline silicon as intrinsic layer in thin film solar cells, Solid State Commun., 1999, vol. 109, pp. 125-128.

[9] D. Gracin, A. Gajović, K. Juraić, J. Sancho-Parramon, M. Čeh: "Correlating Raman-spectroscopy and high-resolution transmission-electron-microscopy studies of amorphous-nanocrystalline multilayered silicon thin films", Thin Solid Films, 2009, vol. 517, Is. 18, pp. 5453-5458.

[10] Kočka, J., Stuchlíkova, H., Stuchlík, J., Rezek, B., Mates, T., Švrcek, V., Fojtík, P., Pelant, I., Fejfar, A., "Microcrystalline silicon - relation of transport and microstructure", Solid State Phenomena, 2001, Vol. 80-81, pp. 213-224.

[11] Kočka, J., Stuchlíkova, H., Stuchlík, J., Rezek, B., Mates, T., Švrcek, V., Fojtík, P., Pelant, I., Fejfar, A., Model od transport in Micro-crystalline silicon", Journal of Non-Crystalline Solids, 2002, Vol. 299-302, pp. 355-359.

[12] Moutinho, H.R., Jiang, C.-S., Perkins, J., Xu, Y., Nelson, B.P., Jones, K.M., Romero, M.J., Al-Jassim M.M., "Effects of dilution ratio and seed layer on the crystallinity of microcrystalline silicon thin films deposited by hot-wire chemical vapor deposition", Thin Solid Films, 2003, Vol. 430, Issues 1–2, pp. 135–140.

[13] A. Belfar, R. Mostefaoui, "Simulation of n1-p2 Microcrystalline Silicon Tunnel Junction with AMPS-1D in a-SiC:H/μc-Si:H Tandem Solar Cell, Journal of Applied Science, 2011, pp. 10.3923.

[14] S. Tripati, R. O. Dusane, "AMPS-1D simulation studies of electron transport in mc-SI:H thin films", Journal of Non-Crystalline Solids, 2006, vol. 352, pp. 1105-1108.

[15] R. H. Franken, R. L. Stolk, H. Li, C. H. M. Van der Werf, J. K. Rath, R. E. I. Schropp, "Understanding light trapping by light scattering textured back electrodes in thin film n-i-p-type silicon solar cells", Journal of Applied Physics, 2007, vol. 102, pp. 014503.

[16] D. Stieler, V. D. Dalal, K. Muthukrishnan, M. Noack, E. Schares, "Electron mobility in nanocrystalline devices", Journal of Applied Physics, 2006, vol. 100, doi: 10.1063/1.2234545.

[17] B. Van Zeghbroeck, "Mobility Carrier Transport, Principles of semiconductor devices", ECEE University of Colorado, 2011.

[18] V. Tudić, "Modeling of Electric characteristics of the photovoltaic amorphous-nanocrystalline silicon cell", doctoral thesis, FER Zagreb, 2014, 245 pp.

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

The investigation of influence of localized states on a-Si:H p-i-n photodiode transient response to blue light impulse with blue light optical bias

Marko Čović[#], Vera Gradišnik[#] and Željko Jeričević[*]
[#]Engineering Faculty/Department of Electrical Engineering, Rijeka, Croatia
[*]Engineering Faculty/Department of Computer Engineering, Rijeka, Croatia
zeljko.jericevic@riteh.hr

Abstract - The series of experiments measuring the transient response of a-Si:H pin photodiode to light impulses superimposed to constant light (optical bias dependence of modulated photocurrent method - OBMPC) of the same wavelength (430 nm) and various reverse voltages on photodiode was performed in order to characterize localized states of the energy gap of amorphous silicon and their influence on photocurrent degradation. The responses were analyzed as a sum of decaying exponential functions using the least squares method and a generalized Fosse's algorithm. This type of response is typical for independent relaxation processes running at the same time. Experiments and subsequent data processing illustrate feasibility of the method and results for the transient response of a-Si:H pin photodiode. The results strongly suggest two energy levels between 0.32 eV and 0.45 eV. These results were obtained applying the optical ac blue and dc blue bias light in a low frequency regime.

I. INTRODUCTION

Exponential decay is typical for a single relaxation processes in physics and first order chemical reactions in chemical kinetics, as well as in biology. In more complex situations where few independent processes of this type are going on in parallel, the summary measurements from the system consist of the sum of exponential functions. For example, in a mixture of radionuclides, each one decays independently and with a different rate controlled by its half life. Although the separation of exponentials from summary signal looks deceptively simple, it is actually a tough numerical problem because of the non-orthogonality of exponentials. Attempts to separate exponentials with close half times iteratively by nonlinear least squares usually results in large number of iterations and no convergence. For the analysis reported here we used the least squares method with a linearization step based on numerical integration. After the linearization, the solution of the multi-exponential problem is obtained by solving an over-determined system of linear equations followed by finding the roots of polynomial. The number of exponentials in the signal dictates the degree of polynomial, rank of the linear system, and multiplicity of numerical integration. The advantage of accurate linearization is that the separation of exponentials becomes a noniterative procedure and the condition number of the linear system can be used to control the quality of solution. The procedure is completely generalized and initial guesses are not necessary. It is also simple to implement non-negativity constrains on a solution.

The complex nature of localized states, such as native and metastable defects in a-Si:H [1], has an influence on a-Si:H p-i-n photodiode transient response. Other authors used the multiexponential trapping rate and modulated photocurrent (MPC) technique [2] to determine parameters of localized states throughout the entire energy gap by employing frequency and temperature scans. We examined the nature and the kinetics of light-induced defects creation in a-Si:H films and photodiode and their influence on photocurrent degradation. We measured and analyzed the transient response of a-Si:H p-i-n photodiode to blue light impulse superimposed to the blue light optical bias (optical bias dependence of modulated photocurrent method – OBMPC [2, 3]) at various reverse bias voltages and one frequency. By means of OBMPC, the trap and recombination localized states parameters throughout the entire energy gap can be identified. To the low-frequency MPC data the deeper recombination centers also contribute [2]. The purpose of this work is to identify the nature and role of trapping and recombination process of mobile carriers. This was done under condition of bias and modulated blue light of weak illumination intensity, at low modulation frequency and at applied low reverse bias voltages on a-Si:H p-i-n photodiode. In this regime, the localized states with deeper energy levels in low frequency regime can be identified. Based on experimentally obtained results, the photodiode transient responses are reconstructed with the help of numerical modeling.

Under described experimental conditions, the measured transient responses show the presence of one or two decaying exponential functions corresponding to two energy levels between 0.32 eV and 0.45 eV.

II. THEORY AND RESULTS

The basic idea of the computing method was first proposed by Foss [4], and was later used by Matheson [5] in chemistry and Jericevic [6] in biology. In all above mentioned papers the method was developed and applied to special cases, no general solution was developed,

although Foss claimed it is possible to construct one. The general solution was finally developed by Jericevic [7], and it does represent the foundation on which here described data processing is based.

The approach outlined here offers unprecedented flexibility addressing important problems of accurate and fast (real time) processing for multiple models (for an arbitrary number of exponential terms) using the same computing methodology. Previous implementations required that specialized subroutines for each individual case (mono-exponential, bi-exponential, tri-exponential, etc) have to be written.

The major steps in out methodology are:

1. Linearization by numerical integration method
2. Solution of linear system of equations with or without non-negativity constrains
3. Determining coefficients and roots of polynomial equation(s) based on linear system solutions.
4. Decay constants computed in the previous step are used to compute pre-exponential terms.
5. Testing and verifying the results by detailed error analysis.

The detailed development of a general solution is presented in [7] and here we are giving a brief description of the final result only:

General solution for $y = \sum_{i=1}^{N} A_i e^{-k_i t}$

$$p_1 = -\sum_{i=1}^{N} k_i \qquad (1)$$

$$p_n = -\sum_{i=1}^{\frac{N!}{(N-n)!n!}} \prod_{1}^{n} k_m \qquad m \in ({}^{N}C_n) \qquad (2)$$

$$p_N = -\prod_{i=1}^{N} k_i \qquad (3)$$

$$p_{N+1} = \sum_{i=1}^{N} A_i \qquad (4)$$

$$p_{N+2} = \sum_{i=1}^{i=N} A_i \sum_{j=1; j\neq i}^{N} k_j \qquad (5)$$

$$p_{N+n} = \frac{1}{(n-1)!}\sum_{i=1}^{N} A_i \sum_{j=1;}^{\frac{(N-1)!}{(N-n)!(n-1)!}} \prod_{1}^{n-1} k_m \qquad (6)$$

$$m \in ({}^{N-1}C_{n-1};\ m \neq i)$$

$$p_{2N} = \frac{1}{(N-1)!}\sum_{i=1}^{N} A_i \prod_{j=1;\, j\neq i}^{N} k_j \qquad (7)$$

Where vector of parameters p is a solution of linear system based on consecutive N-fold numerical integration of multi-exponential equation in time t (N is the number of exponential terms in the summary signal y). From the first N polynomial equations in k (vector of decay constants) is calculated. This set of polynomial equations are presented in brief as (1) to (3). After vector k is known, the computing of parameter vector A (vector of population of states) from the last N equations (4) to (7) becomes a linear problem. The details and complete development of the solution is in [7]. Our rationale for using the general multi-exponential solution is that we did not postulate the number of components (N) in advance. Instead, we let the data determine N by fitting with an increasing number of exponentials and subsequently used the best fit for photodiode characterization.

The a-Si:H p-i-n photodiode, due to the presence of localized states in the energy gap, shows the multi-exponential decay in transient response on light pulse when the light is taken off. The number of components included in response was not known in advance.

In our experiments, the a-Si:H p-i-n photodiode is illuminated with monochromatic blue LED (Kinghbright FULL COLOR RGB LAMPS, 430 nm, I_F=20 mA) light consisting of a constant (bias) and pulsed (probe) light beams of the same relatively weak intensity and low frequency. Consequently, the electron-hole pair photogeneration rate contributes to a constant and a pulsed (transient) mobile carrier density component. For blue light pulse and bias illumination, electron-hole pairs are generated near the front surface. Our sample of a-Si:H p-i-n is illuminated from p-type layer electrons. We assume that the majority carriers are in i-type layer. Due to trapping and release interaction of free carriers with band gap localized states, the resulting transient photocurrent will have time delay with respect to the light excitation. Also, the photocurrent decay will happen long before the photocurrent reduction which is due to free carrier recombination. The carrier from shallow trap are reemitted soon after capture, but the carriers in a deep trap reside longer, are practically lost, and are the major cause for the transient photcurrent decay (base-line current tail).

The transition of mobile carriers, detected in the experiment, relies only on the exchanges between localized gap states and extended states. The hole, at blue light absorption, move directly into the front contact and their contribution to the transient photocurrent is small [3]. The dc illumination determines the position of quasi-Fermi level for electrons, E_{tn} and holes, E_{tp} and can be deduced from the measurement of the dc photocurrent from [8, 9]

$$E_C - E_{tn} = kT \cdot \ln(\mu_n N_C A q \xi / I_{phdc}) \qquad (8)$$

where is the mobility μ_p = 10 $cm^2V^{-1}s^{-1}$, the effective density of states N_c = 10^{20} cm^{-3}, ξ is bias voltage dependent electric field, and attempt-to-escape frequency ν_0 =1·10^{12} s^{-1}. The time response characteristic for dc part

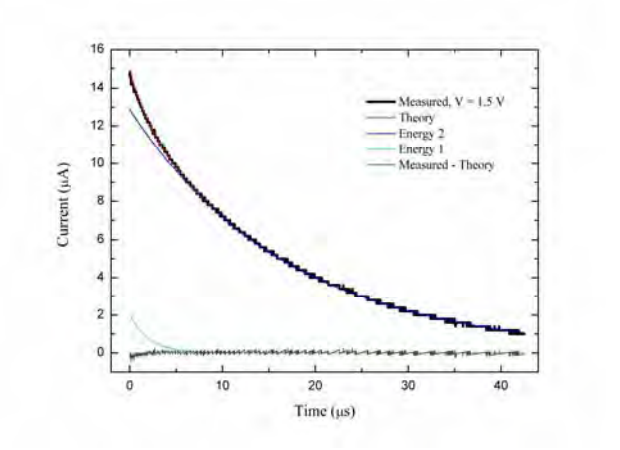


Figure 1. Measured and calculated switch-off transient response of a-Si:H p-i-n photodiode on blue light pulse at blue bias light at 1.5 V bias reverse voltage. Measured is a summary signal (experimental), Theory is fitted function, Energies are fitted components and Measured – Theory is a difference between experimental measurements and fitted function.

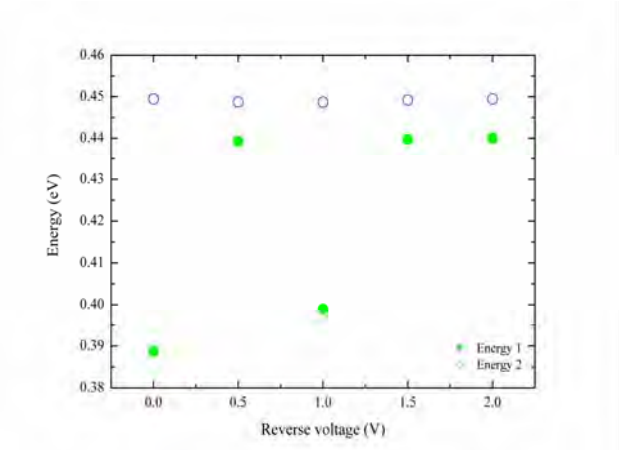


Figure 2. Energies of two energy levels obtained from measured transient responses of a-Si:H p-i-n photodiode by blue light pulse at blue bias light at reverse bias voltages from 0 V to 2 V.

of generation rate of the gap state at the energy E is given by [9] $1/\tau(E) = c_n n_{dc} + c_p p_{dc} + e_n(E) + e_p(E)$, where n_{dc}, p_{dc} are free electron and hole density, c_n, c_p capture coefficients of electrons, holes, $e_n(E)$, $e_p(E)$ emission frequency toward the conduction, valence band. The occupation function of the gap states f_{dc} at constant illumination change from 1 to 0 in two steps which occurring at two quasi-Fermi levels of trapped carriers, E_{tn} and E_{tp}, at which the emission of electron and hole are equal to characteristic capture frequency, ω_c. The energy gap divide, dependently on localized states energy positon, two quasi-Fermi levels, in electron trapping states ($E > E_{tn}$), hole trapping states ($E < E_{tp}$) and recombination states ($E_{tp} < E < E_{tn}$).

At bias voltage V_i, the characteristic frequency, ω_c^i, which is the capture rate of electrons and holes into each type or probed gap state is given by [2]

$$\omega_c^i = n_{dc} c_n^i + p_{dc} c_p^i \qquad (9)$$

where n_{dc} (p_{dc}) are free carrier electron (hole) density and $c_n^i (c_p^i)$ the capture coefficient for an electron (hole) changes. The applied reverse bias voltage scan provides the spectroscopy of the gap states instead of temperature scan.

To characterize the transient behavior two other energies have to be introduced: the characteristic time response of the gap states and the characteristic time, which is the period of the ac signal.

In the MPC (Modulated PhotoCurrent method) experiment the ac behavior is characterized with two other energies, comparing the inverse of characteristic time response of the localized gap states $1/\tau(E)$ and the angular frequency ω of the ac signal. If $\omega<\omega_c$ than $1/\tau(E)>\omega_c>\omega$. The distribution energy described by relation from [2, 8]

$$E_C - E_{\omega n} = kT \ln(\nu_0 / \omega) \qquad (10)$$

is close to E_{tn}^i, $E_{\omega n}^i \approx E_{tn}^i$. In region where $\omega << \omega_c^i$ and ω are comparable to ω_c^i is the so called low frequency regime.

All the states have characteristic response time shorter than the period of the ac signal. The phase shift is low and is not induced by trapping-and-release events of the MPC. The energies fall in recombination centers and the information on gap states at E_{tp} and E_{tn} is done.
In the low frequency (LF) regime the DOS (Density Of States) at the quasi Fermi level of trapped carriers is as we propose the modified expression from one done in [9]

$$N(E_{tn}) \approx \frac{G_{dc}}{k_B T} \frac{t_d}{(T_p/2)} \qquad (11)$$

where the delay time t_d and pulse period T_p, G_{dc} the dc generation rate, k_B Boltzmann constant, T temperature.

The transient photocurrent decay is due to transit time, which is voltage dependent. Due to localized states the transit time has two component, corresponding to two energies, and expressed [3] as

$$t_t = t_0 + \sum_i N(E_i)\sigma v_{th} \left(\frac{d}{\mu_n \xi} \right) \nu_i^{-1} e^{(E_i/kT)} \qquad (12)$$

where time that electron spend in conduction band is t_0, thermal velocity v_{th}, electron capture cross section σ, discrete localized states $N(E_i)$ at E_i energy levels, electric field ξ, electron mobility μ_n, intrinsic layer width d.

The a-Si:H p-i-n photodiode structure used in our experiment is well described in [10] and their transient response on light pulse of blue, green and red light at 2V reverse bias in [11].

The p-i-n structure was deposited on a transparent conductive oxide (TCO) coated glass from undiluted SiH4 by plasma-enhanced CVD and is as follows: glass/TCO/*p*-type (5 nm)/*i*-type (300 nm)/*n*-type (5 nm)/Al back contact as described in [10]. The n-type layer was made by adding phosphine and the p-type by adding diborane to the gas mixture. The back contact was aluminum deposited by the evaporation. The active surface area of the photodiode was 0.81 cm^2. Photo-illumination was

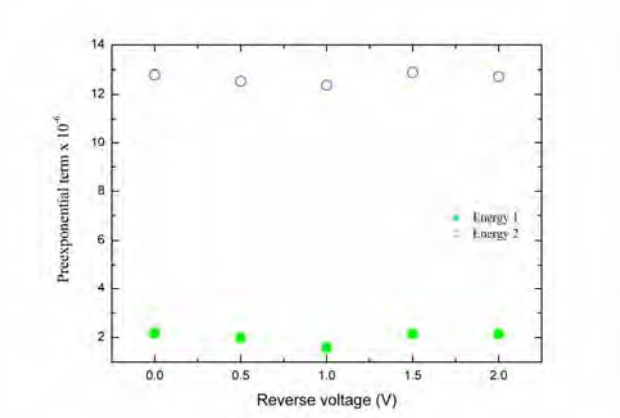


Figure 3. Preexponential terms obtained from measured transient responses of a-Si:H p-i-n photodiode by blue light pulse at blue bias light at reverse bias voltages from 0 V to 2 V.

obtained through the bottom p-type layer. The transient response of a-Si:H photodiode was measured as a response to the light pulses superimposed to the constant light (optical bias dependence of modulated photocurrent method - OBMPC) of the same wavelength of 430 nm and at reverse voltages from 0V to 2V.

Samples were measured at the room temperature. Photocurrent was measured directly on the 10 kΩ load resistor. The two blue (B) light LED for the probe (ac) and the pump (dc bias) from Multicolor LED lamp were used in the experiments, emitting at 430 nm. The emitted photons are of energies higher than the band gap energy. The optical light powers were defined at 20 mA LED bias current and the probe pulse period 3 ms with 50% duty cycle.

The measured a-Si:H photodiode switch-off transient response on a 3 ms light pulses of blue light at bias blue light and at 1.5 V reverse bias is shown in Fig. 1. The responses at bias voltages of 0 V, 0.5 V, 1 V, 1.5 V and 2V were analyzed as a sum of decaying exponential functions using the least squares method and a generalized Fosse's algorithm as described above. From Fig.1 it is evident that for 1.5 V reverse bias voltage, the states with shallower energy (E_1 = 0.4397 eV) and corresponding preexponential term $2.133 \cdot 10^{-6}$ have a small contribution to total photocurrent. The deeper energy state (E_2 = 0.4492 eV) with a $1.2886 \cdot 10^{-5}$ preexponentail term prevails in the transient response. It can be concluded that second energy states behave as deep acceptor localized states. The calculated value of DOS is $5 \cdot 10^{14}$ $cm^{-3}eV^{-1}$ centered at E_C-E_i=0.45 eV with E_C-E_{tn}=0.66 eV and E_C-$E_{\omega n}$=0.5 eV.

The calculated energies of localized energy levels are presented in Fig. 2. The summary measurements from the system consist of the sum of two exponential functions in all the cases, as shown in Fig. 1. The pre-exponential factor yields the information about presented species of localized states is shown in Fig 3. The preexponentail term of the deeper gap states and those nearer midgap states was higher than those of the shallower energy levels. These results agree with those obtained by oher authors [2], where the capture coefficients of the gap states closer to the midgap were higher than those of the shallow energy levels.

III. CONCLUSION

Our experiments indicate that the used frequency of modulated light is lower then the critical frequency and therefore fall in a low frequency regime. The experiments are done using the same light intensity for biased and modulated light. Under those conditions the trap centers act as a recombination centers and influence the a-si:H p-i-n photodiode transient response.

The results of OBMPC experiment have been collected and analyzed in time domain to characterize the behavior of a-Si:H p-i-n photodiode. For the selected light pulse frequency and the low bias voltages on photodiode, the LF regime can be used to determine DOS values and energy levels. Our results show the presence of two energy levels and their influence on a-Si:H p-i-n transient response on blue light pulses at bias blue light. The preexponentail term of the deeper gap states and those nearer midgap states was higher than those of the shallower energy levels. These results agree with those obtained by other authors [2], where the capture coefficients of the gap states near midgap found higher than those of the shallow energy levels.

ACKNOWLEDGMENT

We thank the referees whose comments improved the paper and DKJ for helping with English

REFERENCES

[1] Melskens, J. et al., *IEEE J. Photovolt.*, 6 (2014) 1331-1336 Year: 2014, Volume: 4, Issue: 6 Pages: 1331 - 1336, DOI: 10.1109/JPHOTOV.2014.2349655

[2] Pomoni, M., Kounavis, P., *Phil. Mag.*, 94:21, (2015) 2447-2471.

[3] Shen, D.S., Wagner, S. , J. Appl. Phys. 79, (1996) 794 – 801.

[4] Foss, S.D., Biometrics, 26 (1970) 815-821.

[5] Matheson, I.B.C., Anal. Instr., 16 (1987) 345-373.

[6] Jericevic, Z. et al., Adv. Cell Biol., 3 (1990) 111-151.

[7] Jericevic, Z. "Method for Fitting a Sum of Exponentials to Experimental Data by Linearization Using Numerical Integration Approximation, and Its Application to Well Log Data", USP #7,088,097, 2006.

[8] Kounavis, P. , J. Appl.Phys., vol. 77, (1995) 3872-3878.

[9] Kleider, J.-P. et al., *Phys. Stat. Sol.C,* 5 (2004) 1208-1226.

[10] Gradisnik, V. et al., *IEEE Trans. Electron Devices,* 49, (2002) 550 - 556.

[11] Gradisnik, V et al., *IEEE Trans. Electron Devices,* 53, (2006) 2485 – 2491.

Analysis of Electrical and Optical Characteristics of InP/InGaAs Avalanche Photodiodes in Linear Regime by a New Simulation Environment

Tihomir Knežević and Tomislav Suligoj
University of Zagreb, Faculty of Electrical Engineering and Computing, Micro and Nano Electronics Laboratory, Croatia
tihomir.knezevic@fer.hr

Abstract - The linear characteristics of the InP/InGaAs avalanche detectors are modeled and numerically analyzed by developing a new TCAD-based simulation environment. Temperature dependency of the impact ionization coefficients in InP are fitted for 200 K to 300 K temperature range. Adjustment of the model parameters for the simulations of the dark current sources in InP and InGaAs materials is performed in the same temperature range. Optical constants of the InGaAs material used in the layer stack are fitted to account for the absorption in the material for a range of wavelengths between 0.9 and 1.7 μm. Dark current and *I-V* characteristics under illumination are simulated and analyzed. Impact of the operating temperature on responsivity, breakdown voltage and dark current are analyzed. Excess noise factor is also calculated. Process simulations of Zn diffusion into InP are included in the TCAD simulator and the impact of the real diffusion profiles on the diode characteristics are assessed. The dark current for the structure with diffused Zn p^+ region decreases by a factor of 1.7 compared to the structure with box-like constant concentration p^+ region extracted at operating temperature of 200 K at 90% of V_{BR}.

I. INTRODUCTION

Low-light detection in the near-infrared range is commonly achieved by InP/InGaAs heterostructures employed in separate absorption, grading, charge and multiplication (SAGCM) avalanche photodiodes (APD). Avalanche photodiodes can be operated in linear or "Geiger" mode. In comparison to the standard *pin* photodiodes, InP/InGaAs APDs operated in linear mode have a higher sensitivity which promotes them to a device of choice for the optical communication systems [1]. Photodiode detectors operated in "Geiger" mode, above breakdown voltage, are called single-photon avalanche diodes (SPADs). Single-photon detection in the wavelength range above 1 μm is important for quantum cryptography [2], eye-safe laser ranging (Light Detection And Ranging – LIDAR) [3], time-resolved spectroscopy, photon-counting optical communication [4], etc. In "Geiger" mode, photogenerated carriers can produce self-sustaining avalanche. Avalanche is stopped by using a quenching circuit [5] and a SPAD is ready to detect a new photon.

In SAGCM structures, InGaAs ($In_{0.53}Ga_{0.47}As$) is used as an absorbing region and its lattice is matched to the one of InP. Bandgap of the InGaAs layer is 0.75 eV at room temperature and the layer is used for detection of light with wavelengths in range between 0.9 and 1.7 μm. SAGCM structure provides a way to limit the dark current due to the tunneling in a narrow bandgap InGaAs layer. High-field multiplication region is located in the low-doped InP layer while the charge layer limits the spread of the electric field to the absorption region. Reduced tunneling and impact ionization in the InGaAs layer contribute to the decreased dark current and improved overall performance of the photodetector.

Thermally generated carriers, tunneling currents and background photons all give rise to dark current, which decreases the sensitivity of the APD operated in the linear mode. The same is true for SPADs where the unwanted carriers can trigger an avalanche. This introduces noise in the operation of the SPAD called Dark Count Rate (DCR). Device structure e.g. layer stack thicknesses and doping concentrations can impact the performance of the APDs by changing the magnitude of the dark current coming from trap-assisted tunneling (TAT) [6]-[8]. On the other hand, thermally generated dark current is commonly reduced by decreasing the operating temperature of the InP/InGaAs APDs.

The key element in designing a high performance APDs is the ability to predict the device behavior for different structure parameters. There are simulator environments and analytical models capable of analyzing and simulating optical and electrical characteristics of the InP/InGaAs photodiodes [6], [9]-[13]. However, none of them exploits the functionality of a TCAD software to model the linear InP/InGaAs APD characteristics in the 200 to 300 K temperature range for realistic p^+ region doping profiles.

A new simulation environment using Sentaurus TCAD is developed. This TCAD-based environment is capable of simulating electrical and optical characteristics of the InP/InGaAs APDs. In this paper, fitted TCAD models enable the simulations of the avalanche generation in InP, thermal generation in InP, InGaAs and InGaAsP and trap-assisted tunneling in InP. Process simulations of Zn diffusion into InP are also implemented in the simulation environment. Optical and electrical simulations of the active part of the structure proposed by Liu et al. [6] are performed for temperatures of 200 K and 300 K. Excess noise factor is simulated for the same structure. Impact of

the realistic p^+ region diffusion profiles on the electrical characteristics is also assessed.

II. Device Simulations

A. *Fitting of the physical model parameters for InP/InGaAs APD device simulations*

Commercially available Sentaurus TCAD software from Synopsys is used for the simulations of the InP/InGaAs structure. The TCAD software is capable of performing both device and process simulations. Device simulator [14] can be used for electrical, optical and thermal simulations of the user-defined geometry with different materials. However, simulator provides full functionality of all the models for Si materials. In order to use the physical models for simulations of InP, InGaAs and InGaAsP materials for low temperatures, the physical model parameters should be fitted and properly tested.

Doping profile of the analyzed InP/InGaAs APD structure is shown in Fig. 1 (a). The parameters of the structure such as doping concentrations and layer thicknesses are almost identical as those in [6]. Buffer layer with thickness of 2 µm is defined on top of the bulk n^+ InP region. Buffer layer is followed by the InGaAs absorption region with thickness of 3 µm. Between the InGaAs layer and InP there is a InGaAsP grading layer, which serves to improve the transient characteristics of the device. Field stop or charge region with thickness of 1.2 µm and doping concentration of $2.1 \cdot 10^{16}$ cm^{-3} is used to limit the spread of the electric field into the InGaAs layer. Low-doped n^- InP layer is defined on top of the charge layer. P^+ region is defined as a highly doped box-like p^+ region with the constant doping concentration of $2 \cdot 10^{18}$ cm^{-3}, in the first approximation. Multiplication region is defined to be 0.5 µm thick. Band diagram of the structure at 0 V, 300 K is shown in Fig. 1 (b). The difference in the bandgap of the InP and InGaAs materials causes valence and conduction band discontinuities. Valence band discontinuity can limit the speed of the device since the generated holes have to overcome the potential barrier. Grading layer decreases the valence discontinuity increasing the response speed of the device.

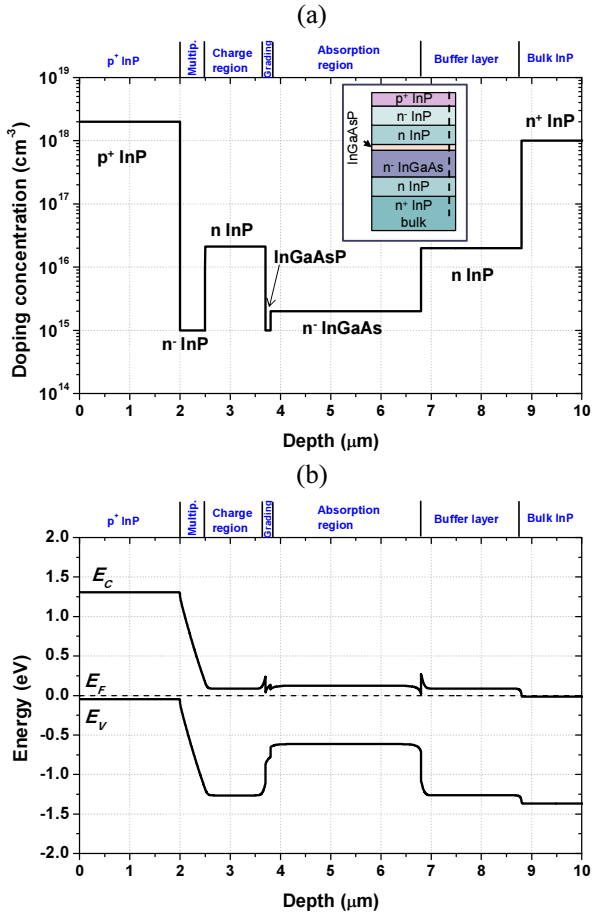


Figure 1. Cross-section of the simulated structure. (a) Doping profile of the SAGCM InP/InGaAs APD. (b) Band diagram of the structure at 0 V; 300 K.

In order to obtain the current-voltage characteristics of the InP/InGaAs diode, the impact ionization coefficients for InP material must be fitted first. TCAD simulator provides different models capable of modeling impact ionization coefficients in wide electric field and temperature range. However, the temperature dependency of the used Okuto-Crowell model for impact ionization in InP could not be obtained just by fitting the appropriate coefficients. Therefore, the parameters for the electron and hole impact ionization coefficients a, b and δ are fitted in the different temperature steps to the impact ionization values from the analytical model from [10]. In [10] the analytical expression was constructed to obtain best fit to the measured data and represents a quasi-physical model. The comparison of the fitted impact ionization coefficients for holes and electrons to the values obtained by the analytical model for temperatures of 200 K and 300 K is plotted in Fig. 2. Excellent fit to the experimental data can be achieved. Using the fitted impact ionization coefficients, the current-voltage characteristics of the structure proposed in Fig. 1 are simulated. Extracted breakdown voltage (V_{BR}) at 300 K is 79 V. Measured V_{BR} for the same structure is 75 V [6] which is in good agreement to the simulations.

Identifying the dark current sources at different temperatures is very useful for proper modelling and optimization of the electrical and optical characteristics of APDs operated in linear mode. Current-voltage characteristics of the linear APDs contain the information on the dark current sources. Fitting the model parameters of TAT for InP and SRH for InP, InGaAs and InGaAsP for various temperatures is a difficult task due to the lack of the

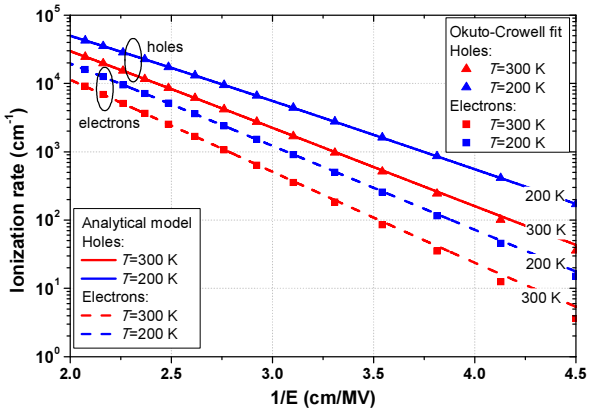


Figure 2. Fitted Okuto-Crowel impact ionization coefficients for electrons and holes for temperatures of 200 K and 300 K.

experimental data. However, there are plenty of literature sources describing the impact of these dark current sources on the DCR of SPADs at various temperatures. We used numerical computation to calculate probabilities that a hole or an electron will initiate an avalanche in [15]. Electron and hole avalanche probabilities are used together with the carrier generation rate profiles obtained from TCAD simulations to calculate DCR. Parameters of the TCAD models for SRH from InP, InGaAs and InGaAsP and TAT from InP are fitted and their contributions to DCR show excellent agreement for various temperatures and device geometry parameters as compared to the measured and simulated data.

B. Electrical and optical analysis of the InP/InGaAs APD

The fitted parameters for the dark current generation are then used in current-voltage simulations of the structure depicted in Fig 1. Simulation results of the current-voltage characteristics for temperatures of 200 K and 300 K are shown in Fig. 3. For both temperatures, the majority of the dark current is originating in InP. The contribution to the dark current from InGaAs and InGaAsP starts at approximately 40 V, which is the punchthrough voltage when those regions become fully depleted. Contribution to the dark current from InGaAsP is negligible at both temperatures. For the temperature of 300 K, the contribution to the total dark current from the InGaAs is comparable to the contribution to the dark current from InP. On the other hand, the contribution of the dark current from InGaAs at 200 K is almost two orders of magnitude lower than the contribution to the dark current from InP. Furthermore, there is an increase of the dark current coming from InP that starts at around 35 V. The origin of this dark current is TAT from InP layer. These results are in agreement to the results reported in the literature where the dominant mechanism that determines the DCR at lower temperatures is TAT. TAT is caused by the high electric field in the multiplication region and can be controlled by decreasing the trap concentration in the region or by adjusting the geometry of the device and increasing the multiplication region thickness.

Optical simulations of the structure are performed and the results of the current-voltage characteristics are shown in Fig. 4 (a). Simulations are done for temperatures of 200 K and 300 K. Complex refractive index for

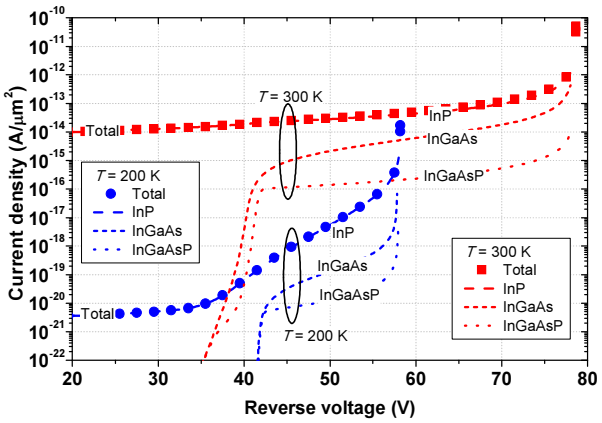
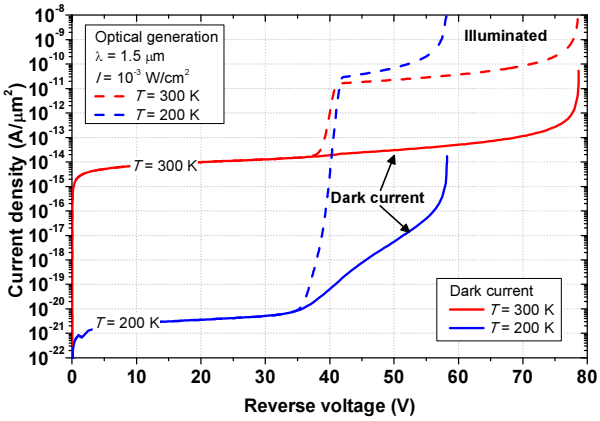


Figure 3. Current-voltage characteristics of the proposed InP/InGaAs APD structure at temperatures of 200 K and 300 K. Symbols: total current; lines: contribution to the dark current from InP, InGaAs and InGaAsP.

(a)



(b)

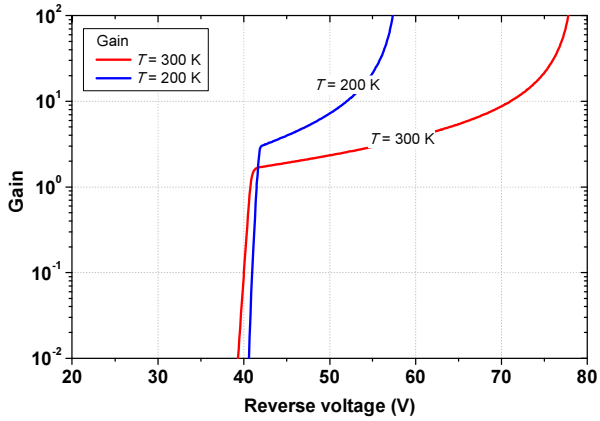


Figure 4. (a) Current-voltage and (b) gain characteristics of the illuminated InP/InGaAs APD for temperatures of 200 K and 300 K.

$In_{0.53}Ga_{0.47}As$ material is not available in the simulator so the fitting to the complex refractive index from [16] is performed. The structure is exposed to the light with wavelength of 1.5 μm and the intensity of light is 10^{-3} W/cm^2. For both temperatures, the punchthrough voltage is around 40 V. On the other hand, the breakdown voltage changes from 58 V to 79 V for the temperatures of 200 K and 300 K, respectively. The punchthorugh voltage of 60 V is reported in [6]. The difference can be attributed to the variations of the doping profiles and thicknesses of the multiplication region. Current-voltage characteristics of the illuminated structure is used to calculate the gain characteristics of the InP/InGaAs diode. Unity gain is defined to be at the diode voltage where the quantum efficiency reaches 80 %. Gain characteristics are depicted in Fig. 4 (b). Since the punchtrough voltage is almost the same for both temperatures and the breakdown voltage is smaller at 200 K, the multiplication gain increases more swiftly for the device operated at 200 K.

C. Excess noise characteristics

Statistical fluctuation of the avalanche process generates noise in the electrical current. In the operation of the linear device, knowledge on the noise that the APD introduces in the electronic system is of utmost importance. Analytical expression for calculation of the excess noise factor (F) is derived in [17]:

$$F = M\left[1 - (1-k)\cdot\frac{(M-1)^2}{M^2}\right], \qquad (1)$$

where M is multiplication gain and k is the ratio of the maximum ionization coefficients of electrons and holes.

Excess noise factor is calculated for InP/InGaAs APD structure for temperatures of 200 K and 300 K and is depicted in Fig. 5. Multiplication gain as a function of voltage is determined by the exposure to the light with wavelength of 1.5 µm and intensity of 10^{-3} W/cm^2. Ionization coefficient profiles for electrons and holes are extracted from 1D simulations for the same bias voltages. Excess noise factors at M = 10 are 4.3 and 4.6 for temperatures of 200 K and 300 K, respectively. Using the APD diode at lower temperatures improves the overall sensitivity of the device for optical detection due to the decrease of the dark current. Lowering the temperatures also decreases the excess noise factor. Other than the decrease of the temperature, excess noise factor depends on the structure of the layer stack of the APD, which is not analyzed in this paper.

III. PROCESS SIMULATIONS

A. *Fitting of the Zn diffusion model parameters*

Zn diffusion is commonly used in the formation of the *pn*-junctions in InP/InGaAs APDs. Process simulations of the Zn diffusion into InP are also added to the TCAD simulation environment. Currently, Zn diffusion into InP is not modeled in the Sentaurus Process [21]. Therefore, we performed calibration of the diffusivity models in order to simulate the Zn diffusion. Device simulations using the Zn diffusion profiles are important for obtaining realistic electrical and optical characteristics of APDs. Process simulations of the Zn diffusion along with device simulations can also be used in the design of the guard rings of InP/InGaAs APDs.

SIMS profiles of Zn diffusion into InP are reported in various literature sources [18]-[20]. Mechanisms governing Zn diffusion and the diffusion models are discussed in [18]. The diffusion is dominated by interstitial-substitutional mechanism where Zn diffuses as a singly ionized interstitial. Zn diffusivity is proportional to the hole concentration and the background concentration can significantly reduce the Zn diffusion. Using the SIMS profiles from the literature and calibrating the simulator parameters, Zn diffusion was simulated in Sentaurus Process simulator.

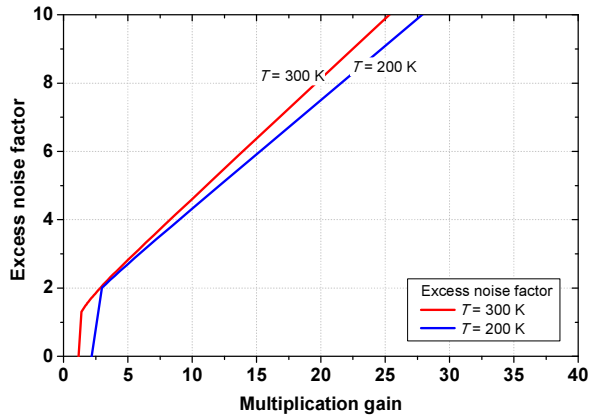


Figure 5. Excess noise factor of the simulated InP/InGaAs APDs for temperatures of 200 K and 300 K

Sentaurus Process offers a number of different diffusion and impurity activation models that could be used for modeling of the Zn diffusion. We focused on the constant diffusion model and Fermi diffusion model. Contrary to constant diffusion model, Fermi diffusion model takes into account the dependency of diffusivities on the electron (hole) concentrations [21]:

$$\frac{\partial C_A}{\partial t} = \nabla \sum_{X,c} D_{AX^c} \left(\frac{n}{n_i}\right)^{-c-z} \nabla \left(C_A^+ \left(\frac{n}{n_i}\right)^z\right), \qquad (2)$$

$$D_{AX^c} = D^0_{AX^c} \exp\left(-\frac{D^E_{AX^c}}{kT}\right), \qquad (3)$$

where C_A is the concentration of substitutional dopands, C_A^+ is the active portion of C_A, c is the charge state of point defect, z is the charge state of dopant A, n_i is the intrinsic concentration, n is the electron (hole) carrier concentration, X is either interstitial or vacancy, k is the Boltzmann constant, T is the temperature, $D^0_{AX^c}$ and $D^E_{AX^c}$ are Fermi diffusion constants that can be calibrated in the process simulator. Dopant activation is modeled by solid solubility model. Parameters of the solid solubility model are taken from [18].

Process simulations of the Zn diffusion using the proposed models are compared to the SIMS data from [18] and depicted in Fig. 6. Sulfur background concentration of $2{\cdot}10^{16}$ cm^{-3} in InP is used. Diffusion assumes constant surface concentration of $8{\cdot}10^{18}$ cm^{-3}. The diffusion time is 30 min at 475° C. Zn diffusion profile obtained by the constant diffusion models largely underestimates the shape of the real diffusion profile obtained from SIMS. Femi diffusion model with c = 2 has a sharper decrease of the Zn concentration near the *pn*-junction than the real SIMS profile. On the other hand, Fermi diffusion model with c = 1 can excellently fit the SIMS data. Values of the fitting constants $D^0_{AI^c}$ and $D^E_{AI^c}$ are 10^{-3} cm^2/s and 1.75 eV, respectively.

The results of the additional verification of the fitted Fermi diffusion model are plotted in Fig. 7. (a) and (b). In Fig. 7 (a) the diffusion is simulated with temperatures of 475° C and 500° C and the durations of 30 min and 15 min, respectively. Bulk doping concentration is $2{\cdot}10^{16}$ cm^{-3}. Excellent matching of the simulated profile to the SIMS

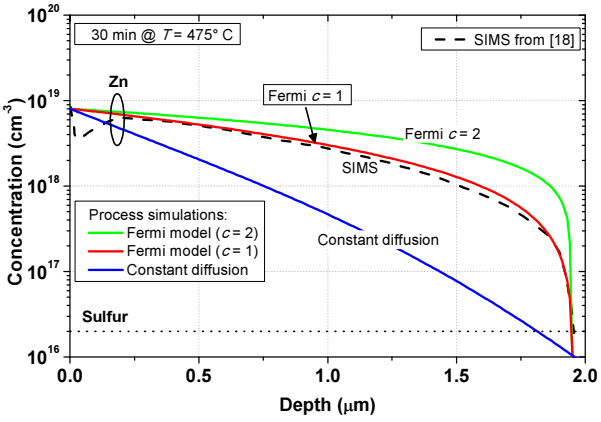


Figure 6. Comparison of the Zn diffusion profile obtained by constant diffusion model, Fermi model with c = 1 and c =2 with SIMS Zn profiles from [18]

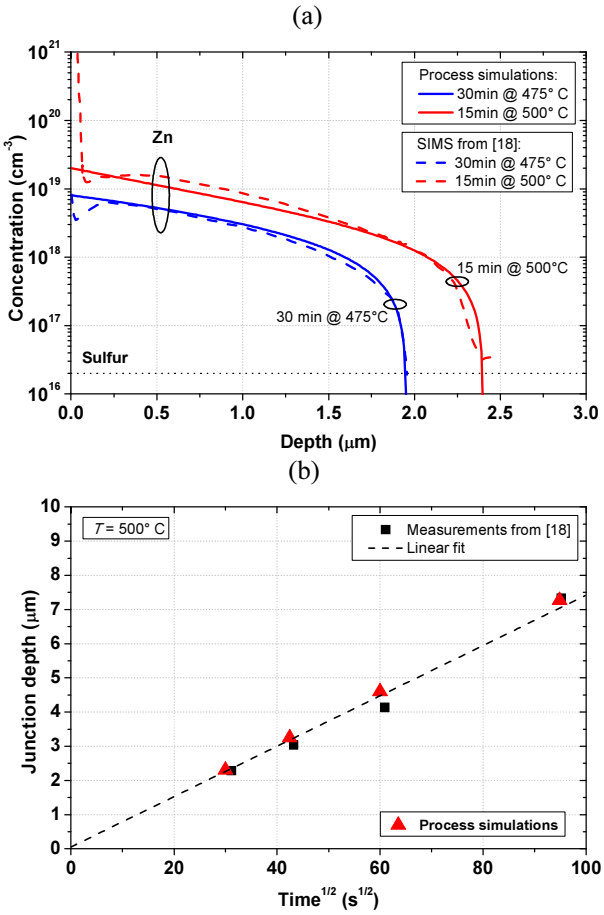


Figure 7. Verification of the fitted Fermi diffusion model parameters. (a) Comparison of simulated doping concentration profile with Zn SIMS profile from [18]. (b) Comparsion of the simulated junction depth with the extracted junction depth from [18].

measurement can be achieved for both diffusion parameters. *PN*-junction depth is extracted versus the square root of the diffusion time for diffusion temperature of 500° C and depicted in Fig. 7 (b). The simulated *pn*-junction depth is compared to the measurements [18]. Junction depth displays the same behavior versus $t^{1/2}$ and is a linear function of $t^{1/2}$. Simulation results show that calibrated Fermi diffusion model provides a realistic model capable of assessing the Zn diffusion profile.

B. Device simulations of APDs with realistic p^+ region doping profile

The APD fabrication is simulated with the fitted parameters of the Fermi model for Zn and the obtained diffusion profiles are later used in the device simulations. Doping concentration profiles of structures where p^+ region is realized as a box-like constant concentration profile and as a diffused Zn profile are depicted in Fig 8. Junction depth in both cases is 2 μm. Temperature used in the simulations of the Zn diffusion is 500° C. The simulated diffusion time is 11.2 min that is needed to achieve 2 μm junction depth. Constant surface concentration of $2 \cdot 10^{19}$ cm^{-3} is assumed.

Current-voltage characteristics of the structures with both p^+ region defined as a constant concentration and diffused Zn region are depicted in Fig. 8. Breakdown voltage for structure with p^+ region simulated with diffused Zn profile increases approximately by 1.5 V for both temperatures. Dark current at 90 % of V_{BR} for device

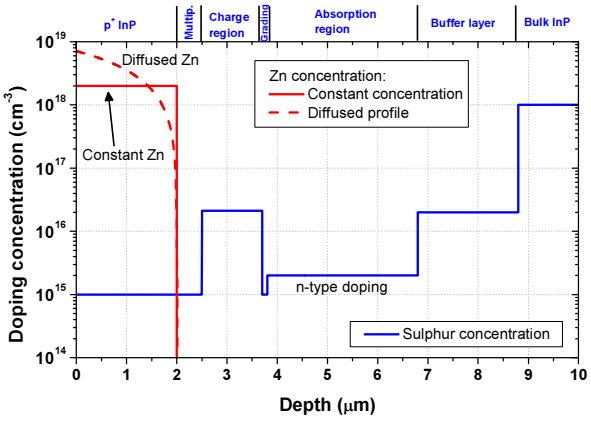


Figure 8. Cross section of the structure with p^+ region realized as a box-like constant concentration profile and diffused Zn profile

operated at 300 K is $1.26 \cdot 10^{-13}$ A/μm^2 and $1.27 \cdot 10^{-13}$ A/μm^2 for structures with box-like constant concentration p^+ region and diffused Zn p^+ region, respectively. On the other hand, dark current at 90 % of V_{BR} at 200 K is $1.5 \cdot 10^{-17}$ A/μm^2 for structure with box-like constant concentration p^+ region and $8.9 \cdot 10^{-18}$ A/μm^2 for structure with diffused Zn p^+ region. The dark current of the structure with diffused Zn p^+ region decreases by a factor of 1.7 compared to the structure with box-like constant concentration p^+ region. The decrease of the dark current is a result of the reduced TAT. For structure with realistic p^+ region, the maximum electric field decreases, resulting in a smaller TAT. Reduced dark current can increase the sensitivity of the APD. The change in the breakdown voltage and current-voltage characteristics shows the importance of using the realistic diffusion profiles in the simulations of the device especially at lower temperatures. This is expected to become even more important in 2D simulations of the APD devices.

IV. CONCLUSION

Comprehensive TCAD simulation environment capable of simulating both device and process characteristics for InP/InGaAs APDs is demonstrated. This simulation procedure can be used in analysis and optimization of complex InP/InGaAs APD structures both in 1D and 2D. The knowledge of the realistic diffusion

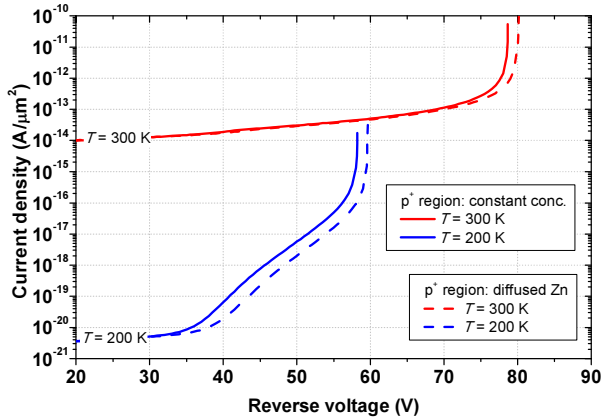


Figure 9. Comparison of the current-voltage characteristics for the structure with box-like constant concentration p^+ region and structure with diffused Zn p^+ region at temperatures of 200 K and 300 K

parameters is an essential parameter for a successful design and analysis of the guard rings at the periphery of the APD.

Analysis of the linear characteristics of a InP/InGaAs APD is demonstrated. Fitted model parameters for the impact ionization model, complex refractive index for InGaAs, TAT model for InP and SRH models for all the materials in the layer stack are used in this analysis. Process simulations of the diffusion of Zn into InP are added to the TCAD simulation environment. Impact of the real diffusion profile on the linear characteristics is also analyzed. Real diffusion profile changes the breakdown voltage and impacts the TAT in InP region. Decrease of the dark current for a factor of 1.7 at 90 % of V_{BR} at 200 K is obtained by using the realistic diffused Zn profile. The importance of using the realistic Zn diffusion profiles is expected to be more critical in 2D simulations of the InP/InGaAs APDs.

REFERENCES

[1] J. C. Campbell, "Recent Advances in Telecommunications Avalanche Photodiodes", Journal of Lightwave Technology, vol. 25, no. 1, pp. 109-121, Jan. 2007.

[2] N. Gisin, G. Ribordy, W. Tittel, H. S. Zbinden, "Quantum cryptography", Reviews of Modern Physics, 74 (1), pp. 145-195, Jan 2002.

[3] U. Schreiber, C. Werner, "Laser Radar Ranging and Atmospheric Lidar Techniques", Proceedings of SPIE, SPIE, 1997, Dec. 1997

[4] S. Verghese et al., "Geiger-mode avalanche photodiodes for photon-counting communications," Digest of the LEOS Summer Topical Meetings, 2005, pp. 15-16.

[5] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," Applied Optics, Vol. 35, Issue 12, pp. 1956-1976, 1996.

[6] Y. Liu, S. R. Forrest, J. Hladky, M. J. Lange, G. H. Olsen and D. E. Ackley, "A planar InP/InGaAs avalanche photodiode with floating guard ring and double diffused junction" Journal of Lightwave Technology, vol. 10, no. 2, pp. 182-193, Feb 1992.

[7] S.R. Forrest, R.G. Smith, and O.K. Kim, "Performance of $In_{0.53}Ga_{0.47}As$/InP avalanche photodiodes," IEEE J. Quantum Electron.,vol. QE-18, pp. 2040-2048, 1982.

[8] F. Acerbi, M. Anti, A. Tosi and F. Zappa, "Design Criteria for InGaAs/InP Single-Photon Avalanche Diode" IEEE Photonics Journal, vol. 5, no. 2, pp. 6800209-6800209, April 2013.

[9] C. L. F. Ma, M. J. Deen and L. E. Tarof, "Multiplication in separate absorption, grading, charge, and multiplication InP-InGaAs avalanche photodiodes", IEEE Journal of Quantum Electronics, vol. 31, no. 11, pp. 2078-2089, Nov 1995.

[10] J. P. Donnelly et al., "Design Considerations for 1.06-μm InGaAsP-InP Geiger-Mode Avalanche Photodiodes" IEEE Journal of Quantum Electronics, vol. 42, no. 8, pp. 797-809, Aug. 2006.

[11] X. Jiang, M. A. Itzler, R. Ben-Michael and K. Slomkowski, "InGaAsP–InP Avalanche Photodiodes for Single Photon Detection", IEEE Journal of Selected Topics in Quantum Electronics, vol. 13, no. 4, pp. 895-905, July-aug. 2007.

[12] M. Anti, F. Acerbi, A. Tosi and F. Zappa, "2D simulation for the impact of edge effects on the performance of planar InGaAs/InP SPADs", Proc. SPIE 8550, Optical Systems Design 2012, 855025

[13] M. Anti, F. Acerbi, A. Tosi and F. Zappa, "Integrated simulator for single photon avalanche diodes," Numerical Simulation of Optoelectronic Devices (NUSOD), 2011 11th International Conference on, Rome, 2011, pp. 47-48.

[14] Sentaurus Device User Guide, Synopsys, Mountain View, CA, USA, Mar. 2016.

[15] T. Knežević, T. Suligoj, "Examination of the InP/InGaAs Single-Photon Avalanche Diodes by Establishing a New TCAD-based Simulation Environment", submitted.for publication

[16] S. Adachi, "Physical Properties of III-V Semiconductor Compounds: InP, InAs, GaAs, GaP, InGaAs, and InGaAsP", John Wiley & Sons, 1992

[17] Mclntyre R.J., "Multiplication Noise in Uniform Avalanche Diodes", IEEE Trans. Electron Devices, ED13, pp. 164-168 (1966)

[18] G. J. van Gurp, P. R. Boudewijn, M. N. C. Kempeners and D. L. A. Tjaden, "Zinc diffusion in n-type indium phosphide" Journal of Applied Physics, Vol. 61, pp. 1846-1855, 1987.

[19] S. Y. Yang and J. B. Yoo, "Characteristics of Zn diffusion in planar and patterned InP substrate using Zn_3P_2 film and rapid thermal annealing process", Surface and Coatings Technology, Vol. 131, Issues 1–3, Pages 66-69, 2000.

[20] H. S. Marek and H. B. Serreze, "Diffusion coefficients and activation energies for Zn diffusion into undoped and S-doped InP", Applied Physics Letters, vol. 51, pp. 2031-2033, 1987

[21] Sentaurus Process User Guide, Synopsys, Mountain View, CA, USA, Mar. 2016.

Design of Passive-Quenching Active-Reset Circuit with Adjustable Hold-Off Time for Single-Photon Avalanche Diodes

I. Berdalović*, Ž. Osrečki*, F. Šegmanović*, D. Grubišić**, T. Knežević* and T. Suligoj*
* Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
** Laser Components DG, Inc., Tempe, Arizona, USA
tomislav.suligoj@fer.hr

Abstract - Single-photon avalanche diodes (SPADs) are gaining popularity in applications where low intensity light needs to be detected. Since they are used in Geiger mode, where the self-sustaining avalanche needs to be quenched, an important part of the detection circuitry is the quenching circuit. First, we examine the operation of a basic passive quenching circuit consisting of the SPAD and two series resistors and measure the SPAD's dark count rate. Then we implement a passive quenching circuit with active reset (PQAR). Without a sufficiently long hold-off time between quenching and reset the circuit does not operate properly. Because of that, a hold-off time is introduced by means of an adjustable time delay circuit. The behavior of the PQAR circuit for different hold-off times is then examined, and the minimum hold-off time of 1 µs, which still allows for correct operation with the given circuitry is determined. Finally, a comparison is made between the passive and the PQAR circuit, focusing on the advantages of active over passive reset.

I. INTRODUCTION

A single-photon avalanche diode (SPAD) is a solid-state photodetector used for detecting low-intensity optical signals [1]. Low cost, miniature size, higher quantum efficiency and low voltage operation made the SPAD a replacement for photomultiplier tubes (PMT) in many applications today. The SPAD is basically a *p-n* junction reverse biased above the breakdown voltage (as opposed to avalanche photodiodes operated below the breakdown voltage) and thus operated in Geiger-mode, where each electron-hole pair can trigger an avalanche multiplication process [2]. The avalanche current rises swiftly until quenched by an external circuit. The leading edge of the current pulse gives information about photon arrival time. However, the avalanche multiplication process can also be triggered by thermally generated electron-hole pairs inside the active region or by the charge released from deep-level traps. These mechanisms give rise to dark count rate (DCR) and represent the noise of a SPAD. Charge released from deep-level traps can cause afterpulsing [3].

The avalanche, once triggered, is self-sustained until quenched by a quenching circuit, during an interval called the quenching time. The quenching circuit also has to detect the avalanche, produce a readable output signal and prepare the diode for new detections, all that during a time interval called dead time [4, 5]. The dead time limits the SPAD's maximum operating frequency, because the diode is not able to detect photons in that period. There are two basic types of quenching circuits: passive and active [4]. Passive quenching circuits shut down the avalanche process by reducing the diode's voltage below the breakdown voltage by means of a ballast resistor connected in series with the diode. The avalanche current creates a voltage drop on the resistor, thereby reducing the voltage on the diode below the breakdown voltage [6]. Active quenching circuits use external circuitry to shut down the avalanche, and an improved performance can be achieved, with circuit complexity as a disadvantage. After quenching is complete, the diode voltage must be restored to its operating value. This can be done passively (long reset time, simple circuit) and actively (small reset time, complex circuit).

The type of quenching circuit determines the type of SPAD operation: free-running or gated. In free-running mode, the SPAD is constantly biased above the breakdown voltage, as opposed to gated mode operation, where the bias voltage is periodically lowered. In gated mode, the diode is only active when a photon needs to be detected, therefore, the frequency of incoming photons must be known before detection. In free-running mode, the diode can detect randomly incoming photons and is only limited by the dead-time needed to quench the avalanche from a previous detection and restore the operating voltage. The diode voltage is restored to its operating value in a period called the reset time. During the reset time interval, the diode is still biased above the breakdown voltage and the avalanche can be triggered, so there must be a certain period after quenching and before reset where leftover charge is removed from the diode. This time interval is called hold-off time [4]. All residual charge can cause afterpulsing during and after reset, thereby reducing the possibility of detecting the next photon. As a trade-off between the circuit complexity and performance, a passive quenching active reset (PQAR) circuit passively quenches the avalanche and actively restores the SPAD's operating voltage, shortening the reset time [7]. There are different techniques for afterpulsing reduction in PQAR circuits, some of them

This work was supported by the Croatian Science Foundation under contract no. 9006.

being variable-load [8] and gated mode PQAR quenching circuits [9]. Monolithic quenching circuit design results in lower parasitic capacitances and higher performance [10].

In this paper, a PQAR circuit with adjustable hold-off time is designed. First, with a simple passive quenching circuit, the diode's voltage and avalanche current waveforms are measured. An active reset mechanism is then implemented using a switching transistor in parallel to the ballast resistor. However, if the active reset starts during passive quenching, the avalanche is not properly quenched and the circuit does not function as wanted. The introduction of a hold-off time between quenching and reset results in a properly functioning quenching circuit. As a means of hold-off time optimization, a quenching circuit with adjustable hold-off time is designed with off-the-shelf components. With negligible complexity, the circuit can be used to determine minimum hold-off time for a certain SPAD with respect to afterpulsing probability and counting frequency. Dark counts are used as a trigger for avalanche multiplication in all measurements.

II. QUENCHING CIRCUITS

A. Passive Quenching Circuit

The passive quenching circuit with passive reset is the basic circuit to shut down the avalanche of the photodiode. As shown in Fig. 1, it consists of a reverse biased SPAD in series with a large ballast resistor R_A=560 kΩ and a small resistor R_C=1 kΩ.

The SPAD used in this and all subsequent circuits is a SAP500 from Laser Components DG [11]. It is operating at a voltage V_{OP}=141.1 V, which is higher than the breakdown voltage V_{BR}=139.1 V by the overvoltage $V_{EX}=V_{OP}-V_{BR}$=2 V. When an avalanche is triggered, the current through the diode increases and the voltage drop across the resistor R_A increases, causing the diode voltage to decrease below the breakdown voltage. As that happens, the avalanche is quenched and the diode current becomes negligible. The voltage drop across the resistor R_C, shown in Fig. 2 a), is in fact caused by the diode current, so the cathode voltage waveform is the same as the diode current waveform, shown enlarged in Fig. 3 (curve a). We can see that the quenching time, i.e. the time until the avalanche current drops, is approximately 50 ns.

Since R_A is loaded with a significant capacitance, namely the oscilloscope input capacitance and other

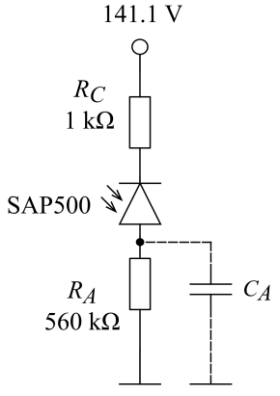


Figure 1. Schematic of the passive quenching circuit.

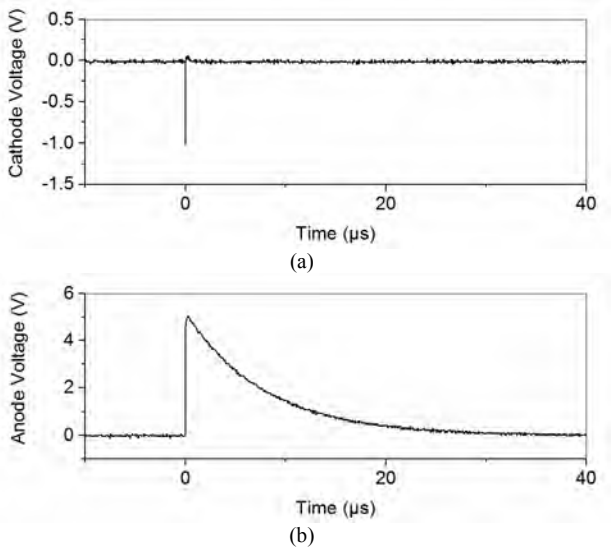


Figure 2. Transient response of passive quenched SPAD: (a) AC coupled cathode voltage and (b) Anode voltage.

parasitic capacitances, the currents through R_C and R_A are different. Most of the diode current is actually capacitance-charging current, and only a fraction of the diode current flows through R_A, producing a voltage drop of around 5 V only, as seen in Fig. 2 b). After the avalanche is quenched, the anode capacitance has to discharge through the ballast resistor R_A and the voltage decreases to zero with an RC time constant determined by the anode capacitance and R_A. As shown in Fig. 2. b), this time constant is rather large so the reset time of the circuit is around 30 μs.

Thus far, we have described the case with an oscilloscope probe connected to the anode. The anode capacitance also determines the peak avalanche current. As shown in Fig. 3 (curve b), the peak current is only about 0.3 mA without the oscilloscope probe connected to the anode, since less current is needed to charge a smaller C_A.

B. PQAR Circuit without Hold-Off

To improve the passive reset, we wanted to implement a passive quenching circuit with active reset. The first idea was to use a constant level discriminator (CLD) to detect the cathode voltage drop. The output of the CLD is connected to the gate of a MOSFET, and when it goes high, it switches the MOSFET on, effectively connecting the anode to the ground and immediately resetting the circuit. The schematic of this circuit is shown in Fig. 4.

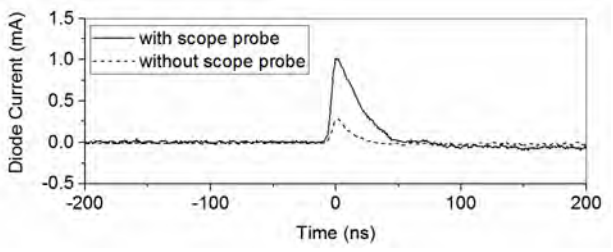


Figure 3. Transient response of passive quenched SPAD: Diode current a) with and b) without oscilloscope probe connected to the anode.

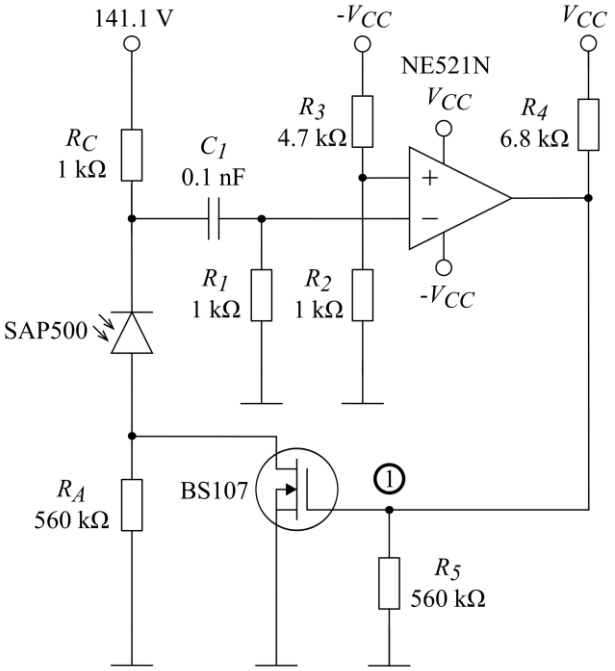


Figure 4. Schematic of the PQAR circuit without hold-off.

The threshold voltage of the CLD is determined using two resistors, R_2=1 kΩ and R_3=4.7 kΩ, connected as a voltage divider from the negative supply to the positive input. This gives us a threshold of about V_{TR}=–0.85 V. The negative input of the CLD is connected to the AC coupled cathode. Thus, when the cathode voltage drops by more than 0.85 V, the CLD is triggered. The output capacitance of the MOSFET (BS107) is 30 pF, which is larger than the oscilloscope capacitance, so this is the dominant capacitance in C_A. As a result, the peak avalanche current is now about 2 mA, producing a cathode voltage drop of nearly 2 V, which is more than enough to trigger the CLD.

Fig. 5 shows the transient response of this circuit.

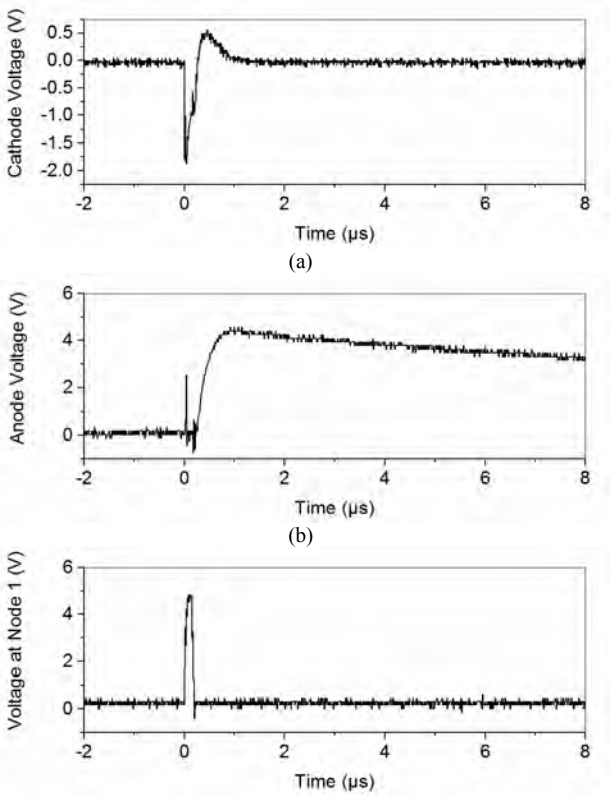


Figure 5. Transient response of PQAR circuit without hold-off: (a) AC coupled cathode, (b) Anode (c) Output of the CLD (node 1).

It can be clearly seen that the anode voltage is indeed reset when the MOSFET is switched on. However, after the MOSFET is switched off, the anode voltage rises again and this time, resets slowly, as in the case of passive reset. This is caused by the fact that reset occurs too soon, before the avalanche is properly quenched. Because of that, a hold-off time is needed between quenching and reset.

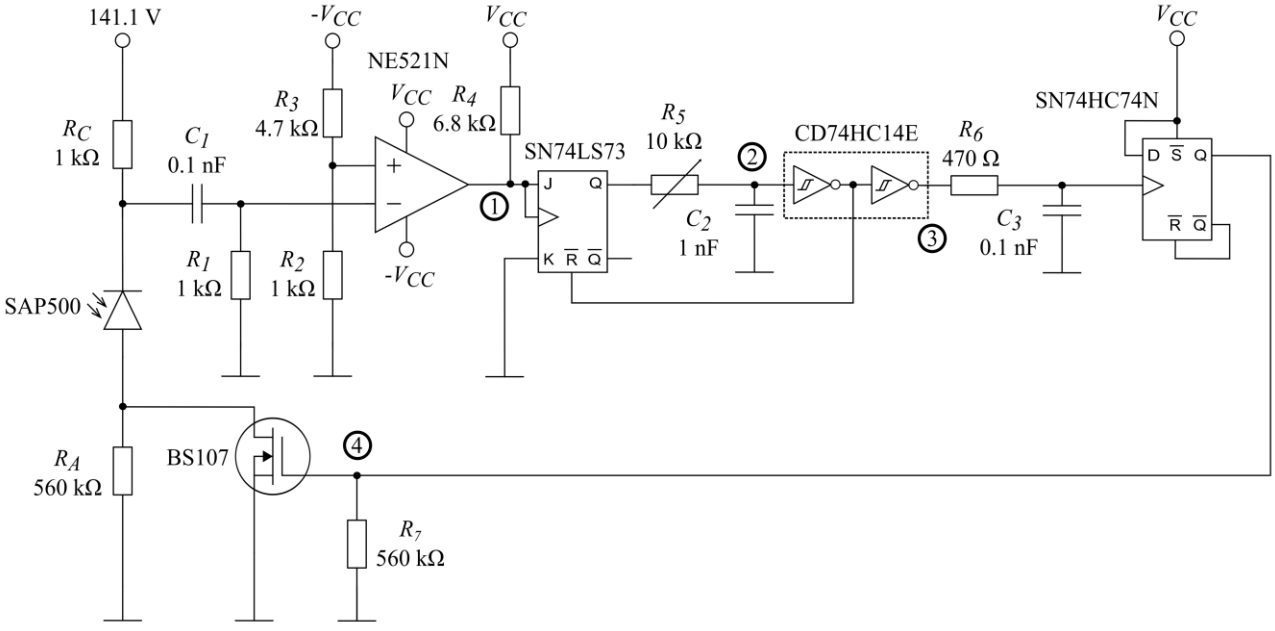


Figure 6. Schematic of the PQAR circuit with adjustable hold-off time, including the pulse generator circuit with the adjustable time delay.

C. PQAR Circuit with Adjustable Hold-Off Time

We have concluded in Section II. B that the PQAR circuit without hold-off between quenching and reset does not operate as wanted. Therefore, we have implemented a pulse generator circuit with adjustable time delay between the output of the CLD and the gate of the MOSFET. The purpose of this circuit is to create a voltage pulse at the gate with adjustable duration after the output of the CLD goes high. The schematic of the whole quenching circuit including the pulse generator is shown in Fig. 6. As in the previous circuit, the cathode voltage drop caused by an avalanche event triggers the CLD (NE521), and its output goes high while the AC coupled cathode voltage is below its threshold voltage. This in turn triggers the J-K flip-flop (74LS73), its Q output goes high and charges the capacitor C_2 with an RC time constant determined by R_5 and C_2. When the capacitor voltage reaches the threshold voltage of the inverting Schmitt triggers (74HC14), the J-K flip-flop is reset asynchronously. The voltage at node 3 goes high and triggers the pulse shaper (74HC74), which subsequently creates a 12 ns pulse at the gate of the MOSFET, switching it on. The RC network consisting of R_6 and C_3 suppresses the ringing at node 3 and ensures the proper operation of the pulse shaper. The pulse shaper is needed to generate a short pulse at the gate, since node 3 stays high considerably longer, until the voltage of the discharging capacitor C_2 falls below the Schmitt trigger threshold. For a higher count rate, it is desirable that the MOSFET is switched on for as short a period of time as possible, in our case about 12 ns. By changing the resistance R_5 we can charge C_2 with different time constants, which gives us a simple way of adjusting the delay of the gate pulse, i.e. the hold-off time.

The voltage waveforms at certain nodes of the described circuit for a hold-off time of approximately 5 μs are shown in Fig. 7.

Fig. 8 shows the anode and cathode voltages of the same circuit, but using the maximum value of resistor R_5, which is 10 kΩ. This value gives us a maximum hold-off time of approximately 10 μs. Obviously, the longer the hold-off time, the greater the possibility that another avalanche event may occur during that time, i.e. before reset. If that is indeed the case, the new avalanche will occur while the diode is still below its operating voltage, which means that the avalanche current will be significantly smaller. That in turn means that the voltage drop on the cathode will be smaller as well, and this voltage drop may not be sufficient to trigger the constant level discriminator. Therefore, if an avalanche occurs during hold-off, it will not be detected by the circuitry. An example of that can be seen in Fig. 8. The avalanche current caused by the second dark count results in a voltage drop of around 0.5 V across the resistor R_C, which is not enough to trigger the CLD. As a result, that count is lost. The diode resets 10 μs after the first dark count.

In order to achieve a high counting frequency, the hold-off time must be as short as possible. Because of the intrinsic time delays of the components used, the minimum hold-off time possible with our circuit was determined to be around 300 ns. However, for such short hold-off times, the afterpulsing becomes a possible issue. Afterpulsing occurs when carriers trapped from an avalanche are subsequently released, triggering new unwanted avalanches [12]. That could explain the

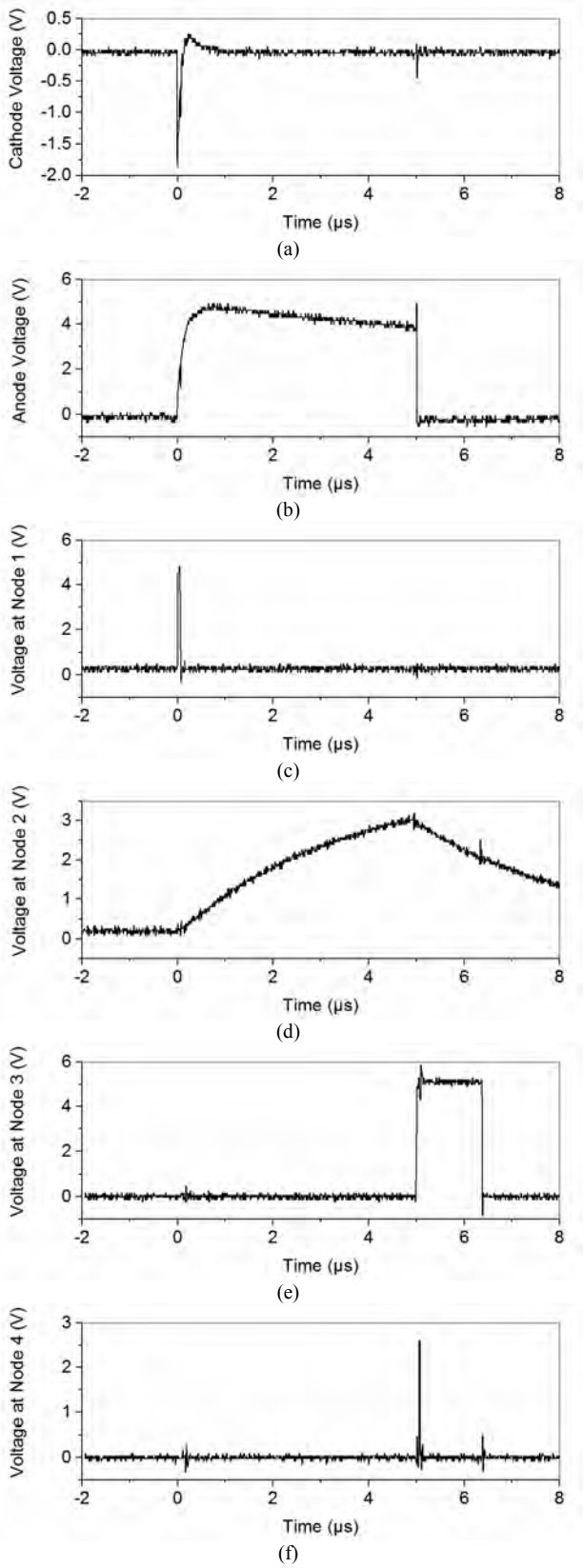


Figure 7. Transient response of PQAR circuit with a hold-off time adjusted to 5 μs in this case: (a) SPAD cathode (AC coupled), (b) SPAD anode, (c) Output of the CLD (node 1), (d) Capacitor C_2 (node 2), (e) Output of the Schmitt triggers (node 3), (f) MOSFET gate (node 4).

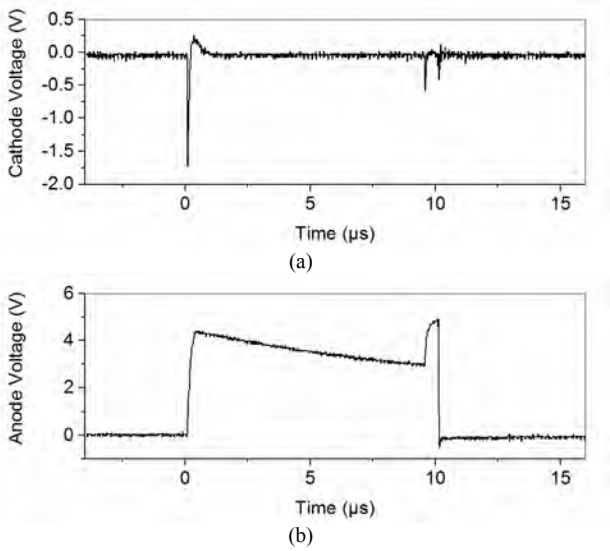


Figure 8. An example of an avalanche pulse occured during hold-off for the PQAR circuit with a hold-off time adjusted to 10 μs: (a) AC coupled cathode voltage and (b) Anode voltage.

waveforms obtained with a hold-off time of around 500 ns, shown in Fig. 9. A new unexpected avalanche occurs immediately after reset in some cases, which could be an afterpulse. This effect is much more pronounced for shorter hold-off times. Thus, for a hold-off time of 500 ns, we have obtained a cascade of multiple avalanches, as shown in Fig. 9. The hold-off time could be further reduced by lowering the overvoltage V_{EX} and thus decreasing the afterpulsing probability. However, a lower V_{EX} results in a lower detection efficiency, so there is a trade-off between shortening the hold-off time and keeping the detection efficiency high.

For all the reasons above, we choose a hold-off time of 1 μs as the optimum hold-off time for the used SPAD. This time is long enough for the afterpulsing probability to be negligible, but not too long, to minimize the number of dark counts during hold-off and to increase the maximum

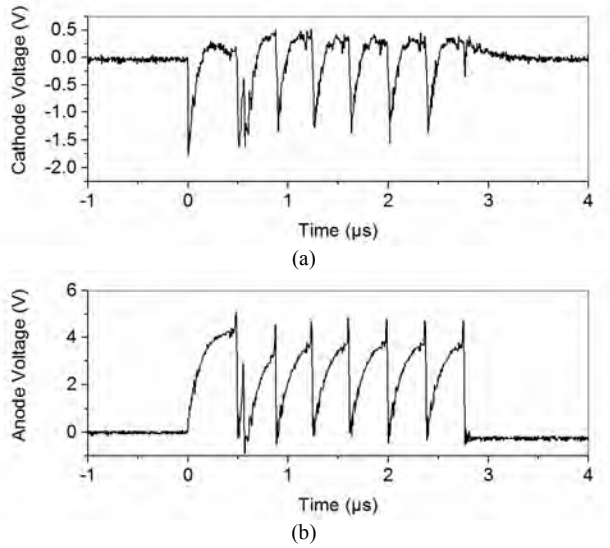


Figure 9. Single shot capture of waveforms of the PQAR circuit with a hold-off time of 500 ns with a cascade of afterpulses: (a) AC coupled cathode voltage and (b) Anode voltage. This event occurs in about 1 in 20 measurements.

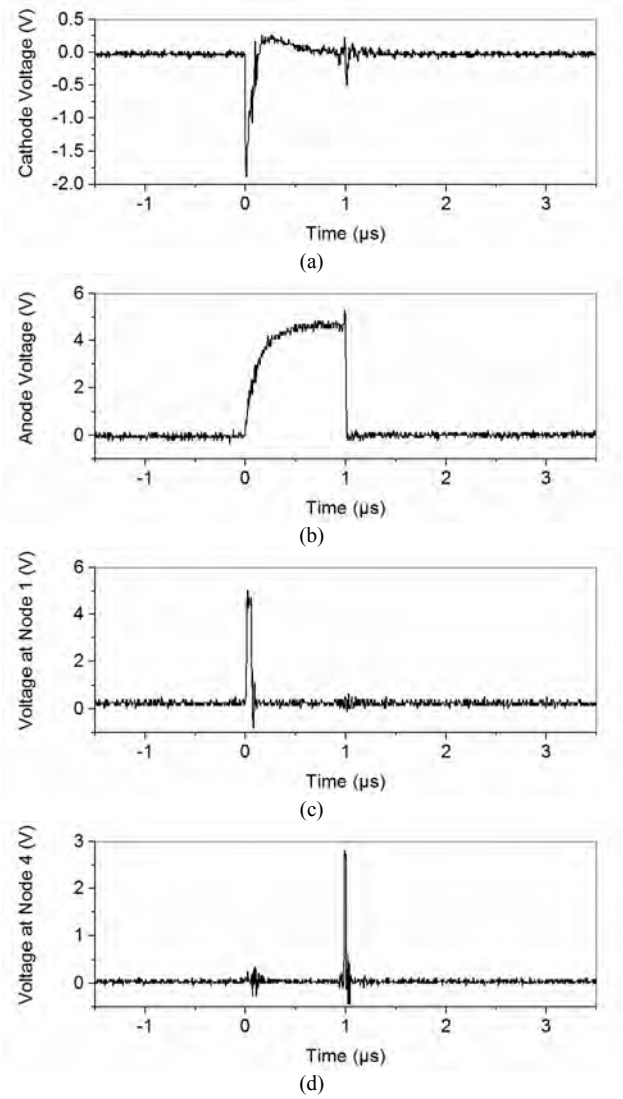


Figure 10. Transient response of the PQAR circuit with the optimum hold-off time of 1 μs: (a) SPAD cathode (AC coupled), (b) SPAD anode, (c) Output of the CLD (node 1), d) MOSFET gate (node 4).

frequency of detected photons. Obviously, for different diodes, this optimum value will vary, and our circuit provides a simple solution for adjusting the desired hold-off time. The voltage waveforms for this final version of the circuit are shown in Fig. 10.

If we compare this PQAR circuit with a hold-off time of 1 μs to the passive quenching circuit using the same resistors in series to the SPAD, we can observe the main advantage of the active over passive reset. While the passive reset time is measured to be around 30 μs, the SPAD with the PQAR circuit is fully reset to its operating voltage in just over 1 μs. That means that the maximum counting frequency of the PQAR circuit is almost 30 times higher than that of the passive circuit, which is a significant improvement. The problems of active reset, such as larger afterpulsing probability, possible detections during hold-off or while the MOSFET is on, are overcome by choosing a suitable hold-off time and by keeping the MOSFET on for very short periods of time, which is achieved by applying short pulses to the gate of the MOSFET.

III. CONCLUSION

The aim of this paper was to analyze the impact of hold-off time on the performance of a PQAR circuit for SPADs. The behavior of the SPAD was examined during the operation with a simple passive quenching circuit. A need for a hold-off time is demonstrated on a simple PQAR circuit where reset starts during quenching. Then, a PQAR circuit with adjustable hold-off time is designed as a means of performance optimization for different SPADs. Finally, the impact of different hold-off times on the performance of the quenching circuit is described, showing that a shorter hold-off time provides a higher counting frequency, but that the hold-off time has to be long enough to prevent false detections. Also, the advantages of active over passive reset, in particular the shorter reset time, are demonstrated.

REFERENCES

[1] H. Dautet et al., "Photon-counting techniques with silicon avalanche photodiodes," Applied Optics, vol. 35, pp. 3894-3900, 1993.

[2] J. Zhang, M. A. Itzler, H. Zbinden, J. W. Pan, "Advances in InGaAs/InP single-photon detector systems for quantum communication," Light: Science & Applications 4, e286, 2015.

[3] M. Stipčević, D. Q. Wang, R. Ursin, "Characterization of a commercially available large area, high detection efficiency single-photon avalanche diode," IEEE Journal of Lightwave Technology, vol. 31, no. 23, pp. 3591-3596, 2013.

[4] A. Gallivanoni, I. Rech, M. Ghioni, "Progress in quenching circuits for single photon avalanche diodes," IEEE Transactions on Nuclear Science, vol. 57, no. 6, pp. 3815-3826, 2010.

[5] M. Stipčević, H. Skenderović, D. Gracin, "Characterization of a novel avalanche photodiode for single photon detection in VIS-NIR range," Optics Express, vol. 18, issue 16, 2010.

[6] B. F. Aull et al., "Geiger-Mode avalanche photodiodes for three-dimensional imaging," Lincoln Laboratory Journal, vol. 13, no. 2, pp. 335-350, 2002.

[7] S. Cova, M. Ghioni, A. Lacaita, C. Samori, F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," Applied Optics, vol. 35, no. 12, pp. 1956-1976, 1996.

[8] S. Tisa, F. Guerrieri, F. Zappa, "Variable-load quenching circuit for single-photon avalanche diodes," Optics Express, vol. 16, no. 3, pp. 2232–2244, 2008.

[9] M. Liu, C. Hu, J. C. Campbell, Z. Pan, M. M. Tashima, "A novel quenching circuit to reduce afterpulsing of single photon avalanche diodes," Proc. SPIE, vol. 6900, no. 5, 2008.

[10] D. Bronzi et al., "Fast sensing and quenching of CMOS SPADs for minimal afterpulsing effects," IEEE Photonics Technology Letters, vol. 25, no. 8, pp. 776-779, 2013.

[11] Laser Components DG, Inc. Pulsed Laser Diodes - Avalanche Photodiodes Catalog, available at: http://www.lasercomponents.com/fileadmin/user_upload/home/Datasheets/lc/kataloge/pld-apd.pdf

[12] M. G. Liu, C. Hu, J. C. Campbell, Z. Pan, M. M. Tashima, "Reduce afterpulsing of single photon avalanche diodes using passive quenching with active reset," IEEE Journal of Quantum Electronics, vol. 44, no. 5, pp. 430-434, 2008.

Impact of the Emitter Polysilicon Thickness on the Performance of High-Linearity Mixers with Horizontal Current Bipolar Transistors

J. Žilak*, M. Koričić, H. Mochizuki**, S. Morita** and T. Suligoj*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Electronics, Microelectronics, Computing and Intelligent Systems, Micro and Nano Electronics Laboratory, Zagreb, Croatia
** Asahi Kasei Microdevices Co. 5-4960. Nobeoka, Miyazaki, 882-0031, Japan
jzilak@zemris.fer.hr

Abstract - The impact of the emitter polysilicon etching in Tetramethyl Ammonium Hydroxide (TMAH) on the characteristics of high-linearity mixers fabricated with the low-cost Horizontal Current Bipolar Transistor (HCBT) is analyzed. During emitter formation, the thick layer of α-Si is deposited over the whole wafer, which is then etched-back in the TMAH. The emitter thickness depends on the TMAH etching time and impacts the HCBT's electrical characteristics. Active down-converting mixers with open-collector topology based on Gilbert cell are fabricated with two types of HCBTs with different TMAH etching time using the lowest-cost HCBT technology with CMOS *n*-well region for *n*-collector. Measurements of mixers' characteristics are done on-wafer by using the multi-contact probes. The mixers achieve maximum IIP3 of 20.2 dBm and conversion gain of 4 dB. Differences in performance characteristics between two mixer types are small indicating that HCBT's circuit performance sensitivity on the emitter thickness variations is relatively small.

I. INTRODUCTION

The improvement of the high-frequency response of the CMOS devices, enabled by the downscaling, led to the widening of the CMOS technology usage in wireless and other radio frequency (RF) analog circuit applications. In order to keep its low-cost and high-volume production, scaling techniques and further CMOS development require increasing investments [1]. On the other hand, bipolar technologies are suitable for the mentioned RF applications at coarser technology nodes due to better high-frequency characteristics, noise factor and higher gain [2], [3]. Hence, the solution to the very cost-sensitive demands of the RF integrated circuits (RFICs) market is the addition of bipolar devices to the coarser lithography CMOS technology. It is critical that such integration is done with minimum number of new masks and process steps, keeping the fabrication costs as low as possible.

The Horizontal Current Bipolar Transistor (HCBT) technology, developed with a novel technological approach as add-on to 180 nm CMOS process, is an example of such low-cost integration. The integration is accomplished with addition of only 2 or 3 lithography masks and few process steps. The fabricated HCBTs with very high peak cutoff frequency (f_T) of 51 GHz and maximum frequency of oscillations (f_{max}) of 61 GHz, along with collector-emitter breakdown voltage (BV_{CEO}) of 3.4 V has been demonstrated [4]. Moreover, the high-voltage transistors are integrated in the process at zero cost [5].

The RF mixers are widely used in mobile and base station transceivers, as well as general purpose RF systems. A spectral efficiency and an intermodulation distortion, generated due the nonlinear nature of active elements, are important effects in modern systems imposing high demands on the RF mixers [6]. Hence, high-linearity mixers with low power consumption are preferable in the integrated mixer design. On the other hand, high performance mixers can be fabricated with a minimum number of large area passive components placed on the chip. Therefore, mixer design has shown to be suitable for RFIC design in novel HCBT technology and high-linearity active mixers with input 3rd order intercept point (IIP3), IIP3 = 23.8 dBm, have recently been demonstrated [7].

In this paper, the impact of the emitter polysilicon etching in Tetramethyl Ammonium Hydroxide (TMAH) on the characteristics of high-linearity mixers fabricated

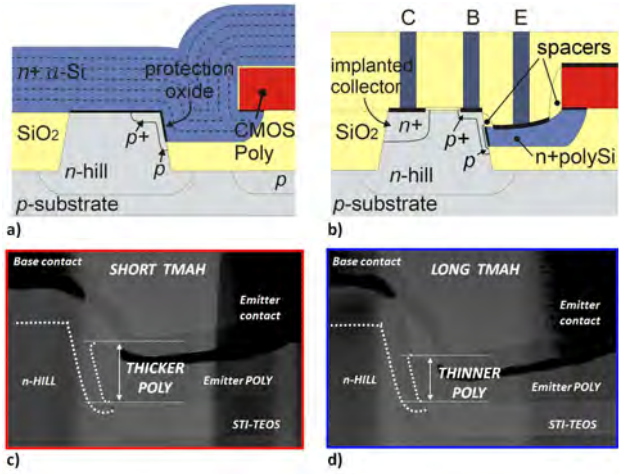


Figure 1. Cross-section of the HCBT: a) α-Si deposition prior to TMAH etching, b) final structure, c) TEM of the intrinsic region with the short-TMAH, d) TEM of the intrinsic region with long-TMAH

This work was supported by the Croatian Science Foundation under contract no. 9006.

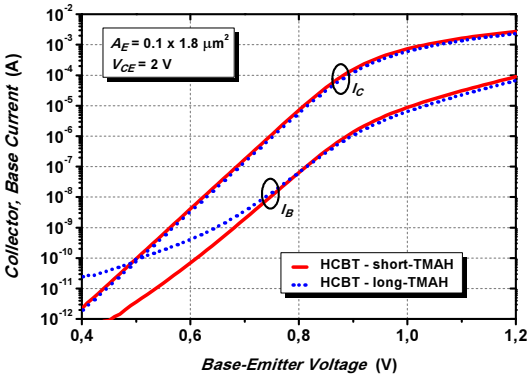


Figure 2. Measured Gummel plots of the HCBTs with the short- and long-TMAH etching times with an emitter area of 0.1×1.8 μm^2 at V_{CE} = 2 V.

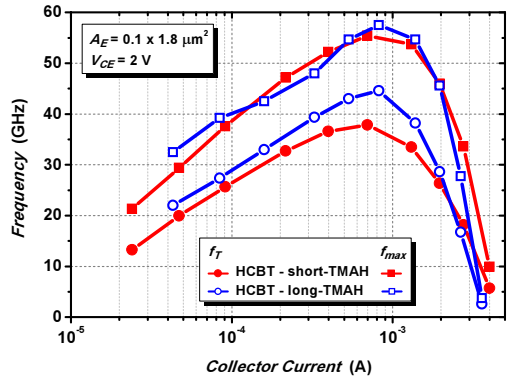


Figure 3. Cutoff frequency (f_T) and maximum frequency of oscilations (f_{max}) versus collector current of the HCBTs with the short- and long-TMAH etching times, with an emitter area of 0.1×1.8 μm^2 at V_{CE} = 2 V.

with the lowest-cost HCBTs using CMOS *n*-well region as *n*-collector with only 2 additional masks is analyzed.

II. Fabricated HCBT Structures

The HCBT fabrication sequence and its integration with the CMOS baseline process using a *base-after-gate* scheme are described in detail in [8]. All examined mixers are designed and fabricated by using the CMOS *n*-well collector HCBT with a single polysilicon region. It is a lowest-cost version of HCBT since it uses only 2 additional lithography masks and skips the additional *n*-collector implantations. A cross-section of the HCBT structure is shown in Fig. 1b. The emitter formation as a subject of this study is described in more detail as follows. The *in situ* phosphorus doped amorphous silicon (α-Si) (Fig. 1a) is deposited and etched-back by using Tetramethyl Ammonium Hydroxide. The deposited polysilicon layer has to be thick enough to achieve planar surface of the deposited film. A thin native oxide layer, grown during the pre-deposition annealing, serves as a protection layer keeping the *n*-hill from the TMAH etching. At the same time, the oxide is thin enough allowing the current flow in or out of the *n*-hill. The TMAH etching is time-controlled and the final polysilicon thickness is determined by the etching duration. Two TEM micrographs of fabricated HCBT structures with different TMAH etching times are shown in Fig. 1. The shorter time (short-TMAH) results in a thicker polysilicon layer (Fig. 1c) and the longer etching time (long-TMAH) results in a thinner polysilicon layer (Fig. 1d).

Electrical characteristics of unit transistors, which differ only in TMAH etching times, are measured. The Gummel characteristics of both HCBT types are shown in Fig. 2. The HCBT with long-TMAH has smaller base current (I_B) at base-emitter voltage V_{BE} = 0.9 V, which is the bias point around peak f_T. Consequently, long-TMAH HCBT has higher peak current gain $\beta_{\text{long-TMAH}}$=133 in comparison to the $\beta_{\text{short-TMAH}}$=117. It is due to the longer distance between intrinsic emitter and the extrinsic base (p^+ base) and consequently reduced electron injection into the base contact. Moreover, the charge sharing between intrinsic and extrinsic base acceptors is reduced if the intrinsic transistor is at larger distance from the extrinsic base, as explained in [8]. The reduced charge sharing effect is also beneficial for the f_T and $f_{\max}$ (Fig. 3) resulting in a higher f_T in the case of the long-TMAH HCBT (thinner emitter polysilicon). The $f_{\max}$ benefits from the increased f_T, but its increase is offset by an increased base resistance in case of thinner emitter at the long-TMAH HCBTs [8].

III. HCBT Mixer Design

High-linearity mixers are designed and fabricated using both unit HCBTs with short- and long-TMAH etching. A down-converting active mixer, using a double-balanced Gilbert cell topology [9] with an open-collector output, is designed and its schematic is shown in Fig. 4. Main circuit parts are a differential pair (Q1 and Q2) as a transconductance stage that transforms the input voltage signal to the current, and a switching quad consisting of 4 transistors (Q3~Q6), that commutates the current providing frequency conversion. The differential Local Oscillator (LO) signal, is generated by the LO buffer circuit. It should be noted that 128 of unit HCBT transistors are connected in parallel in the design for each Q1~Q6 transistor in the scheme. The current mirror (Q7 and Q8) sets the bias current. Degeneration emitter resistors (R_E) are used in order to improve linearity, since it is a critical parameter of mixers in wireless transceivers. The input impedances are designed to be 50 Ω.

The open-collector design provides flexibility in setting parameters such as output impedance, conversion gain and linearity. Moreover, it can be used effectively

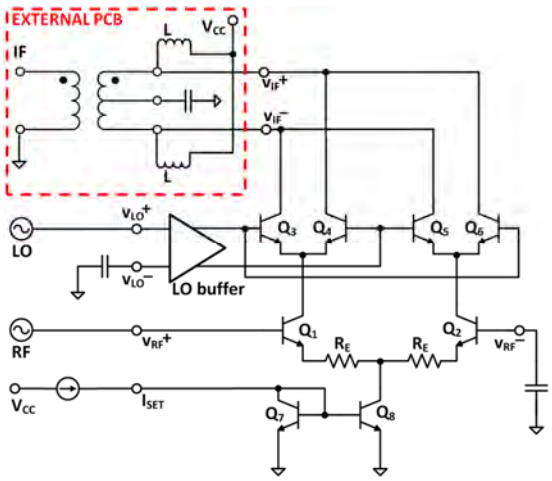


Figure 4. Double-balanced active mixer based on a Gilbert cell with an open collector, which is implemented on-wafer, except the output network that is added on external PCB, as marked.

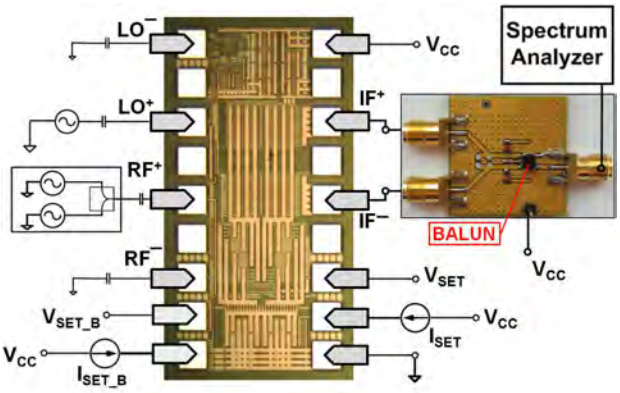


Figure 5. Measurement setup used for the on-wafer measurements with mixer chip photo and photo of additional PCB

with both differential and single-ended filter configurations that are usually connected to the mixer outputs. Mixer performance is, in this case, highly dependent on its output network. It has to reduce the high output impedance of the switching quad transistors and to provide an interface to the load, which assures adequate power transfer with a minimum number of external components [10].

The voltage swing at the output of the LO buffer circuit has to be large enough to switch on and off the switching quad transistors completely (Fig. 4). The buffer is amplifying the input LO signal and provides the single-ended to differential conversion. That is accomplished by the differential amplifier, which is the main part of the buffer circuit. Moreover, the buffer improves the isolation between the LO and other mixer's ports (RF and output IF) and sets the proper DC voltage level for the switching quad transistors (Q3~Q6 in Fig. 4). The output stage of the buffer is build by the emitter followers, which needs to assure sufficient current for driving the LO switching quad capacitances.

The conversion gain (CG) and the IIP3 are the most common figure-of-merits in mixer performance characterization. They are determined by the transconductance of the differential pair (g_m) and the value of the emitter degeneration resistance R_E, as well as by the shape and the magnitude of the LO signal generated by the LO buffer. The CG can be approximated by [11]

$$CG \approx \frac{2}{\pi} \cdot \frac{R_L}{1/g_m + R_E}, \qquad (1)$$

where R_L is the load resistance. The higher g_m and R_L result in the higher CG, while the higher R_E reduces it. On the other hand, both high g_m and R_E are beneficial for IIP3 as a linearity figure of merit, as approximated in [11]

$$IIP3 \approx 4\sqrt{2} v_T \cdot \left(1 + g_m R_E\right)^{3/2}, \qquad (2)$$

where the v_T is the thermal voltage. Hence, in order to achieve CG above 0 dB, the mixer current is increased to have a higher g_m by putting the unit HCBTs in parallel. The chosen value of R_E is 20 Ω and it assures high-linearity behavior.

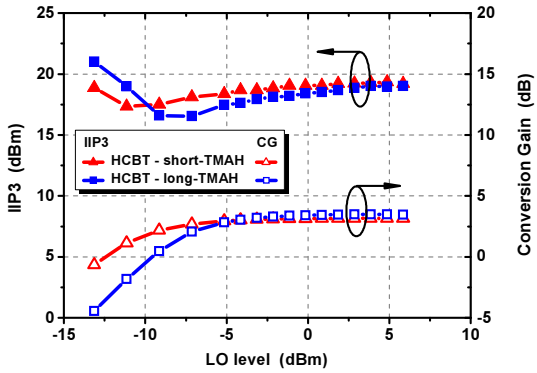


Figure 6. Measured input 3rd order intercept point (IIP3) and conversion gain (CG) vs. LO power level (P_{LO}) of the mixers with R_E = 20 Ω fabricated with the short- and long-TMAH HCBTs at the mixer current of 50 mA. Measurement setup of the two-tone test: P_{RF} = -10 dBm, f_{RF} = 900 MHz, f_{IF} = 20 MHz, Δf = 2 MHz, V_{CC} = 5 V.

IV. HCBT MIXER MEASUREMENTS

The fabricated mixers are measured on-wafer by using the multi-contact probes. The measurement setup is shown in Fig. 5. The used probes have in total 4 ground-signal-ground RF ports connected to the signal pads and several DC probes for the power supply, reference voltages and current connections. Both input ports (RF and LO) are fully differential but they are driven by the single-ended signal with another input grounded. Two RF signal generators are used for the RF port driving, since a two-tone test is required for linearity measurements. The signals are added together by using a power combiner. The mixer current is adjustable via a current source (I_{SET}) connected to the current mirror input. Additional mixer's DC operating point tuning for the optimal performance can be achieved by DC voltages connected to the DC probes.

The open-collector IF outputs are connected to a balun transformer. It is an off-chip component mounted on a small PCB designed for the measurement purpose. The balun is an 8:1 impedance transformer used to convert differential IF outputs to single-ended. When loaded with 50 Ω, the balun present a 400 Ω load to the mixer. The open-collector outputs are biased through the two inductors also placed on the external PCB with the power supply voltage V_{CC} = 5 V. They are required since the maximum DC current of the balun is specified not to exceed the 30 mA and the total mixer current is larger. For the smaller mixer current consumption, the bias could be set through the center pin of the balun's primary winding. The output power is measured by a spectrum analyzer connected to the secondary winding of the balun.

The CG and IIP3 of the examined mixers are measured at 900 MHz RF (f_{RF}) frequency and -10 dBm input power (P_{RF}). The output frequency (f_{IF}) is 20 MHz and the LO frequency (f_{LO}) is set accordingly. The frequency spacing (Δf) used for the two-tone testing is 2 MHz. The used power drive of the LO buffer (P_{LO}) is 0 dBm. The power combiner loss, and losses due to cables used for the RF and the LO port driving are the only losses included in the results deembedding.

The LO buffer circuit function is verified with the measurement of the IIP3 and CG dependence on the LO

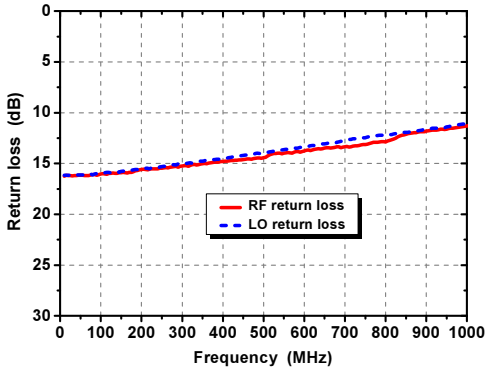


Figure 7. Measured return loss at RF and LO ports of the mixer with R_E = 20 Ω fabricated with the long-TMAH HCBTs at the mixer current of 50 mA.

power level shown in Fig. 6. Results are shown for the mixers with R_E = 20 Ω fabricated with both short- and long-TMAH HCBTs. The IIP3 and CG values are constant around P_{LO} = 0 dB, which is the standard LO drive in the commercial mixers, while small decrease of CG is observed for the LO power levels smaller than -5 dBm. Since the input impedances are designed to be 50 Ω, there is no need for the input matching networks at both RF and LO ports. The return loss measurements, necessary for the input impedances testing, are done by using the vector network analyzer. The measured return loss is less than 10 dB in both mixer types and results of the mixer with R_E = 20 Ω fabricated with the long-TMAH HCBTs is shown in Fig. 7. Hence, more than 90 % of the incident power coming from the RF signal generators is absorbed by the mixer ports in frequency range 10 MHz to 1 GHz.

The measured IIP3 and CG dependence on the mixer current (I_{MIX}) (without the LO buffer and bias circuitry current consumption) of mixers fabricated with both short- and long-TMAH HCBT types are shown in Fig. 8. The total power consumption is 425 mW at a mixer current of 40 mA, including the LO buffer and bias circuitry consumption. The highest IIP3 value of 20.2 dBm is obtained in the short-TMAH HCBT mixer at the mixer current of 45 mA. The mixer with the long-TMAH HCBTs has the IIP3 of 19.6 dBm at the same mixer

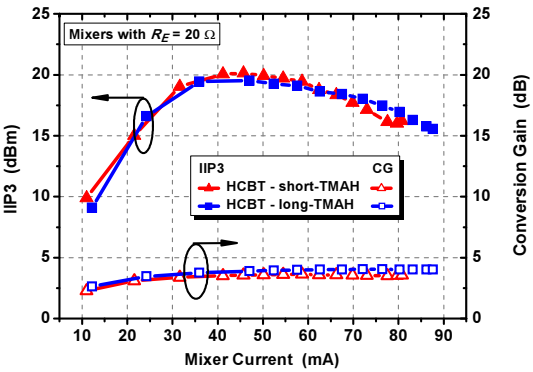


Figure 8. Measured input 3rd order intercept point (IIP3) and conversion gain (CG) vs. mixer current (I_{MIX}) of the mixers with R_E = 20 Ω fabricated with the short- and long-TMAH HCBTs. Measurement setup of the two-tone test: P_{RF} = -10 dBm, P_{LO} = 0 dBm, f_{RF} = 900 MHz, f_{IF} = 20 MHz, Δf = 2 MHz, V_{CC} = 5 V.

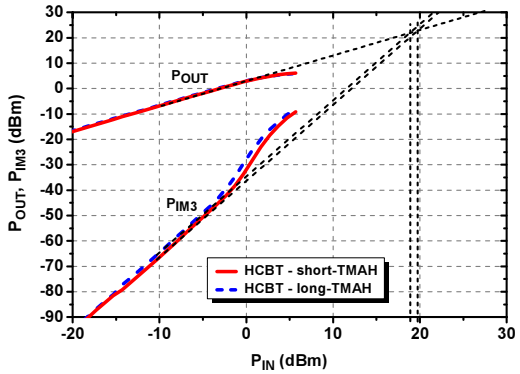


Figure 9. Measured output power (P_{OUT}) and 3rd order intermodulation distortion power (P_{IM3}) vs. input power (P_{IN}) of the mixers with R_E =20 Ω fabricated with the short- and long-TMAH HCBTs at the mixer current of 50 mA. Measurement setup: f_{RF} = 900 MHz, f_{IF} = 20 MHz, P_{LO} = 0 dBm, V_{CC} = 5 V.

current. The shape of the linearity dependence curve on mixer current is similar in both HCBT mixers. On the other hand, CG is higher by around 0.4 dB in the case of mixer with the long-TMAH HCBTs and has a value of 4 dB at the mixer current of 55 mA. The measured IIP3 values of both mixer types are smaller in comparison to the previously reported high-linearity mixers [7], where IIP3 = 23.8 dBm, since those are fabricated with the optimized *n*-hill collector HCBTs. In this work, the mixers are designed with the CMOS *n*-well region as *n*-collector HCBTs that are fabricated with only 2 additional lithography masks, unlike 3 masks used for optimized *n*-hill collector HCBTs.

The output and 3rd order intermodulation distortion power versus input power of both mixer types are shown in Fig. 9. The IIP3 values can be observed as theoretical intercept point of fundamental tone (P_{OUT}) and 3rd order intermodulation tone (P_{IM3}), while the CG values correspond to the difference in P_{OUT} and P_{IN}. The IIP3 and CG start to degrade for P_{IN} above 0 dBm where the large signal effects occur. They are characterized by the 1 dB compression point (P_{1dB}) which is defined as the power of input signal where the power gain is degraded by 1 dB. The measured P_{1dB} values are 2 dBm and 2.7 dBm in both mixers with the short- and long-TMAH HCBTs, respectively.

The similar values of IIP3 and CG in both short- and long-TMAH mixers (Figs. 8 and 9) suggest that the TMAH etching time impact on the mixer performance is relatively small. The emitter degeneration resistor (R_E = 20 Ω), necessary for the high-linearity performance, has a great impact on the mixer characteristics, according to (1) and (2). Hence, the differences in HCBT's TMAH etching times and their impact are masked by the R_E. In order to gain a deeper insight into the effect of TMAH etching time, the mixers without emitter degeneration resistors are also fabricated and the IIP3 and CG dependence on the mixer current (I_{MIX}) are measured. They have identical LO buffers as the mixers with R_E. The results of both short- and long-TMAH HCBT mixers are shown in Fig. 10. The mixers without degeneration have larger CG values, but they are less linear, as expected, in comparison to the ones with degeneration. Since there is

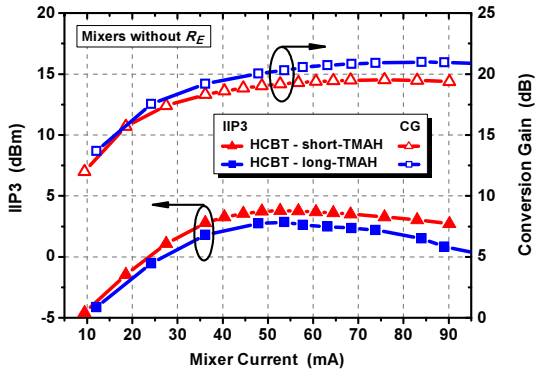


Figure 10. Measured input 3rd order intercept point (IIP3) and conversion gain (CG) vs. mixer current (I_{MIX}) of the mixers without R_E fabricated with the short- and long-TMAH HCBTs. Measurement setup of the two-tone test: P_{RF} = -10 dBm, P_{LO} = 0 dBm, f_{RF} = 900 MHz, f_{IF} = 20 MHz, Δf = 2 MHz, V_{CC} = 5 V.

no R_E, the measured mixer characteristics are more dependent on the HCBT transistor characteristics. The long-TMAH HCBTs have smaller I_B at the same V_{BE} in comparison to the short-TMAH HCBTs. It results in a better current mirroring meaning that for the same reference current (I_{SET}), total mixer current (I_{MIX}) is higher in the case of the long-TMAH mixer. For I_{SET} of 5 mA, the short-TMAH mixers have I_{MIX} around 44 mA and the long-TMAH mixers have I_{MIX} around 58 mA. Considering IIP3 and CG, the differences in the measured values between two mixer types are higher in comparison to the mixers with R_E = 20 Ω, but they are still relatively small. The difference in CG is ΔCG=1.1 dB and in IIP3 is ΔIIP3 = 1.0 dB at the mixer current of 50 mA.

V. CONLUSION

The impact of the TMAH etching time, as an important HCBT technology fabrication step, on the performance of the high-linearity mixers is investigated. The TMAH time variation results in the different polysilicon emitter thicknesses. The HCBT with thinner emitter polysilicon, as a result of the long-TMAH etching time, has a smaller I_B, higher β and f_T. The peak IIP3 and CG values obtained in the mixers fabricated in this lowest-cost HCBT technology are IIP3 = 20.2 dBm and CG = 4 dB. The differences in IIP3 and CG of mixers with HCBTs with 2 TMAH etching times at I_{MIX} = 50 mA are 0.6 dB and 0.4 dB respectively. The additional measurements on the mixers without emitter degeneration resistor R_E also showed small differences in CG and IIP3 (ΔCG=1.1 dB and ΔIIP3 = 1.0 dB), suggesting that the differences in electrical characteristics at transistor level, have a relatively small impact at the circuit level in the case of high-linearity mixers. Furthermore, the overall mixer performance demonstrates the suitability of the lowest-cost HCBT technology with CMOS *n*-well region as *n*-collector for the wireless communication market.

REFERENCES

[1] M. Feng, S.-C. Shen, D.C. Caruth, J.-J. Huan, "Device technologies for RF Front-End circuits in next-generation wireless communications," Proceedings of the IEEE, vol. 92, pp. 354-375, Feb 2004.

[2] P. Deixler *et al.*, "QUBiC4G: a f_T/f_{max} = 70/100 GHz 0.25 µm low power SiGe-BiCMOS production technology with high quality passives for 12.5 Gb/s optical networking and emerging wireless applications up to 20 GHz," in *Proc. BCTM*, 2002, pp. 201-204.

[3] H. S. Bennet *et al.*, "Device and Technology Evolution for Si-based RF Integrated Circuits," *IEEE Trans. Electron. Devices*, vol. 52, pp. 1235-1258, Jul 2005.

[4] T. Suligoj *et al.*, "Horizontal Current Bipolar Transistor (HCBT) with a Single Polysilicon Region for Improved High-Frequency Performance of BiCMOS ICs," *IEEE Electron Device Lett.*, vol. 31, pp 534-536, Jun 2010.

[5] M. Koričić, J. Žilak, T. Suligoj, "Double-Emitter Reduced-Surface-Field Horizontal Current Bipolar Transistor with 36 V Breakdown Integrated in BiCMOS at Zero-Cost," *IEEE Electron. Device Lett.*, vol. 36, pp. 90 – 92, Feb 2015.

[6] B. Razavi, RF Microelectronics, 2nd ed., New York, USA, Paerson Education, Inc., 2012.

[7] J. Žilak, M. Koričić, H. Mochizuki, S. Morita, T. Suligoj, "Impact of Emitter Interface Treatment on the Horizontal Current Bipolar Transistor (HCBT) Characteristics and RF Circuit Performance," in *Proc. BCTM*, 2015, pp. 31-34.

[8] M. Koričić, "Horizontal Current Bipolar Transistor Structures for Integration with CMOS Technology", doctoral thesis, FER, University of Zagreb, Croatia, 2008.

[9] B. Gilbert, "A precise four-quadrant multiplier with subnanosecond response", *IEEE Journal of Solid-State Circuits*, vol. 3, pp. 365-373, 1968.

[10] M. B. Judson, "Low-voltage front-end circuits: SA601, SA602," Philiphs Semiconductors, Aplication Note AN1777, Aug. 1997.

[11] J. Rogers and C. Plett, "Mixers," in *Radio Frequency Integrated Circuit Design*, Boston, MA, USA: Artech Hounce Inc., 2003.

Fully-integrated Voltage Controlled Oscillator in Low-cost HCBT Technology

M. Koričić *, J. Žilak *, H. Mochizuki **, S. Morita ** and T. Suligoj*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Department of Electronics, Microelectronics, Computing and Intelligent Systems, Micro and Nano Electronics Laboratory, Zagreb, Croatia
** Asahi Kasei Microdevices Co. 5-4960. Nobeoka, Miyazaki, 882-0031, Japan
marko.koricic@fer.hr

Abstract - Design of cross-coupled voltage controlled oscillator in low-cost HCBT technology is presented. Beside the low-complexity front-end devices, only 2 metal layers are used and the passives are implemented in the available on-chip structures. Varactors are fabricated as pn-junctions by using the ion implantation from the technology. Symmetric inductors are fabricated by using the topmost metal layer. Since only 2 aluminum metal layers are available, small thickness of the aluminum layer and proximity of the silicon substrate limit the inductor quality factor. Varactor and inductor models for circuit simulations are developed by using the device and electromagnetic simulations, respectively, and are compared to measured characteristics of fabricated devices.

I. INTRODUCTION

Integration of bipolar transistors to the baseline coarser lithography Complementary-Metal-Oxide-Semiconductor (CMOS) processes has become an attractive technological approach for the extension of the technology applications, without a significant increase of fabrication cost [1]. A typical application is in the wireless communication circuits where benefits are gained by the improved high frequency and noise performance of bipolar transistors. Such integration should be cost-effective to meet the demands of very cost-sensitive market. Horizontal current bipolar transistor (HCBT) is integrated with standard 180 nm CMOS technology with only 2 or 3 additional lithography masks and a small number of additional processing steps [2], resulting in the very low-cost BiCMOS technology. At the same time, electrical performance is comparable to more expensive vertical bipolar transistors with implanted base. Furthermore, high-voltage devices are added to the technology at zero-cost [3], [4], extending the portfolio of the available devices.

RF mixer is so far the only published RF circuit implemented in HCBT technology [5]. In order to fabricate other RF front-end circuits, high quality passive components are needed, which: (i) are not available in the technology in its experimental and development phase, and (ii) increase the fabrication cost. In this paper, the design of fully integrated cross-coupled voltage controlled oscillator (VCO) is presented. The goal is to design VCO at the lowest possible cost by using only available front-end process steps needed for HCBT fabrication, i.e., in the bipolar-only version of the HCBT technology, with only the first two metal layers of standard CMOS interconnect processing. In this paper, design, modeling and fabrication of passive components in HCBT technology are done for the first time.

This work was supported by the Croatian Science Foundation under contract no. 9006.

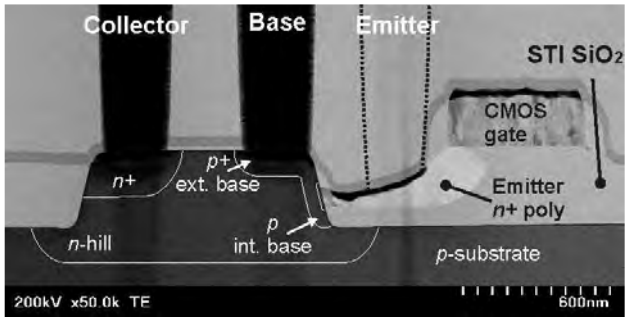


Figure 1. TEM cross-section of fabricated HCBT.

II. VCO DESIGN IN HCBT TECHNOLOGY

VCO is fabricated by using only the process steps for the fabrication of HCBT devices. TEM cross-section of fabricated HCBT is shown in Fig. 1. with marked transistor regions. Details about the process are given elsewhere [2]. Steep collector profile optimized for the performance of high-speed HCBT devices [6] is used for the fabrication of VCO.

Electrical schematic of the designed VCO is presented in Fig. 2a and the photograph of fabricated chip in Fig. 2b. The core of the VCO consists of the cross-coupled differential pair biased by the tail current source (I_0) realized by the simple current mirror. Bias current (I_{BIAS}) is supplied off-chip. Resonant *LC*-tank of the circuit consists of a symmetric inductor and varactors which are fabricated by the front-end process steps available in the technology. Output buffers are implemented as two stages of the emitter followers in order to reduce the parasitic capacitance at the collectors of T_1 and T_2 and to provide capability of driving the output pad and the input impedance of the spectrum analyzer. Also, a separate biasing is used in order to monitor the power dissipation of the VCO core.

In order to sustain the oscillations, the cross-coupled pair generates negative conductance at the collectors of T_1

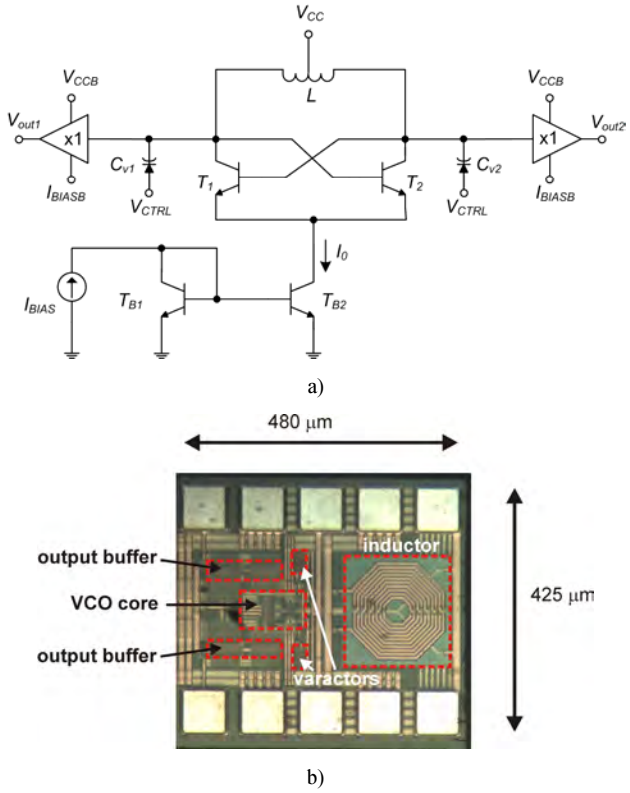


Figure 2. a) Electrical schematic of designed cross-coupled VCO. b) Chip photograph.

and T_2, which replenishes the energy lost in the resistive component of *LC*-tank. The transconductance of transistors (g_m) should be high enough to satisfy the condition:

$$g_m \geq \frac{1}{R_p} \tag{1}$$

where R_p is the equivalent parallel resistance of the inductor. In practice, the g_m is chosen higher than the marginal case in (1), to allow complete current steering and higher output voltage swings.

As a design constraints we choose: a) tail bias current I_0 smaller than 2 mA in order to have power dissipation less than 10 mW including bias circuitry and excluding output buffers, b) output voltage swing around 0.6 Vpp in order to have complete current steering in the cross-coupled pair, c) central frequency f_c=2.5 GHz and d) tuning range of 10% of the central frequency, i.e., Δf=250 MHz.

In the case of complete current steering in the differential pair, output voltage swing is [7]:

$$V_{out,pp} = \frac{4}{\pi} I_0 R_p \tag{2}$$

where R_p is equivalent parallel resistance of the resonant tank. In the first approximation, it is defined by the inductor parameters at the frequency of oscillation:

$$R_p = 2\pi f_{osc} L \left(\frac{1}{Q} + Q \right) \tag{3}$$

TABLE I. SIMULATED ELECTRICAL CHARACTERISTICS OF INTEGRATED INDUCTORS WITH DIFFERENT GEOMETRICAL PARAMETERS

	Inductor 1	Inductor 2
Wire width, W [µm]	12	5.5
Wire spacing, S [µm]	3	0.8
Number of turns, N	3	10
Output diameter, D [µm]	150	170
$L@f_{osc}$, [nH]	1.12	10
$Q@f_{osc}$	1.487	1
Q_{max}	3.6 (@10 GHz)	1 (@2.3 GHz)
$L@f(Q_{max})$, [nH]	1.293	10
$R_p@f_{osc}$, [Ω]	26.2	157
$I_0@V_{outpp}$=0.6 Vpp, [mA]	12.4	1.5

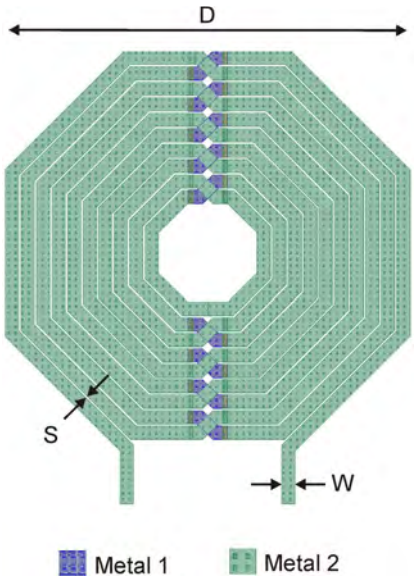


Figure 3. Layout of the symmetric inductor used in the design of VCO.

where f_{osc} is the frequency of oscillation, L the inductance and Q the inductor quality factor. A high Q of the inductor is desirable due to smaller losses, which reflects to smaller power consumption. Furthermore, a phase noise, an important oscillator parameter, is proportional to $1/Q^2$ and it benefits from usage of high quality inductors.

A. Inductor design and modeling

Symmetric inductors shown in Fig. 3 are used in the *LC*-tank of VCO. Inductors are fabricated by using only 2 aluminum layers of 0.5 µm thickness. Since the sheet resistance of these layers is rather high, wide metal lines are used in order to minimize wire resistance and improve *Q*-factor. On the other hand, this increases the capacitive coupling to substrate limiting the maximum achievable Q due to substrate losses. Additionally, wider metal lines require larger spacing between the lines due to reliability of the interconnect processing, yielding a smaller mutual and hence the overall inductance. Furthermore, total inductor area, which is usually specified in terms of an output diameter, increases and the achievable inductance is limited. Results obtained by electro-magnetic (EM) simulations for 2 inductors realized with wide and moderately wide metal lines and with comparable output diameter (D) are shown in Table 1. The equivalent parallel resistance (R_p) is calculated at a target oscillation frequency (f_{osc}=2.5 GHz) from (3) and the tail bias current (I_0) for desired output voltage swing V_{outpp}=0.6 Vpp is

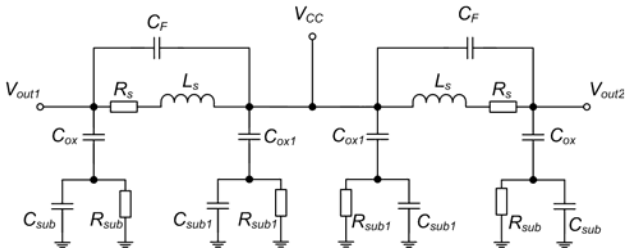


Figure 4. Compact model of the symmetric inductor extracted for the use in the circuit simulation.

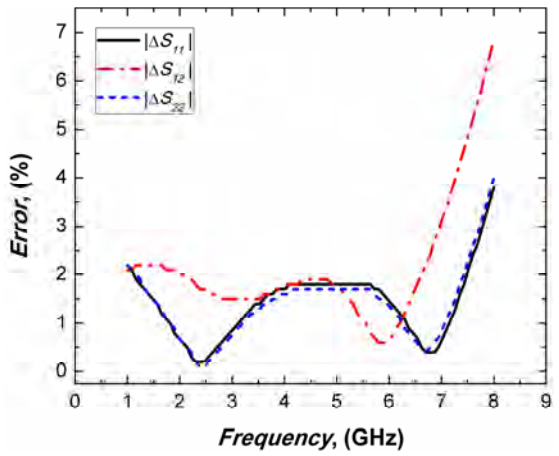


Figure 5. Results of the optimization used for the extraction of inductor model parameters from Fig. 3.

TABLE II. PARAMETERS OF THE INDUCTOR COMPACT MODEL OBTAINED FROM THE OPTIMIZATION

C_F, [fF]	L_s, [nH]	R_s, [Ω]	C_{ox}, [fF]	C_{sub}, [fF]
203	4.44	36	44.6	1.2e-9
R_{sub}, [Ω]	C_{ox1}, [fF]	C_{sub1}, [fF]	R_{sub1}, [Ω]	
602	119	3.1e-6	1506	

calculated by combining (2) and (3). It can be seen that it is difficult to obtain a high Q inductor at frequency as low as 2.5 GHz. In the case of the inductor 1, Q=1.487, which is higher compared to the inductor 2. However, due to a larger width of the metal lines (W), a smaller length of the wire with comparable D, results in a much smaller inductance of the inductor 1. This translates to much higher I_0 and increased power dissipation. The inductor 2 is designed to have the peak Q-factor around the targeted f_{osc}=2.5 GHz. Due to the higher inductance, the I_0 is within the specification of the power dissipation for the given output voltage swing. Therefore, the inductor 2 is used for the design of VCO.

It should be noted that in the case of the inductor 1, a moderately high Q of 3.6 is obtained at f=10 GHz. Since the higher Q can be obtained at higher frequencies, VCO can be designed to operate at a higher frequency and then a frequency divider can be used to obtain the specified lower frequency. For example, if the inductor 1 is used for the design of VCO operating at 10 GHz and V_{outpp}=0.6 Vpp from (2) and (3) we can calculate the required tail current to be I_0=1.5 mA. Furthermore, the phase noise performance would be improved as explained earlier. However, design of VCO at f=2.5 GHz without

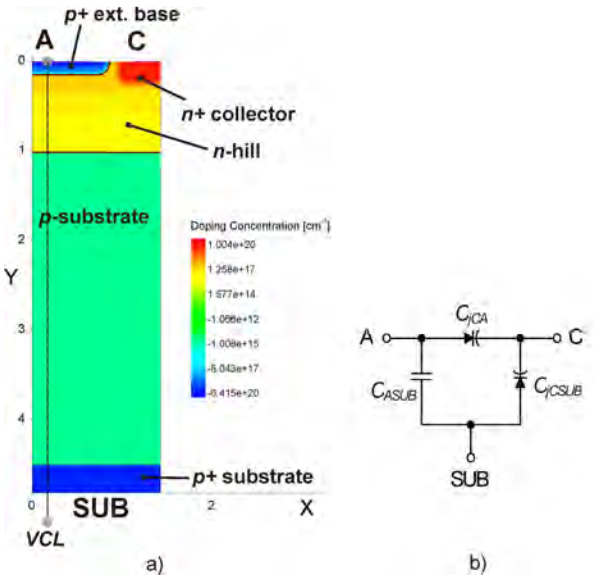


Figure 6. Varactor: a) 2D device simulation model, b) simple π-model of capacitances. Due to symmetry, half of the structure is simulated.

frequency divider is chosen to put a focus only on the performance of VCO.

S-parameters obtained by the EM simulations are used to extract a compact model of the inductor, which is shown in Fig. 4. Optimization goals are defined as absolute differences between S_{11}, S_{12} and S_{22} obtained from the EM simulation and the ones obtained from the compact model from Fig. 4 in the frequency range from 1 GHz to 8 GHz. Relative difference after the optimization is shown in Fig. 5 showing the error lower than 2 % in the frequency band of interest around 2.5 GHz, indicating the suitability of the model for the circuit simulation. The parameters of the compact model obtained by the optimization are listed in Table 2.

B. *Varactor design and modeling*

In order to tune the frequency of the VCO, variable capacitors are needed. Frequency of oscillation is:

$$f_{osc} = \frac{1}{2\pi\sqrt{L(C_p + C_v)}} \tag{4}$$

where C_v is variable capacitance and C_p is overall parasitic capacitance at the output node, including the collector-base, the base-emitter and the collector-substrate capacitance of the cross-coupled pair as well as the buffer input capacitance, the inductor parasitic capacitance and interconnect capacitance (see Fig. 2a). Equation (4) is used to set the value of the varactor minimum and maximum capacitance for the target central frequency (fc) and the tuning range (Δf). Frequency tuning is done by reverse biased diodes, which are fabricated by using the ion implantation for the fabrication of the n-collector (n-hill in Fig. 1) and the p+ extrinsic base of HCBT. Electrical characteristics and the circuit model of the diode are obtained by the TCAD device simulations. Doping profiles under the extrinsic base, which are measured by Secondary Ion Mass Spectrometry (SIMS) are loaded in the device simulator. Cross-section of the simulated

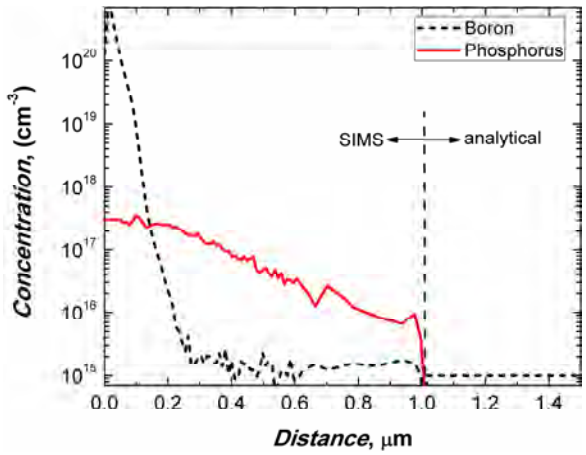


Figure 7. Measured SIMS profile used in the device simulation. Doping profile at the cutline VCL from Fig. 5.

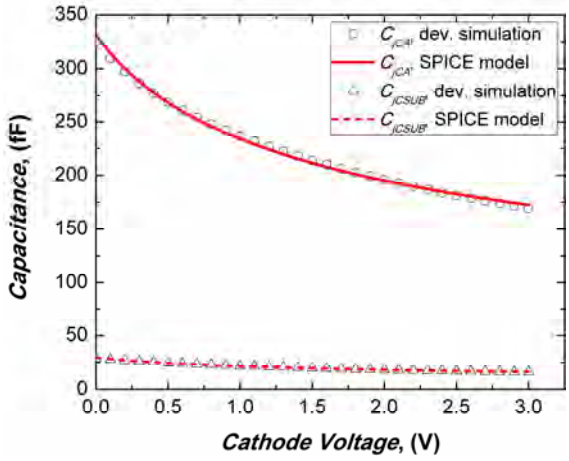


Figure 8. Comparison of varactor CV-characteristics obtained by the device simulations and from the extracted SPICE model.

varactor is shown in Fig. 6a and the profiles used in the simulations obtained at the cutline VCL are shown in Fig. 7. AC simulations are performed at different bias points in order to obtain the capacitance-voltage (CV) characteristics. Capacitances of the structure from Fig. 6a can be represented by a simple π-network shown in Fig. 6b. It consists of two junction capacitances: cathode-anode capacitance (C_{jCA}) and cathode-substrate capacitance (C_{jCSUB}) as well as parasitic anode-substrate capacitance (C_{ASUB}), which is shown to be negligible from the simulation results. In SPICE-like programs the *pn*-junction capacitance in is modeled by:

$$C_{jB} = \frac{C_{jB0}}{\left(1+\frac{V_{app}}{V_{bi}}\right)^{M}} \qquad (5)$$

where C_{jB0} is zero-bias capacitance, V_{bi} built-in potential, M grading coefficient and V_{app} voltage applied between the cathode and the anode. SPICE model parameters are extracted from the device simulation results which are then scaled in the circuit simulations to obtain target f_c and Δf. Finally, the actual physical design of varactor layout is carried out with the area defined by the circuit simulation results. Fitting of the SPICE model to the device simulation after scaling is shown in Fig. 8 The quality factor of varactors (Q_C) is important for total quality factor of resonant *LC*-tank. The total quality factor is:

$$\frac{1}{Q_{TOT}} = \frac{1}{Q_L} + \frac{1}{Q_C} \qquad (6)$$

where Q_L and Q_C are quality factors of the inductor and the varactor, respectively. Q_C depends on the series resistances and hence is difficult to model by the device simulations. In the physical design of the circuit layout, *p*+ and *n*+ regions from Fig. 6a are kept as close as possible with the constraint not to have too low breakdown voltage between the cathode and the anode. The cathode and the anode regions are realized as a long thin slices, which yields slightly larger series resistances (i.e., lower Q_C) compared to the case when *n*+ rings are formed around smaller *p*+ square-shaped regions. The structure with slices is chosen because it is dominated by the cross-section from Fig. 6a and is basically 2D structure, which is better predicted by the 2D device simulations.

TABLE III. TRANSISTOR SIZES USED IN VCO DESIGN

Transistor	T_{B1}	T_{B2}	T_1	T_2
No. of unit transistors	1	8	8	8

TABLE IV. SIMULATED ELECTRICAL CHARACTERISTICS OF VCO

Parasitic capacitance at the output	C_{ps}=0	C_{ps}=80 fF
V_{CC}, [V]	3.3	3.3
I_{BIAS}, [μA]	200	200
I_0, [mA]	1.6	1.6
Minimum frequency, f_{min}, [GHz]	2.481	2.292
Maximum frequency, f_{max}, [GHz]	2.818	2.575
Tuning range, Δf, [MHz]	337	283
Central frequency, f_c, [GHz]	2.650	2.434
Output signal power, P_{out}, [dBm]	0.81	-0.47
Phase noise @100kHz offset, *PN* [dBc/Hz]	-73.5	-73.2

C. VCO design

VCO is designed by using the inductor and the varactor models described in the previous sections. The size of the varactors is scaled in order to obtain the desired oscillation frequency (f_{osc}) and the tuning range (Δf). HCBT model used in the design is previously developed standard Gummel-Poon model, which is already used in the design of RF mixers [5]. Unit size transistors are used and are connected in parallel depending on the required collector current. Transistor sizes in terms of number of parallel unit transistors are given in Table 3. The bias current supplied of the chip in Fig. 2a is I_{BAS}=200 μA. The size ratio of T_{B1} and T_{B2} sets the tail bias current I_0=1.6 mA. The sizes of the transistors T_1 and T_2 are chosen the same as in the tail current source in order to avoid large current density when current is completely steered in one of the branches. The results of simulated electrical performance of VCO are summarized in Table 4. Table includes the column with assumed parasitic capacitance at the output node which is used in simulations (C_{ps}). It includes the wire capacitance and all possible discrepancies in capacitance in transistor and varactor model. Its value is set to fit the measured data.

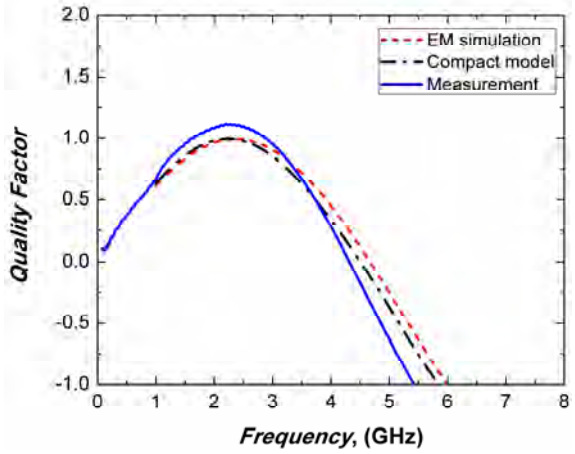


Figure 9. Comparison of measured and modeled inductor Q-factor.

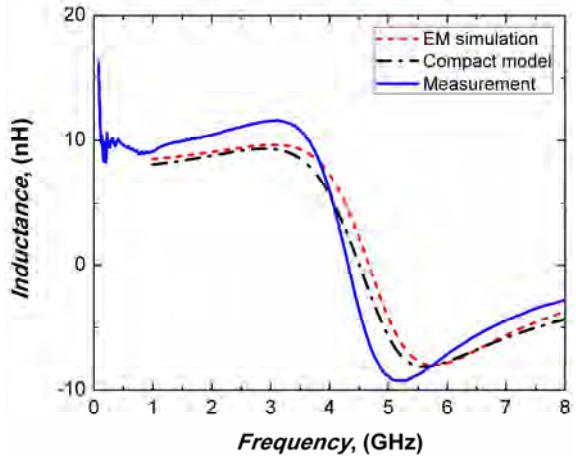


Figure 10. Comparison of measured and modeled inductance of the inductor.

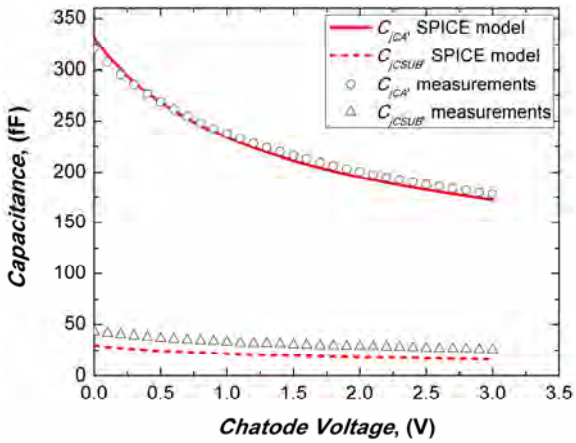


Figure 11. Comparison of measured and modeled CV-characteristics of the varactor

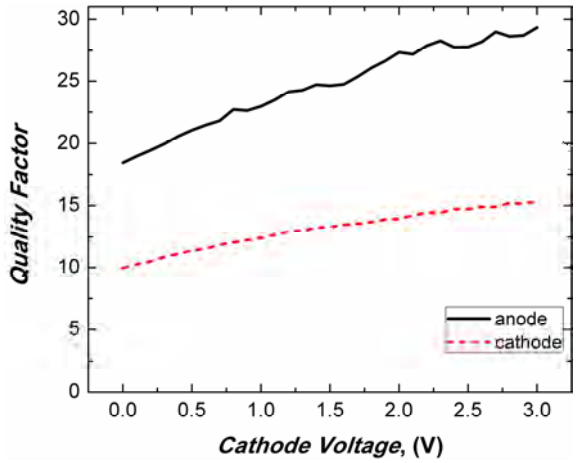


Figure 12. Measured quality factors of varactor at the anode and the cathode terminals at frequncy f=2.5GHz.

III. MEASUREMENT RESULTS

Comparison of measured inductor Q-factor and inductance up to 8 GHz with the ones obtained by EM simulations and compact model used in the design phase are shown in Figs. 9 and 10, respectively. Good fitting between EM simulation and the compact model can be observed as expected from the optimization results shown in Fig. 5. Slightly higher value of inductance and Q-factor is obtained in measurements. De-embedding of the measurement results is done by using open-pad structure only, and portion of the wire connecting inductor to the pad adds small series inductance. Furthermore, the stack of the substrate, metal and dielectric layers and their electrical parameters which are set as the input to the EM simulator might introduce some error. Nevertheless, a satisfactory fit of the measured by modeled electrical characteristics is achieved.

Comparison of the CV characteristic obtained by the SPICE model used in the simulations and the measured S-parameters of fabricated varactor is shown in Fig. 11. Good fitting of the measured C_{jCA} by the model is accomplished. The measured C_{jCSUB} is underestimated in the simulation due to the underestimated peripheral component, which is not used in the 2D device simulation structure employed for the model development (see Fig. 6a). On the other hand, the voltage dependency is well reproduced and model can be easily calibrated by adjusting the parameter C_{jB0} in (5). The extracted quality factor of the varactor (Q_C) from the measured S-parameters at 2.5 GHz is shown in Fig. 12. The Q_C measured at the cathode plays a more significant role since it is connected to the output node. Q_C above 10 is achieved for all applied control voltages, which is rather low value, but according to (6), LC-tank quality factor (Q_{TOT}) is still dominated by the inductor in our design.

Measured output signal power and frequency of oscillation dependency on control voltage (V_{CTRL}) at I_{BIAS}=200 µA of fabricated VCO are shown in Fig. 13 and the results are summarized in Table 5. Measurements are made directly on-chip by using multi-contact probes. Maximum and minimum f_{osc} for V_{CTRL} of 0 V and 3 V are 2.53 GHz and 2.295 GHz, respectively, resulting in the tuning range (Δf) of 235 MHz, which is 9.75 % of the central frequency (f_c=2.41 GHz). Measurement results fit well to the results of the simulated VCO with assumed parasitic capacitance C_{ps}=80 fF reported in Table 4. The assumed C_{ps} takes into account the discrepancy of C_{jCSUB} in Fig. 11 as well as the wire capacitance at the output node. The measured output power shows discrepancy compared to the simulation results, which is attributed to the actual buffer output impedance driving the 50 Ω input

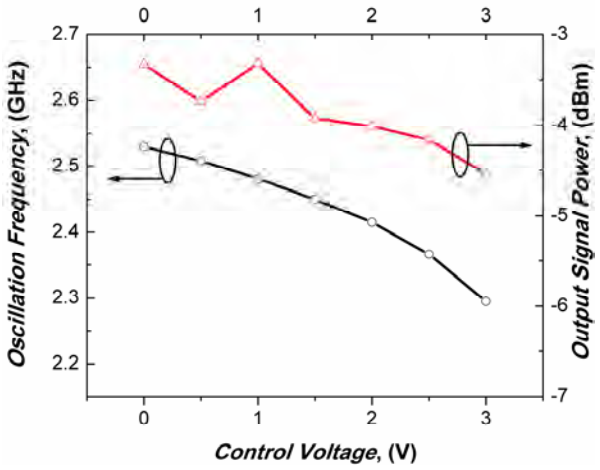


Figure 13. Measured oscillation frequency and output signal power dependency on control voltage of fabricated VCO.

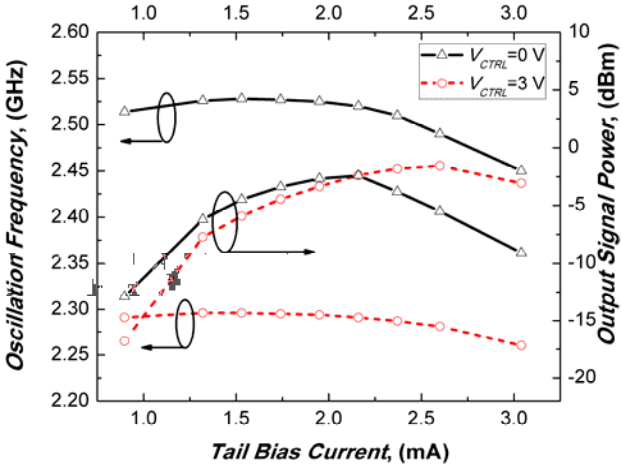


Figure 14. Measured oscillation frequency and output signal power dependency on tail bias current of fabricated VCO.

impedance of the spectrum analyzer as well as to the reduction of the Q_{TOT} by the resistance of the interconnect wires that is not taken into account in the circuit design. Reduction of the P_{out} for larger V_{CTRL} can be also partly attributed to the reduction of Q_{TOT} of the *LC*-tank. V_{CTRL} is applied at the anode of the varactor, and since the cathode voltage is approximately V_{CC}, the cathode-anode voltage is decreased when V_{CTRL} is increased resulting in the decrease in Q_C (see Fig. 12). Hence, according to (6), Q_{TOT} decreases. Measured P_{out} and f_{osc} dependencies on the tail bias current (I_0) of the cross-coupled pair for two marginal control voltages (V_{CTRL}) which define the tuning range, are shown in Fig 14. Frequency tuning range and P_{out} are relatively constant for the I_0 between 1.5 mA and 2.5 mA. P_{out} rises for larger I_0, which agrees with (2), and then falls-off when transistors enter the high current regime.

Direct phase noise measurement method by using the spectrum analyzer, which is available at the moment, is not suitable for characterization of free-running VCOs [8]. Therefore, the simulation results are given in this paper. Results show the phase noise of -73.2 dBc/Hz at the

TABLE V. MEASURED ELECTRICAL CHARACTERISTICS OF VCO

V_{CC}, *[V]*	3.3
I_{BIAS}, [μA]	200
I_0, [mA]	1.74
Minimum frequency, f_{min}, [GHz]	2.295
Maximum frequency, f_{max}, [GHz]	2.53
Tuning range, Δf, [MHz]	283
Central frequency, f_c, [GHz]	2.434
Output signal power, P_{out}, [dBm]	-3.3
Phase noise @100kHz offset, *PN* [dBc/Hz]	Not available

100 kHz offset from the carrier frequency (see result for C_p=80 fF in Table 4), which is rather high value. Since the phase noise is proportional to $1/Q_{TOT}^2$ it can mainly be improved by incorporating high-*Q* inductors and varactors.

IV. CONLUSION

Design of the cross-coupled voltage controlled oscillator in low-cost HCBT technology is presented. Varactors and inductors are designed and modeled by device and electromagnetic simulations and fabricated by using the available process steps without altering the technology. Good agreement between modeled and measured electrical characteristics of fabricated passive components is accomplished. It is shown that the VCO performance is limited by the low *Q*-factor of the inductor, which is fabricated in the 2nd aluminum metal layer of standard CMOS interconnect.

REFERENCES

[1] H. S. Bennett, R. Brederlow, J. C. Costa, P. E. Cottrell, W. M. Huang, A. A. Immorlica, Jr., J.-E. Mueller, M. Racanelli, H. Shichijo, C. E. Weitzel, and B. Zhao, "Device and technology evolution for Si-based RF integrated circuits," IEEE Trans. Electron Devices, vol. 52, no. 7, pp. 1235–1258, Jul. 2005.

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

[3] M. Koričić, T. Suligoj, H. Mochizuki, S. Morita, K. Shinomura, and H. Imai, "Double-Emitter HCBT Structure—A High-Voltage Bipolar Transistor for BiCMOS Integration,", *IEEE Trans. Electron Devices*, vol. 59 , no. 12 pp. 3647 – 3650, Dec. 2012.

[4] M. Koričić, J. Žilak, T. Suligoj, "Double-Emitter Reduced-Surface-Field Horizontal Current Bipolar Transistor With 36 V Breakdown Integrated in BiCMOS at Zero Cost," *IEEE Electron Device Lett.*, vol. 36, no. 2, pp. 90-92, Feb. 2015.

[5] J. Žilak, M. Koričić, H. Mochizuki, S. Morita, T. Suligoj, "Impact of Emitter Interface Treatment on the Horizontal Current Bipolar Transistor (HCBT) Characteristics and RF Circuit Performance," in *Proc. BCTM*, 2015, pp. 31-34.

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

[7] Behzad Razavi, "Oscillators," in *RF Microelectronics*, 2nd edition, Boston, MA, USA, Prentice Hall, 2011.

[8] Keysight Technologies, "Phase Noise Measurement Solutions," Selection guide, USA, 5990-5729EN, Aug. 2014.

Variable-Gain Amplifier for Ultra-Low Voltage Applications in 130nm CMOS Technology

Daniel Arbet, Martin Kováč, Lukáš Nagy, Viera Stopjaková and Michal Šovčík
Department of IC Design and Test
Faculty of Electrical Engineering and Information Technology
Slovak University of Technology
Bratislava, Slovakia
e-mail: daniel.arbet@stuba.sk

***Abstract*—The paper deals with design and analysis of a variable-gain amplifier (VGA) working with a very low supply voltage, which is targeted for low-power applications. The proposed amplifier was designed using the bulk-driven approach, which is suitable for ultra-low voltage circuits. Since the power supply voltage is less than 0.6 V, there is no risk of latchup that is usually the main drawback of bulk-driven topologies. The proposed VGA was designed in 130 nm CMOS technology with the supply voltage of 0.4 V. The achieved results indicate that gain of the designed VGA can be varied from 0 dB to 18 dB. Therefore, it can be effectively used in the many applications such as automatic gain control loop with ultra-low value of supply voltage, where the dynamic range is the important parameter.**

I. Introduction

Advanced nanotechnologies enable very-large-scale integration and bring the opportunity to design ultra low-power analog and mixed-signal integrated systems. In these modern nanoscale CMOS processes, also the power supply voltage is continuously scaled down. Moreover, the threshold voltage of MOS devices does not decrease with the same slope as the supply voltage. Trends towards ultra-low value of power supply voltage can effectively increase battery life of portable electronics, biomedical implanted devices, hearing-aid devices, etc.

One of the most important building blocks of analog integrated circuits (IC) is a variable gain amplifier (VGA), which is used in many applications in order to stabilize the voltage amplitude of a signal at its output. VGA is usually employed in an automatic gain control (AGC) circuit to maximize the dynamic range of the whole system.

Since the proposed VGA is meant to be used in low-voltage applications, the possibility of using standard amplifier topologies is limited (because the supply voltage is usually less than 1 V). Standard VGA topologies are ussually based on the conventional differential structure [1]–[3]. Since there are four or more stacked transistors, these topologies require a high value of the supply voltage and it is not suitable for ultra-low voltage applications. One possible topology of a VGA suitable for a low supply voltage is based on the pseudo-differential difference amplifier (PDDA). PDDA topology can effectivelly increase the input and output voltage ranges. On the other hand, disadventages of the PDDA include high sensitivity to process and temperature variations (PVT) and low value of CMRR (Common-Mode Rejection Ratio) parameter due to a missing tail source. Therefore, ussually Common-Mode Feedback (CMFB) or Common-Mode Feedforward (CMFF) circuit is employed to stabilize the operational point and to increase CMRR of the PDDA [4], [5].

In the case when the PDDA is based on the two common-source amplifiers [4], the input voltage range is limited by the treshold voltage of the input transistors. In order to increase the input voltage range, a rail-to-rail input stage has to be used. For this purpose, some of unconventional design techniques is needed. In order to overcome this limitation and also to increase the input voltage range of VGA, so-called bulk-driven technique (MOS devices are controlled by bulk instead of gate) was used to design the proposed VGA. Bulk-driven approach has been employed to design a number of analog building blocks [7], [8].

In this paper, design of ultra low-voltage bulk-driven VGA based on pseudo-differential topology is presented and its main parameters are analysed. In Section II, the proposed topology of VGA is presented. Performed small-signal analysis of the VGA is then described in Section III. Achieved results are presented in Section IV. In Section V, the achieved parameters of the developed VGA are summarized and shortly discussed.

II. Proposed Variable Gain Amplifier

A. VGA general description

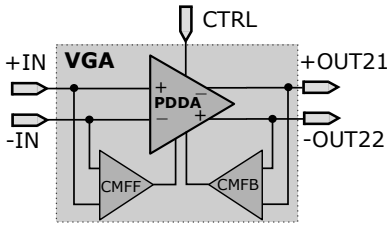


Fig. 1. Block diagram of the proposed VGA

Fig. 1 shows the block diagram of the proposed VGA. The main block representing the VGA core is a bulk-driven fully PDDA (FPDDA) with the gain control input terminal. FPDDA is based on two common-source amplifiers, where a bulk-driven input transistor is used. In order to stabilize the operational point and increase the CMRR of the proposed

VGA, CMFB and CMFF circuits were employed. To achieve good stability of the CMFB loop, frequency compensating capacitors have been used (not depicted in Fig. 1).

B. Proposed Pseudo-Differential VGA topology

Schematic diagram of the proposed ultra-low voltage VGA core circuit is depicted in Fig. 2. The VGA was designed in 130 nm CMOS technology and can reliably operate at the supply voltage of 0.4 V.

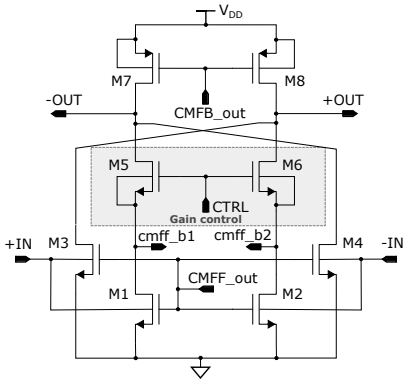


Fig. 2. Schematic diagram of the bulk-driven FPDDA-based VGA

The proposed topology is based on the cascode FPDDA, where bulk-driven (BD) input MOS transistors were used. Input BD transistors are used to obtain the rail-to-rail input voltage range. Unfortunately, bulk transconductance of a MOS transistor is given by the following expression $g_{mb} \approx 0.2g_m$, which means that the gain and gain-bandwidth product (GBW) of the proposed VGA will be decreased. On the other hand, the circuits designed by the bulk-driven approach are useful for ultra low-voltage and low power applications.

Generally, gain of the VGA can be varied by controlling its total conductance or the total output resistance. Thus, in our case, transistors M5 and M6 were employed to control the VGA gain. Voltage change at the VGA control terminal (CTRL) causes a change in current flowing through the input transistors M1 and M2, and thereby regulates their transconductance. Thus, changes in transconductance of the input transistors lead to variation of the total VGA gain. Therefore, the total gain of VGA is directly proportional to transconductance of the gain control transistors (g_{m5} and g_{m6}). Additionally, the gain control transistors together with the input transistors represent a cascode stage and therefore, g_{m5} and g_{m6} influence also the output resistance of VGA. Detailed small-signal analysis of the designed VGA is described in Section III.

C. CMFF circuit

In order to adjust the bias voltage (CMFF_out) of input transistors (M1-M4), a CMFF circuit depicted in Fig. 3 was used. Transistors M9 and M10 represent the input transistors of the CMFF circuit, which follow the signal at the input of VGA. In the case of differential input signal, there is no change in the current flowing trough the diode-connected transistor M11, and voltage at the CMFF output (voltage drop on diode-connected transistor M13) is stable. Difference between values of common-mode voltage at the VGA inputs will change the current through transistor M11. This current is mirrored by transistor M12 to transistor M13 and its voltage drop will change. This principle was used to regulate (gate) bias voltage for input transistors M1-M4 when the common-mode voltage at the inputs of VGA is changed due to PVT. Finally, voltage drop on the transistors M1 and M2 was used as bias voltage for input transistor in the CMFF circuit (M9 and M10). Thank to this biasing technique, the VGA sensitivity to common-mode voltage as well as PVT is reduced.

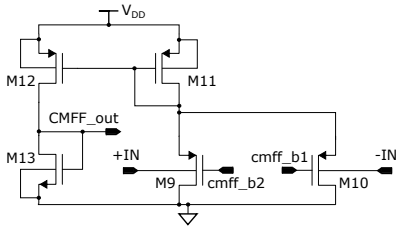


Fig. 3. Schematic diagram of CMFF circuit

D. CMFB circuit

In general, a CMFB is needed to prevent the output voltages from saturating to one of the power rails when the input common-mode voltage varies or there is a change in the circuit operating point due to VGA control voltage. A CMFB usually consist of a common–mode voltage detector and an error amplifier, as presented in [9]. In general, this approache suffers from chip area overhead in terms of resistors implementation and separation function of two main active blocks. In addition, potential incompatibility of conventional structures with ultra–low voltage application is increased (especially with the supply voltage value deep below 1 V), which results in frequently used low voltage realizations like current-based operation CMFB (CB–CMFB) or its improved version [10] combining the common–mode detection and gain functionality if comapared to many published common–mode detection only realizations [9].

The novel self-biased bulk driven CMFB circuit that combines common–mode voltage detection and error amplification capabilities is introduced (Fig. 4), and used as the most important part of the low-voltage VGA based on pseudo–differential structure. The main idea is slightly similar to CB–CMFB, where two differential pairs ($N2x, N3x$) with purposely degraded current sources ($N1x$) of the same polarity are connected in parallel but in our design, current mixing is accomplished in the output node $CMFB_out$ across two current mirrors ($N4x, N5x$). Degraded tail current sources (diode-connected transistors in this case, in opposite to CB–CMFB) was employed to improved entire common-mode gain of the VGA. We have to note that blue dashed feedback consisting of these two transistors can works as a negative feedback (across $N3x$) or a partially positive feedback (across $N4x, N5x$). Additionally, the output node

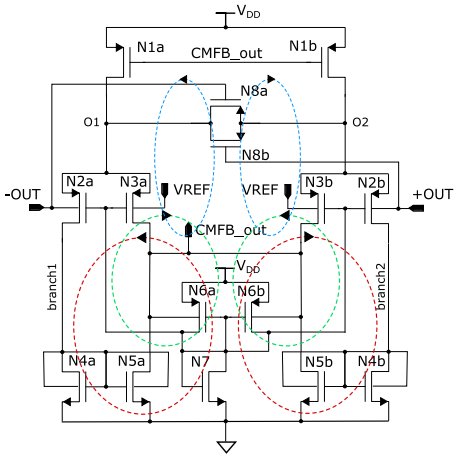


Fig. 4. Schematic diagram of the bulk-driven CMFB: original design (black), modification (red), N–type realization

is jammed between two current sources ($N3x, N5x$) with high output impedance. Therefore, further gain improvement can be expected, which is in the case of inherited low gain bulk–driven configuration, the main effort. Finally, similar to the PDDA, also self-biased approach was employed into CMFB design that slightly contributes to improvement of the entire PSSR of VGA and eliminates the need for a separate bias circuit. Originally, the designed CMFB needs only one bias voltage for transistors $N2x$ and $N3x$ that was in case of self-biased configuration formed by antiseries diode like connected transistors $N6x$ and $N7$. This forms two additional feedback loops, where green dashed and red dashed loop introduces partially positive and negative feedback, respectively. Therefore, design of the CMFB circuit must be carefully taken and rather investigated as part of the whole VGA, especially, if suitable common-mode phase margin requirements have to be fulfilled. To improve its common-mode and differential-mode performance, the topology was extended by a couple of bridged transistors $N8x$ that lightly balance current between two differential pairs.

III. Small-Signal Analysis

A. Small-Signal Analysis of VGA

For better understanding of the gain control technique used in the developed VGA, the low-frequency small-signal analysis has been performed. The small-signal model of the proposed VGA (without CMFB circuit and second stage) is shown in Fig. 5. Since the selected topology is symmetrical, in the small signal analysis, we can consider half-circuit model and evaluate impact of $+V_{in}$ and $-V_{in}$ separately, and then simply sum the contribution of both parts.

The total low-frequency gain of the proposed VGA circuit can be written as

$$A = G_m \cdot R_{out}, \tag{1}$$

where G_m and R_{out} represent the total transconductance and the total output resistance of the proposed VGA, respectively.

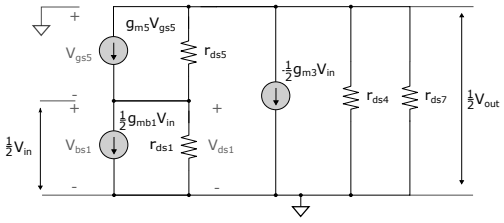


Fig. 5. Small-signal model of proposed VGA

Using the small-signal model (Fig. 5), the total VGA transconductance G_m can be expressed as follows:

$$G_m = g_{mb} \cdot \frac{g_{m5} + g_{ds5}}{g_{ds1} + g_{m5} + g_{ds5}} - g_{mb3} \tag{2}$$

where g_{mb1} and g_{mb3} are bulk transconduntance of transistors M1 and M3, g_{m5} is a transconductance of transistor M5, while g_{ds1} and g_{ds5} are output transconductance of transistors M1 and M5. If we consider that $g_{m3} >> g_{ds5}$, then equation 2 can be rewritten as follows:

$$G_{m1} = g_{mb1} \cdot \frac{g_{m5}}{g_{ds1} + g_{m5}} - g_{mb3} = g_{mb1} \cdot K - g_{mb3} \tag{3}$$

$$K = \frac{g_{m5}}{g_{ds1} + g_{m5}} = \frac{1}{1 + \dfrac{g_{ds1}}{g_{m5}}} \tag{4}$$

One can observe that the total transconductance of VGA (G_m) can be controlled by transconductance g_{m5}, which depends on the control voltage (CTRL). Coefficient K can vary in the range from 0 to 1. If $g_{m5} << g_{ds1}$ the coefficient K form equation 4 is become zero and the total transconductance of VGA will be equal to $-g_{mb3}$. In the opposite case (when $g_{m5} >> g_{ds1}$), coefficient K is equal to 1 and G_{m1} will be equal to $g_{mb1} - g_{mb3}$. This meas that by increasing the transconductance g_{m5}, the total transconductance G_{m1} is decreased.

Besides, the total transconductance G_m also depends on g_{ds1}/g_{m5} ratio. If numerator and denominator of the ratio are divided by I_{ds1}, the following expression can be written:

$$\frac{g_{ds1}}{g_{m5}} = \frac{\dfrac{g_{ds1}}{I_{ds1}}}{\dfrac{g_{m5}}{I_{ds1}}} = \frac{V_{CTRL} - V_{th1}}{2 \cdot (V_{A1} + V_{ds1})}, \tag{5}$$

where V_{CTRL} is the control voltage of VGA, and V_{th1}, V_{A1} and V_{ds1} is the threshold voltage, Early voltage and voltage between source and drain of transistor M1, respectively. From equation 5, it can be observed that g_{ds1}/g_{m5} ratio, which controls the total transconductance G_m, depends on the control voltage of the proposed VGA.

Using the small-signal model depicted in Fig. 5, the output resistance R_{out} of the proposed VGA is expressed as

$$R_{out1} = [(1 + g_{m5} r_{ds1}) r_{ds5} + r_{ds1}] || r_{ds7} || r_{ds4}, \tag{6}$$

where r_{ds1}, r_{ds4}, r_{ds5} and r_{ds7} is the output resistance of transistor M1, M4, M5 and M7, respectively. Since the term $g_{m5}r_{ds1}r_{ds5} > r_{ds1}$, equation 6 can be simplified to

$$R_{out} = \underbrace{g_{m5}r_{ds1}r_{ds5}}_{\text{A}} || \underbrace{r_{ds4}||r_{ds7}}_{\text{B}} \tag{7}$$

In equation 7, terms A and B represent a parallel combination of two resistances. From the VGA design point of view, if term B is maximized, the total resistance R_{out} of the VGA will depend on term A. In such a case, it is possible to vary R_{out} by transconductance g_{m5} that is proportional to the control voltage of the VGA.

B. Small-Signal Analysis of CMFB

To investigate the low frequency common–mode gain of the CMFB circuit (investigation of error amplifier character), the exhausted small–signal model has been derived (Fig. 6), where the individual parameters are also listed.

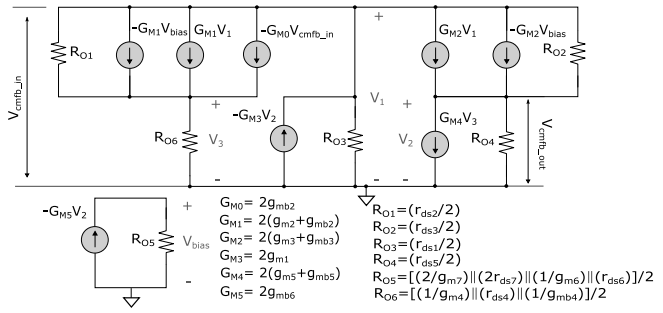


Fig. 6. Low frequency small–signal model of CMFB

Exhausted result of close-loop common-mode gain of CMFB derivation is however, rigorous and too difficult to interpret or extract any handy information because of three existing internal feedback loops, as already discussed. Fortunately, if considering the operational modes of individual transistors (i.e. $N1$ works like linear resistor, $N4, N6, N7$ and $N2, N3, N5$ represent diode-connected transistors and current sources, respectivelly) approximated expression of common-mode gain of CMFB can be derived:

$$Av_{CMFB_{app}} = (\frac{g_{mb2}}{g_{ds3} + g_{ds5}})(\frac{g_{m5} + g_{mb5}}{g_{m4} + g_{mb4}}) \tag{8}$$

Equation 8 is based on the following presumptions:

$$G_{M1} \stackrel{!}{=} G_{M2} \wedge R_{O1} \stackrel{!}{=} R_{O2} \wedge R_{O1} \stackrel{!}{\gg} R_{O6} \wedge G_{M4}R_{O6} \stackrel{!}{=} 1 \tag{9}$$

Thus, equation 8 reveals a few interesting observations: common–mode gain depends only on properties of the differential pairs and current mirrors with no influence of bias circuit and transistors $N1x$. Such a degree of freedom can be useful for finding the solution to comply the sensitive restrictions (Eq.9), where the last one is the most difficult to fulfill. In our case, its value was approximately 1.029 in all corners that leads to inaccuracy of calculation up to cca 2.5 dB. Additionally, one can observe that using gate–driven approach, common–mode gain of the novel self–biased CMFB can be improved further up to $4 - 5times$.

IV. Simulation Results

In this section, simulation results achieved for the designed VGA (including the CMFB and CMFF circuits) using the supply voltage of 0.4 V are presented. The results were obtained from Corner and Monte Carlo (MC) analyses, where the process variation as well as mismatch of devices were taken into account. Since the proposed VGA is suitable for low-frequency applications, there is an assumption that results obtained by post-layout simulation will not differ significantly.

The frequency response of VGA for different values of the control voltage is shown in Fig. 7. It can be observed that for the control voltage of 0.1 V, the VGA gain of about 18 dB and GBW of 1.2 MHz were achieved. These parameters were obtained for the load capacitance of 1 pF. For higher values of the control voltage, the gain decreases down to 6 dB. It is important to note that the bandwidth (BW) does not change with the CTRL voltage, which is one of advantages of the proposed VGA.

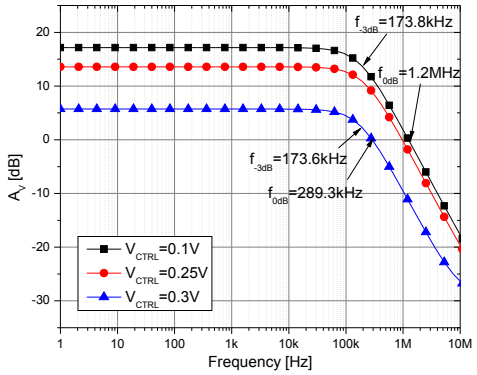


Fig. 7. Frequency response of the VGA

As can be observed from Fig. 8, the VGA gain is varied in the range of the control voltage from 0 V to 0.33 V but it is linear only in the range from 0.3 V to 0.33 V. In the whole range, gain varies from 18 dB to 0 dB. The gain change (for low values of CTRL) caused by process variations is 3.34 dB, which represents a significant result taking into account that the proposed VGA was designed in 130 nm technology. On the other hand, variance of gain for high values of CTRL is substantially higher because the slope of the characteristics is changed by the different process corner. The worst case can be observed in fast-slow (FS) and slow-fast (SF) corner.

Although the supply voltage is very low, gain is relatively stable for low values of CTRL from temperature point of view. For high values of CTRL, deviation of gain caused by temperature increases because the bias voltage of input transistors is temperature dependent. This deviation can be reduce using temperature compensation of the bias voltage.

The important pameter of the PDDA is a CMRR. Fig. 9 shows variations of the CMRR parameter obtained from MC analysis. It can be observed that CMRR varies in the range

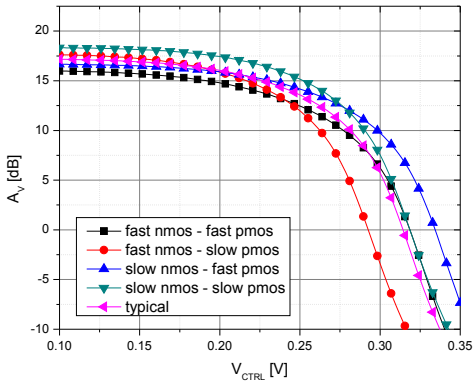


Fig. 8. VGA gain vs the control voltage in all process corners

from -40 dB to -85 dB, while the mean value is about -60 dB. This is very good results generally, which was achieved thank to both CMFF and CMFB circuits being employed.

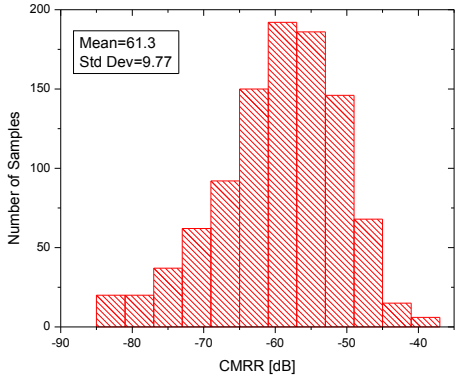


Fig. 9. Variations of CMRR evaluated by MC analysis

Since the self-biased technique was used in the proposed VGA, the PSRR parameter is also improved. Fig. 10 shows the variation of the PSRR parameter, where the process variation and mismatch of all devices were taken into account. One can observe that the PSRR parameter varies in the range from -75 dB to -25 dB, while the mean value of PSRR is about -45 dB.

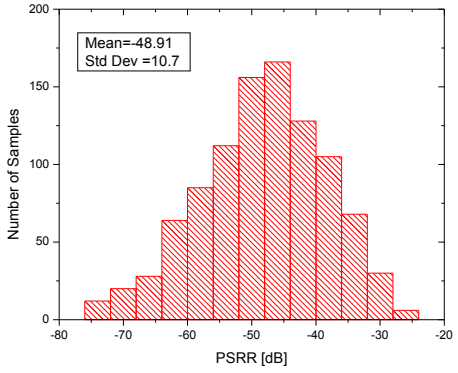


Fig. 10. Variation of PSRR obtained from MC analysis

Since the designed VGA consists a novel bulk-driven CMFB, we individually investigated its performance. Two most important characteristics are depicted in Fig. 11 and Fig. 12. Fig. 11 shows the frequency response across all corners, where the load capacitance of 0.5 pF was considered and the reference voltage $VREF$ was set to $V_{dd}/2$. The common–mode gain between 17.38 dB and 21.29 dB was achieved that is however, further increased by g_{mM7} and g_{mM8} transconductances. The achieved value of common-mode gain is sufficient to hold the common-mode output voltage approximately in the range from 188.5 mV to 199.8 mV in all corners. However, the phase of common–mode gain transfer function (not shown) moves in the range from 0 to 360 deg and therefore, compensation task could be quite challenging. One can also observe that the minimum bandwidth of 23.43 kHz was achieved at slow-slow (SS) corner. This value is sufficient for biomedical and audio applications, where the bandwidth of approximately 1 kHz and 20 kHz is required, respectively.

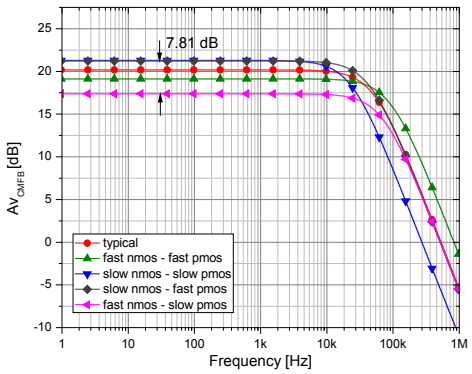


Fig. 11. Frequency response of CMFB

Fig. 12 shows the CMFB output control voltage V_{CMFB_out} versus the VGA output voltage. It is obvious that CMFB is characterized by good performance in terms of the output control voltage range, where CMFB exhibits 0.12 V linear range behavior at 0.4 V supply voltage. We have to also note that FS corner belongs to the worst case corner in terms of common-mode gain and CMFB output voltage operation region.

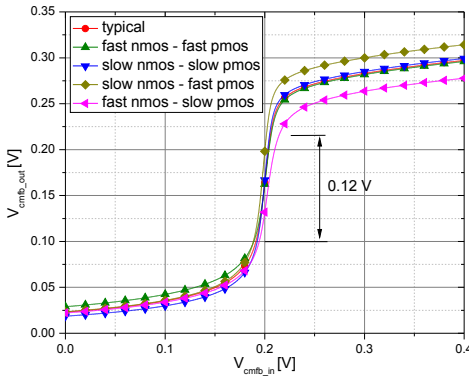


Fig. 12. CMFB output voltage vs. common output voltage of VGA

Main parameters of VGA and CMFB circuits obtained from Corner Analysis are summarized in Fig. 13 and Fig.14, respectively. In the case of the proposed VGA, FF corner represents the worst case. In FF corner, the gain of 15.98 dB and current consumption of 19.57 μA were achieved. A summary of all characteristic parameters i.e. static current consumption, BW, GBW and the maximum ripple of CMFB output voltage $\triangle V_{CMFB_out}$ driven by $\pm$ 200 mV differential output voltage of CMFB, can be found in Fig. 14. Because diode-connected transistors were used instead of current–mode operation transistors, one can observe substantial current fluctuation across individual corners affecting also BW and GBW. The worst case static consumption of approximately 12.5 μA was observed for fast-fast (FF) corner. Since the proposed VGA is based on two input differential pairs and uses the cross-coupled topology, the total current consumption does not depend on the control voltage CTRL, which is considered as advantage.

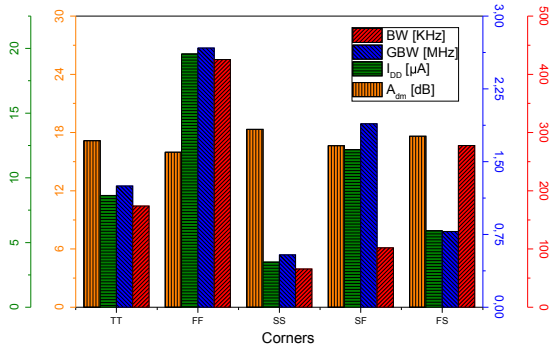


Fig. 13. Summary of the VGA main parameters across all corners.

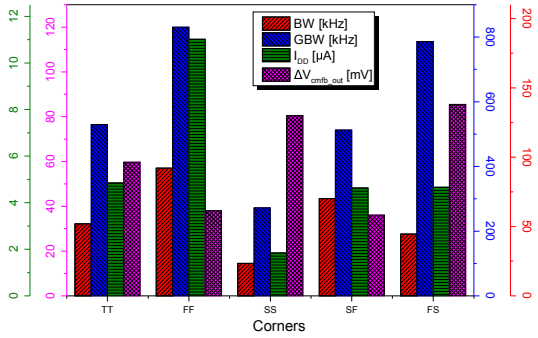


Fig. 14. Summary of the CMFB main parameters across all corners.

V. Conclusion

The novel VGA circuit based on PDDA topology for low-voltage applications, designed in 130 nm CMOS technology, was presented. As demonstrated, the designed VGA represents a building block for low-power and low-voltage applications, where the supply voltage of less than 0.6 V, differential signal processing as well as high dynamic range and low distortion are required. The future work will be led towards design of a logarithmic function generator in order to increase the linear-in-decibel range as well as on further improvement of the main VGA parameters in terms of ExG requirements.

TABLE I
Main parameters of the proposed VGA

Parameter	Condition	Min	Typ	Max
V_{CTRL} [V]	$V_{DD} = 0.4\ V$	0	-	0.33
A_v tun. range [dB]	$0 \div 0.33$	0	-	18
A_v Lin-in-dB [dB]	$0.3 \div 0.33$	0	-	8
GBW [Hz]	$C_L = 1\ pF$	$289k$	-	$1.2M$
BW [Hz]	$C_L = 1\ pF$	$173.6k$	-	$173.8k$
CMRR [dB]	$V_{DD} = 0.4\ V$	-40	-60	-85
PSRR [dB]	$V_{DD} = 0.4\ V$	-25	-45	-75
A_v, cmfb [dB]	$C_L = 0.5\ pF$	17.4	20.1	21.29
GBW, cmfb [kHz]	$C_L = 0.5\ pF$	290	570	820
BW, cmfb [kHz]	$C_L = 0.5\ pF$	20	50	90
Lin. op. range, cmfb [V]	-	0.12	0.13	0.15
Total power [μW]	$V_{DD} = 0.4\ V$	1.39	3.45	7.84

Acknowledgment

This work was supported by the Slovak Republic under grants VEGA 1/0762/16 and VEGA 1/0823/13.

References

[1] H. D. Lee, K. A. Lee, and S. Hong, "A Wideband CMOS Variable Gain Amplifier With an Exponential Gain Control," *Microwave Theory and Techniques, IEEE Transactions on*, vol. 55, no. 6, pp. 1363–1373, June 2007.

[2] P.-C. Huang, L.-Y. Chiou, and C.-K. Wang, "A 3.3-V CMOS wideband exponential control variable-gain-amplifier," in *Circuits and Systems, 1998. ISCAS '98. Proceedings of the 1998 IEEE International Symposium on*, vol. 1, May 1998, pp. 285–288 vol.1.

[3] T. Yamaji, N. Kanou, and T. Itakura, "A temperature-stable CMOS variable-gain amplifier with 80-dB linearly controlled gain range," *Solid-State Circuits, IEEE Journal of*, vol. 37, no. 5, pp. 553–558, May 2002.

[4] A. Suadet and V. Kasemsuwan, "A 1 Volt CMOS Pseudo Differential Amplifier," in *TENCON 2006. 2006 IEEE Region 10 Conference*, Nov 2006, pp. 1–4.

[5] M. Shahabi, R. Jafarnejad, J. Sobhi, and Z. Daei Kouzehkanani, "A novel low power high CMRR pseudo-differential CMOS OTA with common-mode feedforward technique," in *Electrical Engineering (ICEE), 2015 23rd Iranian Conference on*, May 2015, pp. 1290–1295.

[6] F. Khateb and S. Vlassis, "Low-voltage Bulk-driven Rectifier for Biomedical Applications," *Microelectron. J.*, vol. 44, no. 8, pp. 642–648, Aug. 2013.

[7] G. Raikos, S. Vlassis, and C. Psychalinos, "0.5 V bulk-driven analog building blocks," *International Journal of Electronics and Communications*, vol. 66, no. 11, pp. 920 – 927, 2012.

[8] J. Carrillo, G. Torelli, R. Prez-Aloe, and J. Duque-Carrillo, "1-V rail-to-rail CMOS OpAmp with improved bulk-driven input stage," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 3, pp. 508–516, 2007.

[9] J. Carrillo, G. Torelli, M. Dominguez, R. Perez-Aloe, J. Valverde, and J. Duque-Carrillo, "A Family of Low-Voltage Bulk-Driven CMOS Continuous-Time CMFB Circuits," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 57, no. 11, pp. 863–867, Nov 2010.

[10] F. Castano, G. Torelli, R. Perez-Aloe, and J. Carrillo, "Low-voltage rail-to-rail bulk-driven CMFB network with improved gain and bandwidth," in *Electronics, Circuits, and Systems (ICECS), 2010 17th IEEE International Conference on*, Dec 2010, pp. 207–210.

Relaxation Oscillator Calibration Technique with Comparator Delay Regulation

J. Mikulić*, G. Schatzberger* and A. Barić**
* ams AG, Graz, Austria
** University of Zagreb/Faculty of Electrical Engineering and Computing, Zagreb, Croatia
josip.mikulic@ams.com

Abstract – This paper presents an improved technique for the calibration of the relaxation oscillators with respect to the delay of the comparators. The drawbacks of the conventional topology for the relaxation oscillators are analyzed. Based on the analysis, the circuit modification which resolves the effects of the comparator delay in the trimming procedure is proposed. The simulations in ams 0.18µ CMOS technology exhibit more than 5x the improvement in the precision compared to the conventional topology, evaluated in the temperature range from -40 to 125 °C.

I. INTRODUCTION

A stable clock reference is one of the basic building blocks of every digital and mixed-signal circuit. Together with the ever-increasing trend of the semiconductor industry, the clock references tend to be implemented as a full on-chip solution in the systems that require low power consumption and low production cost [1-5]. As a drawback compared to the crystal oscillators, they suffer from the reduced accuracy, most of the time being in the range of several percentage points [1].

Although some techniques have been adapted in order to minimize the influence of the process and the supply voltage variations [1-3], they rely on the stable references over the temperature, which are not always at the disposal for low-cost, full on-chip solutions. On the other hand, the techniques for the process and temperature compensation of the references, such as [6], have a limited success, and are always inferior to the performance of the crystal oscillators. Therefore, it is obvious that the trimming procedure should be considered for the increased accuracy.

In this paper, the digital trimming technique is described, similar to the technique presented in [4]. Combined with the LUT (Look-Up Table) and a temperature sensor, it enables the calibration of the oscillator in the entire temperature range. This principle is illustrated using the conventional relaxation oscillator topology, shown in Fig. 1, which has the advantage of having the calibration possibility with the referent currents, in contrast to ring and harmonic oscillators where the trimming of the resistors and capacitors would be needed. Finally, the modification of the topology which ensures more accurate results of the calibration procedure is proposed, with a negligible increase of the circuit complexity and power consumption.

This paper is organized as follows. In Section II the conventional topology is described and analyzed. The Section III proposes the improvement of the design, while in Section IV the improved topology is verified with the simulations. The final conclusions are given in Section V.

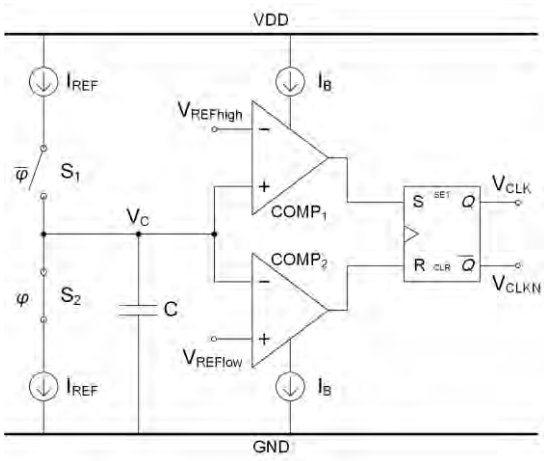


Figure 1. The conventional relaxation oscillator topology.

II. CONVENTIONAL OSCILLATOR ARCHITECTURE

A. Timing Analysis

The conventional topology of the relaxation oscillator is shown in Fig. 1. As seen in Fig. 2, the capacitor C charges and discharges with the current I_{REF}. The switches S_1 and S_2 alternate the charging and discharging process every half-cycle. The two comparators, biased with the current I_B, activate the set and reset signals of the SR flip-flop at the moments when the capacitor voltage V_C gets higher and lower than the referent voltages $V_{REFhigh}$ and V_{REFlow}, respectively. From Fig. 2 it can be seen that the duration of one period is determined by the slew rate of the capacitor voltage ($SR_{VC} = C/I_{REF}$), the difference of the

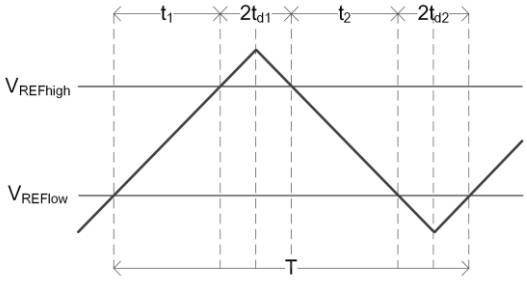


Figure 2. One period of the capacitor voltage V_C.

referent voltages ($\Delta V_{REF} = V_{REFhigh} - V_{REFlow}$), and the time needed for the comparators to activate the set and reset signals (t_{d1}, t_{d2}).

If we neglect the switching delays of the flip-flop and the switches, and assume that $t_{d1} = t_{d2} = t_d$, the expression for the period is then calculated as follows:

$$T = 2\left(\frac{C\left(V_{REFhigh} - V_{REFlow}\right)}{I_{REF}} + t_{d1} + t_{d2}\right) = \frac{2C\Delta V_{REF}}{I_{REF}} + 4t_d \quad (1)$$

The comparator delay t_d observed in (1) represents a serious problem in low-power and high-precision designs [1]. First of all, it takes a significant portion of the total period duration, unless an excessive amount of power is consumed. Furthermore, it is never precisely known, being influenced by parasitics, as well as the process, temperature, supply and referent voltage changes. It also has a negative influence on the calibration procedure, explained in the following section.

B. Calibration Method

From (1) we can observe that the oscillation period can be modified with the current I_{REF}. A straightforward way to realize this is a current DAC (Digital to Analog Converter) controlled by the temperature sensor. The block scheme of such a system is shown in Fig. 3. As a result, the current I_{REF} is defined as

$$I_{REF} = B \cdot I_{REF0}\,, \quad (2)$$

where I_{REF0} is the value of the unaltered referent current, while B is the referent current multiplication factor. Equation (1) then becomes:

$$T = \frac{2C\Delta V_{REF}}{B \cdot I_{REF0}} + 4t_d\,. \quad (3)$$

From (3) we can derive the expression for the factor B needed to trim the oscillator to the desired period (T_{des}) at a certain temperature:

$$B = \frac{T_0 - 4t_d}{T_{des} - 4t_d}\,, \quad (4)$$

where T_0 represents the measured period with $B = 1$.

From (4) we can observe that for the precise calibration, the parameter t_d has to be known upfront. As this is never the case, the method for bypassing this effect is developed in the following section.

III. Modified Oscillator Architecture

A. Comparator Delay Analysis

At the beginning, the analysis of the comparator delay is conducted. For this purpose, two complementary symmetrical OTA (Operational Transconductance Amplifier) topologies shown in Fig. 4 are employed. Using the symmetrical OTA topology has the advantage of reduced systematic offset and better transfer characteristic in contrast to the basic, asymmetrical OTA topology. The complementary topologies are suitable because of the different voltage levels at the inputs of each comparator.

In Fig. 5 the time diagram of the low-to-high transition of the clock voltage V_{CLK} is shown. Currents and voltages correspond to the OTA in Fig. 4(a), i.e. to $COMP_1$ in Fig. 1. For $t < 0$, the voltage at the inverting input of the comparator ($V_{REFhigh}$) is larger than the voltage at the non-inverting input (V_C). From this it follows that the current I_{D1} is larger than I_{D2}, the transistor M_{10} is in the linear re-

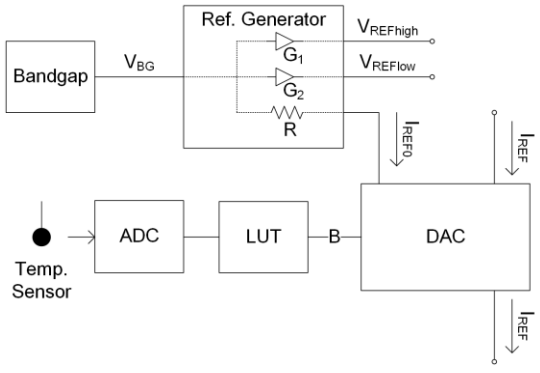


Figure 3. The referent current and voltage generator, together with the calibration circuitry.

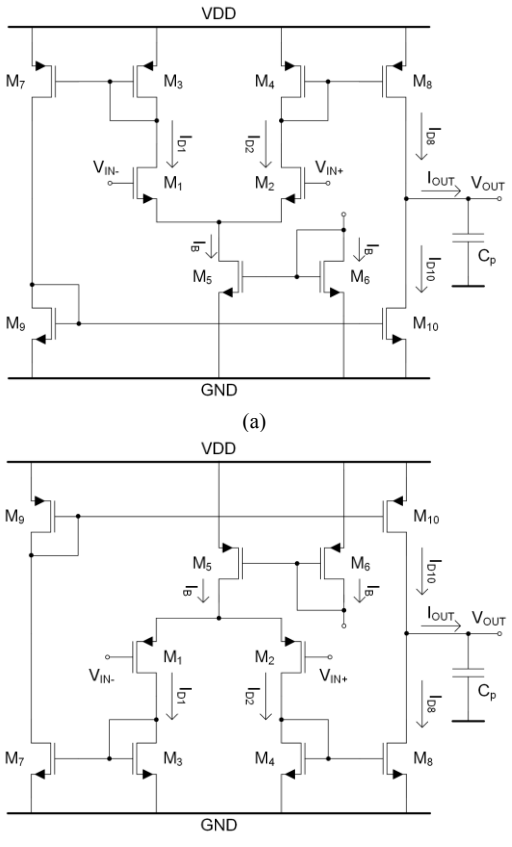


Figure 4. The symmetrycal OTA (a) with nMOS input pair ($COMP_1$) (b) with pMOS input pair ($COMP_2$) .

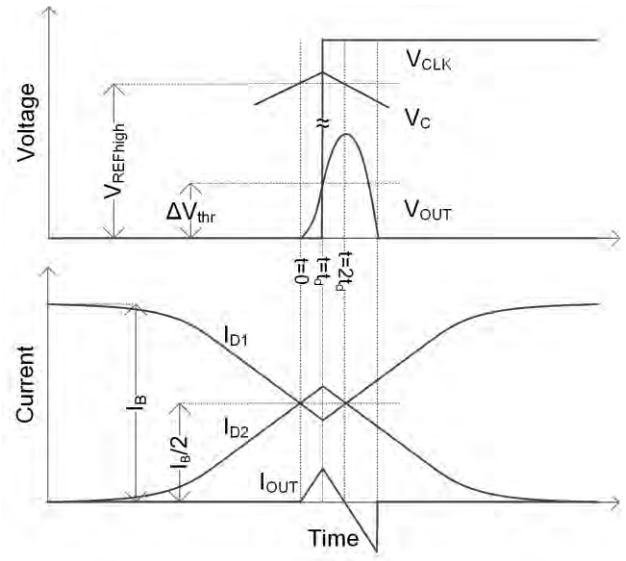


Figure 5. The time diagram of the comparator signals during the low-to-high transition of the clock voltage V_{CLK}.

gion, while the transistor M_8 is in the saturation region. As a consequence, the voltage V_{OUT} is at the ground level. At the moment $t = 0$, the voltages at the inputs are equal, as well as the currents I_{D1} and I_{D2}. At this moment, the transistor M_{10} starts to enter the saturation region, while the output current I_{OUT}, which charges the parasitic capacitance C_p at the output, becomes equal to the difference of the currents I_{D8} and I_{D10}. The voltage V_{OUT} then rises and reaches the ΔV_{thr} level of the flip-flop at $t = t_d$, which then changes the output state of the flip-flop from low to high. At this moment, the states of the switches S_1 and S_2 also reverse, and the voltage V_C now starts to decrease with the identical slew rate. As a result, the waveforms of the currents I_{D1} and I_{D2} are mirrored around $t = t_d$. They become equal again at $t = 2t_d$, after which they discharge the voltage V_{OUT} to the ground level.

In order to describe the behavior quantitatively, first we use the equation for the strong inversion of the MOS transistor, with neglected channel length modulation effect:

$$I_D \approx K_T (V_{GS} - V_T)^2 , \tag{5}$$

where I_D is the drain current, K_T is the technology and design constant, V_{GS} is the gate-source voltage, and V_T is the threshold voltage. Using (5), we can derive the following equation for the differential currents of the transistors M_1 and M_2:

$$\frac{I_{D1,2}}{I_B} \approx \frac{1}{2}\left(1 \mp \Delta V \sqrt{\frac{K_T}{I_B}} \sqrt{2 - \frac{K_T \Delta V^2}{I_B}}\right), \tag{6}$$

where

$$\Delta V = V_{IN+} - V_{IN-} = V_C - V_{REFhigh} . \tag{7}$$

Furthermore, we can safely neglect the higher order effects in (6) related to ΔV^2, as the voltage ΔV in this application will be low. As a result, the expression in (6) is reduced to

$$\frac{I_{D1,2}}{I_B} \approx \frac{1}{2}\left(1 \mp \Delta V \sqrt{\frac{2K_T}{I_B}}\right). \tag{8}$$

For $t > 0$, the current I_{OUT}, which charges and discharges parasitic capacitances at the output node of the comparator, can be written as

$$I_{OUT} = I_{D8} - I_{D10} \approx I_{D2} - I_{D1} . \tag{9}$$

Equation (9) combined with (8) leads to

$$I_{OUT} \approx \Delta V \sqrt{2K_T I_B} . \tag{10}$$

For $0 < t < t_d$, the differential voltage ΔV is equal to

$$\Delta V(t) = \frac{I_{REF}}{C} \cdot t . \tag{11}$$

Combining (10) with (11), we obtain the current I_{OUT} as a function of time for $0 < t < t_d$:

$$I_{OUT}(t) \approx \frac{I_{REF}}{C} \cdot \sqrt{2K_T I_B} \cdot t . \tag{12}$$

The delay t_d of the comparator is experienced as the time needed for the output voltage to charge up to the threshold voltage ΔV_{thr} of the SR flip-flop (approximately half of the supply voltage). As the following is valid:

$$\Delta V_{thr} = \frac{1}{C_p} \int_0^{t_d} I_{OUT}(t) dt , \tag{13}$$

the expression (13) combined with (12) then gives

$$\Delta V_{thr} \approx \frac{1}{C_p} \int_0^{t_d} \frac{I_{REF}}{C} \cdot \sqrt{2K_T I_B} \cdot t dt . \tag{14}$$

After the integration, we obtain the following expression for the delay t_d of the comparator

$$t_d = \sqrt{\frac{2C\Delta V_{thr} C_p}{I_{REF}\sqrt{2K_T I_B}}} . \tag{15}$$

A similar analysis can be performed for the high-to-low transition. Although in this case the parameters C_p, ΔV_{thr} and K_T slightly differ, this difference can be neglected for the purpose of this simplified analysis.

B. Calibration Method with Modified Architecture

After combining (1) and (15), the expression for the duration of the clock period becomes

$$T = \frac{2C\Delta V_{REF}}{B \cdot I_{REF0}} + 4\sqrt{\frac{2C\Delta V_{thr} C_p}{B \cdot I_{REF0}\sqrt{2K_T I_B}}}\,. \qquad (16)$$

If we set the bias current I_B in the following relation with I_{REF0}

$$I_B = \kappa \cdot B^2 I_{REF0}\,, \qquad (17)$$

where κ is an arbitrary constant, the expression for the clock period then becomes equal to

$$T = \frac{1}{B} \cdot \left(\frac{2C\Delta V_{REF}}{I_{REF0}} + \frac{4\sqrt{C\Delta V_{thr} C_p \sqrt{\frac{2I_{REF0}}{\kappa \cdot K_T}}}}{I_{REF0}} \right). \qquad (18)$$

Each parameter in (18), excluding the factor B, is a constant determined by the design and process at a certain temperature. If we compare (18) with (3), we can see that the delay of the comparators can be modified by varying the factor B in the same way as the rest of the expression. As a result, the expression (4) reduces to

$$B = \frac{T_0}{T_{des}}\,, \qquad (19)$$

from which it follows that an improved precision will be achievable, since the only parameter required for the calibration is the measured period T_0.

C. Square Function Realization

The improved calibration method, according to (17), requires the square relationship between the bias current I_B and the parameter B. Exponential behavior of the MOS transistors in weak inversion will be exploited in order to realize this relationship. The schematic of the proposed circuit is shown in Fig. 6.

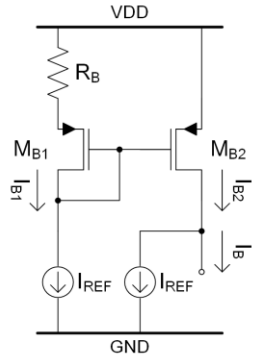


Figure 6. The square function generator.

First, we start from the current equation valid for MOS transistors in weak inversion. The currents I_{B1} and I_{B2} are equal to

$$I_{B1,2} = I_0 \exp\left(\frac{-V_{GS1,2}}{nkT/q}\right), \qquad (20)$$

where I_0 is the constant determined by the design and process, n is the emission coefficient, k is the Boltzmann constant, T is the temperature in Kelvin and q is the charge of an electron. From (20), we can calculate the relationship of the two currents:

$$I_{B2} = I_{B1} \exp\left(\frac{I_{B1}R_B}{nkT/q}\right) \approx I_{B1}\left(1 + \frac{I_{B1}R_B}{nkT/q}\right). \qquad (21)$$

for $I_{B1}R_B/(nkT/q) < 1$. If we consider that

$$I_{B1} = I_{REF} = B \cdot I_{REF0}\,, \qquad (22)$$

together with

$$I_B = I_{B2} - I_{REF}\,, \qquad (23)$$

the expression for the bias current I_B turns into

$$I_B \approx B^2 I_{REF0}^2 \cdot \frac{R_B}{nkT/q} = \kappa \cdot B^2 I_{REF0}\,, \qquad (24)$$

where

$$\kappa = \frac{I_{REF0}R_B}{nkT/q}\,. \qquad (25)$$

Although in theory κ close to or larger than one would induce the higher order effects in the approximation made in (21), in practice it will push the transistor M_{B2} towards the strong inversion region, compensating for the effects and making the approximation valid again. Also note that the circuit in Fig. 6 can be scaled down to reduce the power consumption.

IV. Simulations

To confirm the observations made in the previous sections, the simulations in ams 0.18µ CMOS technology are performed. The basic topology from Fig. 1 is combined with the reference generators and the calibration circuitry shown in Fig. 3. The improved topology also utilizes the bias current generator from Fig. 6.

For the simulation purposes, the supply voltage V_{DD} is set to 1.8 V, the capacitor C to around 1 pF, the current I_{REF0} and I_B to 1.2 µA, while the referent voltages $V_{REFhigh}$ and V_{REFlow} are equal to 1.2 V and 0.6 V, respectively. The factor κ is approximately set to 1. Considering everything

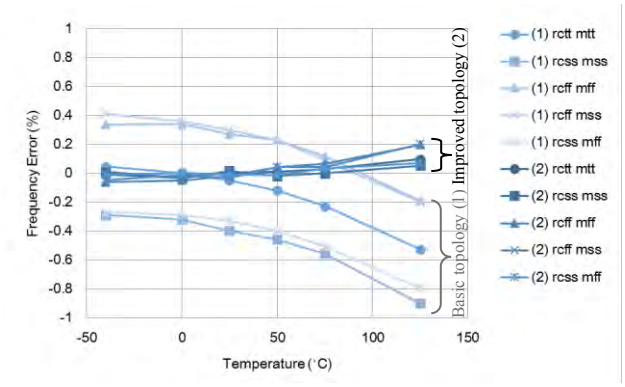


Figure 7. Absolute frequency error after trimming vs. the temperature, plotted for 5 process corners of the basic (1) and the improved (2) topology.

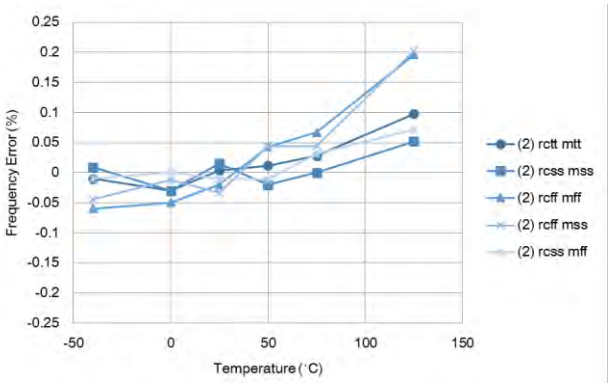


Figure 8. Absolute frequency error after trimming vs. the temperature, plotted for 5 process corners of the improved topology.

mentioned before, it is clear from (1) that the resulting frequency will be somewhere around 1 MHz, not counting the effects of the comparator delay.

Shown in Fig. 7, the simulation results of the calibration procedure demonstrate the improvement. The results are plotted over the temperature range from -40 to 125 °C, with 5 different process corners considered for both topologies. The basic topology exhibits the absolute frequency error from -0.9% to 0.4% in the complete temperature range. The error falls down to a range from -0.05% to 0.2% with the improved topology. That corresponds to the error reduction of more than five times. Moreover, the improved topology exhibits almost ideal behavior in the temperature range from -40 to 75 °C, with the absolute frequency error around ±0.05%. The frequency error for the improved topology is shown separately in Fig. 8.

V. CONCLUSION

A new topology for the improved calibration technique of the relaxation oscillators is proposed. Based on the analytical calculations, the method for neutralizing the comparator delay is developed. The proposed technique is verified with the simulations in ams 0.18µ CMOS process. The results reveal the possibility to achieve a low-power, full on-chip solution, having the accuracy comparable to the accuracy of crystal oscillators.

REFERENCES

[1] Y. Tokunaga, S. Sakiyama, A. Matsumoto, S. Dosho, “An on-chip CMOS relaxation oscillator with voltage averaging feedback,” J. Solid-State Circuits, vol.45, no. 6, pp. 1150-1158, Jun. 2010.

[2] T. Tokairin, et al., “A 280 nW, 100 kHz, 1-cycle start-up time, on-chip CMOS relaxation oscillator employing a feed-forward period control scheme,” in Proc. Dig. Symp. VLSI Circuits, pp. 16-17, Jun. 2012.

[3] K.-J. Hsiao, "A 32.4 ppm/°C 3.2-1.6V self-chopped relaxation oscillator with adaptive supply generation," in VLSI Circuits Symp. Dig. Tech. Papers, pp. 14-15, Jun. 2012.

[4] A. Vilas Boas, A. Olmos, “A temperature compensated digitally trimmable on-chip IC oscillator with low voltage inhibit capability,” in Proc. IEEE Int. Symp. Circuits and System (ISCAS), vol. 1, pp. 501-504, Sep. 2004.

[5] Y. H. Chiang, S. I. Liu, “A submicrowatt 1.1-MHz CMOS relaxation oscillator with temperature compensation,” IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 60, no. 12, pp. 837-841, Oct. 2013.

[6] K. Sandaresan, P.E. Allen, and F. Ayazi, “Process and temperature compensation in a 7-MHz CMOS clock oscillator,” IEEE J. Solid-State Circuits, vol. 41, no. 2, pp. 433-441, Feb. 2006.

A Bootstrap Circuit for DC–DC Converters with a Wide Input Voltage Range in HV-CMOS

N. Mitrovic, R. Enne and H. Zimmermann
Institute of Electrodynamics, Microwave and Circuit Engineering
Vienna University of Technology
Vienna, Austria
e-mail: natasa.mitrovic@tuwien.ac.at, reinhard.enne@tuwien.ac.at and horst.zimmermann@tuwien.ac.at

Abstract - Bootstrap circuits are essential parts of integrated DC–DC converters with an NMOS transistor as high-side switch, which provide voltage overdrive for the gate drivers and to drive the high-side switch's gate. This paper presents a bootstrap circuit which doesn't need an additional supply voltage for charging the bootstrap capacitor to a desired voltage level, but uses the input voltage of the converter. Other advantages of this circuit are proper operation over a wide range of duty ratios and input voltages (from 7.5 V to 45 V). The circuit is designed in 0.18 µm 50 V high-voltage (HV) CMOS technology as part of a buck DC–DC converter for a high output power. Post layout simulations show a voltage dip lower than 80 mV when the high-side NMOS transistor is switched on and the power loss of the bootstrap circuit equal to 160 mW for nominal conditions (input voltage 36 V, duty ratio 15%, bootstrap voltage 4.8 V, input power of drivers 115 mW). The layout dimensions of the bootstrap circuit are 96 µm × 251 µm.

I. INTRODUCTION

Switch mode step down DC–DC converters are widely used due to their high efficiency even at high input power, compared to low-drop-out (LDO) regulators. The use of an NMOS transistor as high-side (HS) switch is preferable rather than having a PMOS HS switch, since it can achieve low on-resistance with usually three times smaller transistor width compared to a PMOS. This means that considerably less area is needed for large-area high-voltage transistors, and, more important, better efficiency. The main disadvantage is that a bootstrap circuit together with the bootstrap capacitor has to be implemented. The bootstrap circuit has to provide enough energy to the bootstrap capacitor so it can properly supply the HS drivers to switch the HS NMOS correctly.

The conventional bootstrap circuits consist of an external diode connected to an additional voltage supply and a capacitor, which is not area efficient and can be potentially a source of instability. Depending on the technology, a bootstrap circuit together with smaller bootstrap capacitor can be implemented on chip, but additional voltages must be provided [1]-[5].

This paper presents a bootstrap circuit which is supplied from the input voltage of the converter, so the need for additional high voltage sources is eliminated. The voltage on the bootstrap capacitor is not determined by the input voltage, which represents one big advantage, so this topology of the circuit is easily adjustable to fit the requirements of wide range DC–DC converters specifications.

The authors would like to thank the Austrian BMVIT via FFG, the ENIAC Joint Undertaking and AMS AG for financial funding in the project eRAMP.

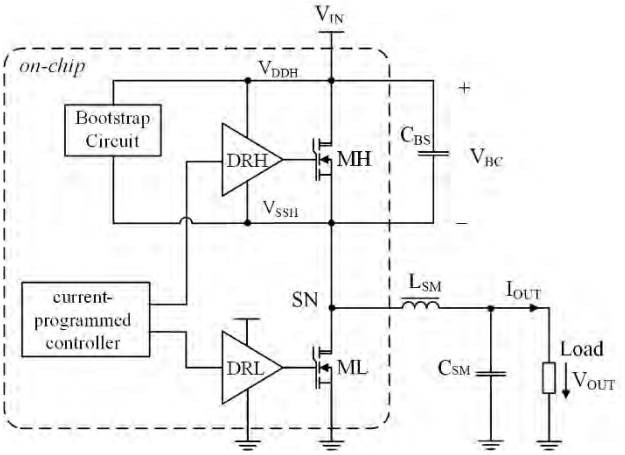


Figure 1. Structure of the buck converter

I. CONVERTER OVERVIEW

Fig. 1 illustrates the simplified structure of the investigated buck converter. The chip design is divided into high-side (HS) switch transistor MH and low-side (LS) switch transistor ML with drivers DRH and DRL, respectively, and regulation loop. The switching node (SN), whose potential is denoted as V_{SSH}, is connected to the off-chip smoothing inductor (L_{SM}) and capacitor (C_{SM}). The drivers for HS and LS switching transistors, are cascade connected tapered inverters. During switching on the power transistors, the drivers have to inject a significant amount of charge into the gates of the power transistors in order to enable their fast turn on and to minimize the switching losses. The regulation loop is based on a current-programmed controller, i.e. peak current-mode control is used, which determines and controls the duty cycle of the converter.

All blocks in the dotted box are realized in an N-well isolated 0.18-µm high-voltage CMOS technology. The circuit is designed for 36 V input voltage and 5.5 V output voltage with 1 A output current. An external inductor and capacitor are used as output filter.

A. Power MOSFETs sizing

Main losses of the converter are usually caused by drivers of the switch transistors and their on-resistance [6]. The power loss in the gate drivers is given by (1), where $e_{G,MH}$, $e_{G,ML}$ are the energy per transistor width needed to switch on the switch transistors, w_{MH}, w_{ML} are the widths of the power MOSFETs and f is the switching frequency. Equation (2) represents the conduction losses of MH and ML, assuming continuous conduction mode (CCM), where $r_{on,MH}$ and $r_{on,ML}$ represent the on-resistances related to the widths, I is the mean output current and δ is the duty ratio.

$$P_{driver} = f \cdot (e_{G,MH} \cdot w_{MH} + e_{G,ML} \cdot w_{ML}) \quad (1)$$

$$P_{res} = I^2 \cdot \left[\frac{r_{on,MH}}{w_{MH}} \cdot \delta + \frac{r_{on,ML}}{w_{ML}}(1-\delta)\right] \quad (2)$$

By minimizing the sum of these two equations, relations for the transistor widths is given in (3) and (4):

$$w_{MH} = I \cdot \sqrt{\frac{\delta \cdot r_{on,MH}}{f \cdot e_{G,MH}}} \quad (3)$$

$$w_{ML} = I \cdot \sqrt{\frac{(1-\delta) \cdot r_{on,ML}}{f \cdot e_{G,ML}}} \quad (4)$$

After replacing the values obtained from simulations and values extracted from the specifications of the converter, the width of the MH transistor is calculated to be w_{MH} = 7 cm, and the width of ML is w_{ML} = 14 cm. Drivers are designed in such a manner that the last inverter in the chain can drive the gate capacitance of the switching transistor, therefore the tapering factor and the number of stages of the inverters are determined relative to the width of the power transistors.

B. Sizing of the bootstrap capacitor

From simulations the electrical charge of Q_G = 5 nC needed for HS MOSFET to switch on is calculated by integrating its gate current during its transition time between off and on state. This charge has to be provided by the bootstrap capacitor, by releasing some charge into the DRH driver which redirects it to the MH gate. The capacitance value of the bootstrap capacitor C_{BS} is determined by setting the voltage drop of V_{BC} being tolerable for this charge loss. In this case, the initial voltage $V_{BC,1}$ is 5 V, the voltage drop was set to be 100 mV, and the capacitance C_{BS} was calculated using the formula for the stored charge of a capacitor, Q = C·V. The initial charge stored in the bootstrap capacitor is $Q_{stored,1} = C_{BS} \cdot V_{BC,1}$ and after discharge we have $Q_{stored,2} = C_{BS} \cdot V_{BC,2}$, $V_{BC,2}$ = 4.9 V, where $Q_G = Q_{stored,1} - Q_{stored,2}$. The bootstrap capacitance is then calculated:

$$C_{BS} = \frac{Q_G}{V_{BC,1} - V_{BC,2}} = 50nF. \quad (5)$$

Since an increase in power consumption of the produced chip compared to the simulated design is expected, the finally chosen bootstrap capacitor is double the calculated value, i.e. C_{BS} = 100 nF.

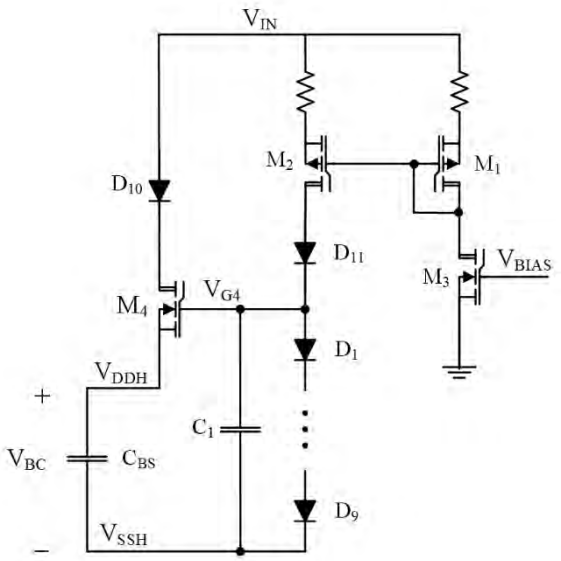


Figure 2. Circuit diagram of proposed bootstrap circuit

II. CIRCUIT IMPLEMENTATION

Fig. 2 shows the bootstrap circuit. The core of this circuit topology consists of the high voltage NMOS transistor M_4, diodes D_1 to D_9 and the current mirror M_1–M_2. The drain current of transistor M_4 directly charges the bootstrap capacitor. The stack of diodes, D_1 to D_9, indirectly defines the maximum achievable bootstrap capacitor voltage, which can be represented as $9 \cdot V_D - V_{th,M4}$ where $V_{th,M4}$ is the threshold voltage of M_4. The high–voltage NMOS transistor M_3 works as a current source, providing the bias current for the current mirror M_1–M_2. The current of the transistor M_3 is determined by the size of the transistor and V_{BIAS} voltage, which can be any low-voltage form DC source that is already used by other regulation circuits or the band gap voltage which is usually used as the reference voltage in the regulation loop of the converter.

When the MH is turned on, switching node SN is on high potential, close to input voltage of converter, V_{IN}. Transistor M_4 is off and charging of the bootstrap capacitor is disabled. Only the transistors M_1 and M_3 conduct very small referent current, which can't be mirrored due to the high potential of node V_{G4}. Capacitor C_1 keeps the gate potential of the M_4 during it's off period, i.e. it holds the V_{GS4} voltage positive and prevents damages to the transistor. Also, diode D_{10} has significant role during this interval, by blocking reverse current of M_4, i.e. preventing M_4 to discharge the bootstrap capacitor. Similar function has the diode D_{11} which prevents C_1 discharging through transistor M_2.

In the opposite interval, when the ML power switch is conducting, charging of the bootstrap capacitor takes place. Transistor M_2 is now conducting, mirroring the current of M_1. One part of the M_2 current serves to charge capacitor C_1. The diodes D_1–D_9 carry current, and they conduct the other part of the current of M_2. The gate-source voltage of M_4 is now positive and above the threshold voltage, so M_4 conducts significant current which charges the bootstrap capacitor.

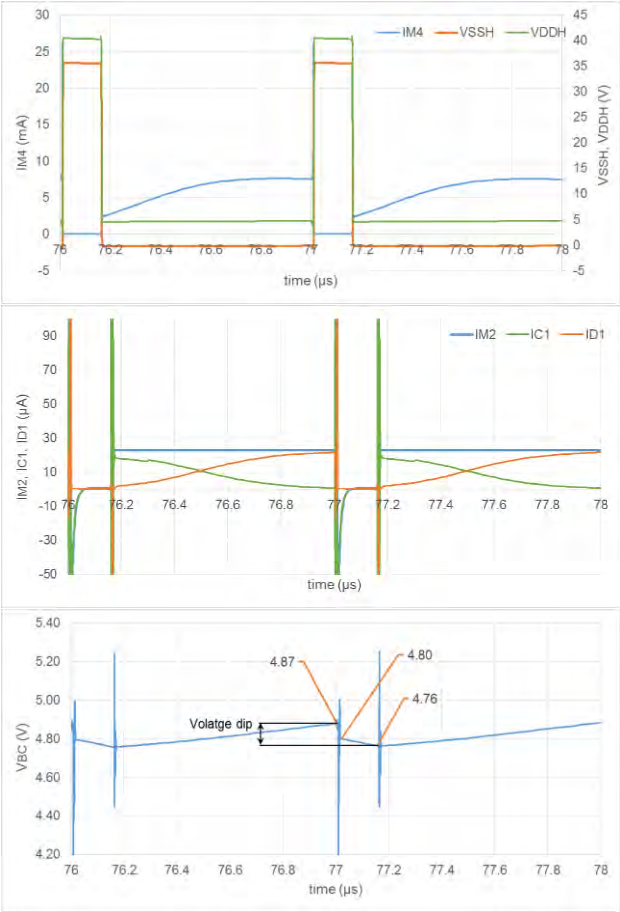


Figure 3. a)-c) Transient waveforms of the voltage and currents of the bootstrap circuit: a) I_{M4} – drain current of the M_4 transistor, V_{SSH} and V_{DDH} – potentials at the bootstrap capacitor terminals; b) I_{M2} – source current of the M_2 transistor, I_{C1} – current of the capacitor C_1, I_{D1} – current of the diode D_1; c) V_{BC} – votage across the bootstrap capacitor.

Transient simulation traces for nominal conditions of the most important signals are shown in Fig. 3. During the time when MH is conducting, voltages V_{SSH} and V_{DDH} are at high potential, so the bootstrap circuit is inactive, M_4 is not conducting, which is clearly visible on the waveform, Fig. 3a. When MH is switched off, the drain current of M_4 is increasing due to raise of its gate voltage, and the bootstrap capacitor is being charged. It also shows how the M_2 current is distributed over the time between capacitor

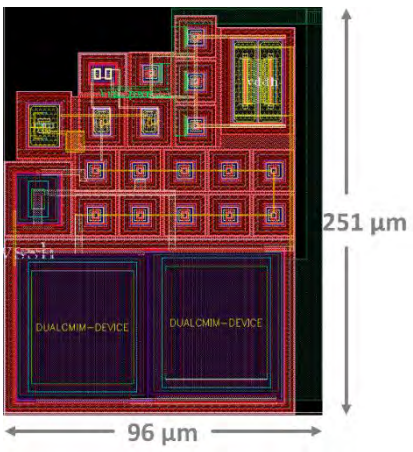


Figure 4. Layout view of the bootstap circuit

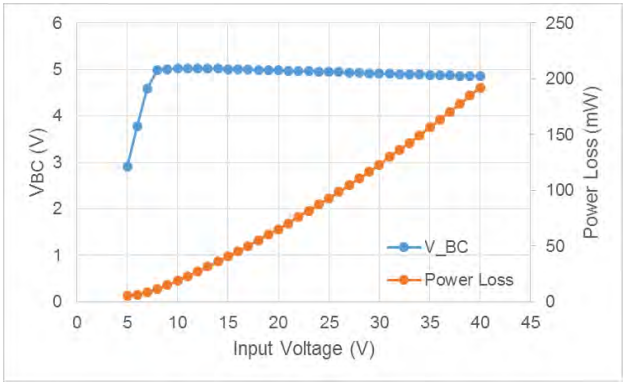


Figure 5. Power loss and bootstrap voltage dependence on the input voltage (duty ratio 0.15)

C_1 and diodes D_1-D_9, Fig 3b. At first, most of the current is flowing into the capacitor C_1, so the voltage V_{G4} is rising, causing then raise in diodes current, which in the second part of the interval becomes dominant. The voltage dip of the voltage V_{BC}, defined like the difference between voltages when MH is off and on, is very low, and it is in the range of 80–100 mV, which can be seen in Fig 3c.

The layout of the bootstrap circuit is given in Fig. 4. The capacitor C_1 which had to be divided into two capacitors connected in series, because of the technology rule about maximum allowed voltage between terminals, occupies most of the area of the circuit.

III. Simulation Results

Post layout transient simulations of the circuit were run with varying two parameters: input voltage of the converter and duty ratio. In both cases two outcomes are interesting to monitor: the final voltage across the bootstrap capacitor, V_{BC}, and the power loss of the bootstrap circuit. All the simulations are performed at switching frequency of 1MHz.

The goal for the final voltage across the bootstrap capacitor is set to be 4.5 V for two reasons. Firstly, this way the proper operation of the drivers is guaranteed, so the reliability that they can always switch is high. The other reason is reduction of the conduction losses of MH transistor, since when it is conducting, its gate-source voltage is equal to the V_{BC}. The goal could have been put even lower, since the drivers can operate properly even with the V_{BC} voltage as low as 2.5 V and the threshold voltage of the MH is 0.7 V, at the cost of efficiency and reliability.

In Fig. 5 it can be observed that power losses are strongly correlated with the input voltage of the converter, with almost linear dependency. The main percentage of the circuit losses is due to the fact that M_4 and D_{10} conduct during the period when the switching node is at low potential, so current of some milliamps at such large voltage difference will make significant loss due to the 36 V input voltage, which is the target input voltage of this converter. Even so, the loss of the circuit which is equal to 160 mW at this operating condition is acceptable in comparison to the 5.5 W converter output power, i.e. the loss of the bootstrap circuit will reduce the efficiency of the whole converter by less than 3%.

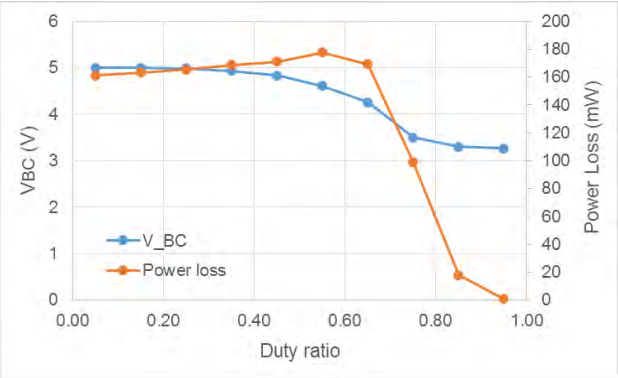


Figure 6. Power loss and bootstrap voltage dependence on the duty ratio (input voltage 36 V)

The maximum achievable bootstrap voltage is mostly limited by the input voltage, since it can be roughly defined as $V_{BC} = V_{in}-V_D-V_{OV}$, where V_D is diode forward voltage and V_{OV} is the overdrive voltage of M_4 transistor. So, as it can be noticed form the graph, there is a minimum input voltage of 7 V required for the bootstrap circuit to operate properly. On the other hand, when the input voltage is bigger than 7 V, the dependence of the bootstrap voltage on the input voltage is negligible, since it is then defined by the forward voltage of the diodes and the gate-source voltage of M_4.

The bootstrap voltage V_{BC} is very dependent on variation of the duty ratio, Fig. 6, since it represents the time interval in which the capacitor is not charged but in which it is being discharged. The power loss of the circuit is not so highly related, but it shows a slight increment with the increase of the duty ratio, which is a consequence of a lower value of V_{BC} and then voltage V_{DDH}, so the voltage drop across M_4 and D_{10} is bigger.

IV. Conclusion

This paper presents a bootstrap circuit for DC–DC converters which is supplied from the input voltage of the converter. The voltage value across the capacitor is not directly related to the input voltage, so the circuit can be used for a wide range of converters. Also, the calculation of the widths of the HS and LS power switch transistors is shown, based on which then the value of the bootstrap capacitor capacitance is calculated. The voltage across the bootstrap capacitance being larger than 4.5 V in the realized circuit is obtained for input voltages already above 7V and duty ratios smaller than 55%. Simulations show also good stability of the bootstrap voltage when the input voltage and the duty ratio are varying.

Acknowledgment

The authors would like to thank A. Steinmair and F. Schrank from AMS AG in Unterpremstätten, Austria, for technical support. The work has been performed in the project eRamp (Grant Agreement N°621270), co-funded by grants from Austria and the ENIAC Joint Undertaking.

References

[1] Seidel, A.; Costa, M.; Joos, J.; Wicht, B., "Bootstrap circuit with high-voltage charge storing for area efficient gate drivers in power management systems, " in *Proc. 40th European Solid State Circuits Conf., ESSCIRC 2014*, Sep. 2014., pp.159-162.

[2] Seidel, A.; Costa, M.S.; Joos, J.; Wicht, B., "Area Efficient Integrated Gate Drivers Based on High-Voltage Charge Storing, " in *Solid-State Circuits, IEEE Journal of*, vol.50, no.7, pp.1550-1559, July 2015.

[3] Xu, J.; Lin Sheng; Xianhui Dong, "A novel high speed and high current FET driver with floating ground and integrated charge pump, " in *Proc. 2012 IEEE Energy Conversion Congress and Exposition (ECCE),* Sep. 2012., pp.2604-2609.

[4] K. Abe, K. Nishijima, K. Harada, T. Nakano, T. Nabeshima, and T. Sato, "A novel three-phase buck converter with bootstrap driver circuit," in *Proc. Power Electronics Specialists Conf.*, Jun. 2007, pp. 1864–1871.

[5] M. Huque, R. Vijayaraghavan, M. Zhang, B. Blalock, L. Tolbert, and S. Islam, "An SOI-based high-voltage, high-temperature gate-driver for SiC FET," in *Proc. IEEE Power Electronics Specialists Conf.*, PESC 2007, Jun. 2007, pp. 1491–1495.

[6] R. Enne and H. Zimmermann, "An integrated low-power buck converter with a comparator controlled low-side switch," in *Proc. IEEE 13th Int. Symp. Design Diagnostics Electron. Circuits Syst.*, 2010, pp. 84–87.

A Fractional-N Subsampling PLL based on a Digital-to-Time Converter

N. Markulic[1, 2], K. Raczkowski[1], P. Wambacq[1, 2] and J. Craninckx[1]
[1]imec, Leuven, Belgium,
[2]Vrije Universiteit Brussel, Brussels, Belgium

Abstract - The paper presents a subsampling PLL which uses a 10-bit, 0.5 ps unit step Digital-to-Time Converter (DTC) in the phase-error comparison path for the fractional-N lock. The gain and nonlinearity of the DTC can be digitally calibrated in the background while the PLL operates normally. During fractional multiplication of a 40 MHz reference to frequencies around 10 GHz, the measured jitter is in the range from 176 to 198 fs. The worst measured fractional spur is -57 dBc and the in-band phase noise performance of the PLL is −108 dBc/Hz. The presented analog PLL in advanced 28 nm CMOS achieves a figure-of-merit (FOM) of -246.6 dB that compares well to the recent state-of-the-art.

I. INTRODUCTION

Frequency synthesizers, typically implemented as phase-locked-loops (PLLs) are omnipresent building blocks used for local oscillator (LO) generation in radio frequency (RF) communication, accurate clock generation in digital circuits, clock recovery, etc. In wireless transceivers, they serve for up/down conversion of the baseband data. A very peculiar aspect of a wireless LO synthesizer is its phase noise and spurious performance. Namely, system level performance in, both, the receive (RX) and transmit (TX) chain are fundamentally limited by the LO phase noise. For example, the limit in error vector magnitude (EVM) in an RX and a TX for high-order modulation schemes is limited by the LO in-band phase noise. In an RX, adjacent channel interferers are reciprocally mixed down onto the desired signal by the LO phase noise and spurs. Moreover, the TX spectral output mask in the receive band is limited by TX LO far-out phase noise, etc. A considerable amount of energy and chip area is therefore typically spent to guarantee low phase-noise LO operation.

The analog subsampling PLL [1] introduced in 2009, shows, even today, an unparalleled synthesizer efficiency amongst the CMOS frequency synthesis state-of-the-art: the lowest integrated phase-noise (i.e. RMS jitter) vs. power consumption. The extreme phase-error detection gain of this architecture reduces the in-band phase-noise and, thanks to that, allows wide bandwidths for efficient VCO noise filtering. In this way, a subsampling architecture achieves the PLL "utopia" in which the output phase-noise is mainly dominated by the reference noise [1]. At the same time, power consumption is reduced thanks to the divider-less operation. However, the architecture's inherent integer-N operation prevents the adoption of this approach in practical wireless transceivers.

Within the fractional-N PLLs, the all-digital systems (for example [2], [3]) show potential of achieving performance similar to the subsampling PLL, although often at the cost of large power consumption of the time-to-digital converter (TDC). Moreover, to achieve highest performance, these systems depend on increasingly complex and power consuming calibration techniques.

We propose a solution that enables fractional-N operation of a subsampling PLL [4,7]. Instead of measuring the fractional residue phase-error with a TDC, we recognize that this error is known *a priori* and can be compensated for by a Digital-to-Time converter (DTC). In this manner, we are able to achieve fractional-N lock, while retaining the key benefits of subsampling operation. The mixed-signal solution that we propose takes advantage of nanoscale CMOS and is not limited by its analog performance. Phase-error detection is done in essence by a switch and a capacitor that benefit from scaling and the charge pump/transconductor can be very simple since, in a subsampling PLL, their noise and linearity do not impact the overall system performance [5].

II. FRACTIONAL-N OPERATION OF THE SUBSAMPLING PLL ENABLED BY A DTC

The integer-N subsampling PLL [1] operates by sampling the (differential) voltage-controlled oscillator (VCO) sinusoid with a repetition rate set by the reference frequency (Figure 1). In a phase-locked state the sampling happens precisely at the zero crossings of the differential sinewave. Any deviation from the timing of the zero crossing results in a non-zero voltage being sampled, which in turn is converted to a correction current fed to the loop filter. Because the sampling events occur only at the edges of the reference, the VCO zero crossings are aligned to these edges and the VCO produces a frequency which is an exact integer-N multiple of the reference.

The basic subsampling PLL cannot synthesize fractional-N frequencies, because it lacks any phase modulation mechanism in the loop (for example a divider). We therefore add a DTC in the reference path of the PLL as depicted in Figure 1. The DTC is used to control the exact moment of sampling, such that it always falls on the expected zero crossing of the VCO, even for non-integer ratios between the VCO frequency and the reference clock.

A simple example with a fractional-N multiplication that differs from integer-N by 0.25 is shown in Figure 1. In the first cycle the sampling edge appears at the same moment as in the integer-N mode. Then, in the second cycle, the sampling edge is delayed by 0.25*Tvco after the reference edge. In the third cycle, the sampling edge is delayed by 0.5*Tvco, then by 0.75*Tvco. Finally, in the

fifth cycle, the sampling should happen 1*Tvco after the reference edge, however, simply skipping a VCO cycle and sampling at the original reference edge yields to the same effect. Since the required PLL frequency and the reference frequency are known, it is possible to calculate the position of any following zero crossings with absolute precision. The DTC should cover at least 1 VCO period of delay.

The digital computation of the necessary phase adjustment, i.e. of the delay that needs to be inserted in the reference path is depicted in Figure 2. The difference between the multiplication factor N and the integer-quantized value is extracted first. A ΔΣ modulator is used to generate the integer quantization of N so that the signal *Diff* is a zero-mean stream accumulated with the desired "phase wrapping" behavior. The accumulated value *Acc* is then appropriately scaled on the available DTC input quantization range.

III. IMPLEMENTATION CONSIDERATIONS

If a fractional-N subsampling PLL, as described in the previous section, were implemented with an ideal DTC, it would have the same performance as an integer-N subsampling PLL. This lies in stark contrast to the case of a traditional mixed signal ΔΣ PLLs, where there is an unavoidable penalty associated with the modulation through the divider. Any practical implementation of the fractional-N subsampling PLL system will, however, be limited by the limitations of the DTC.

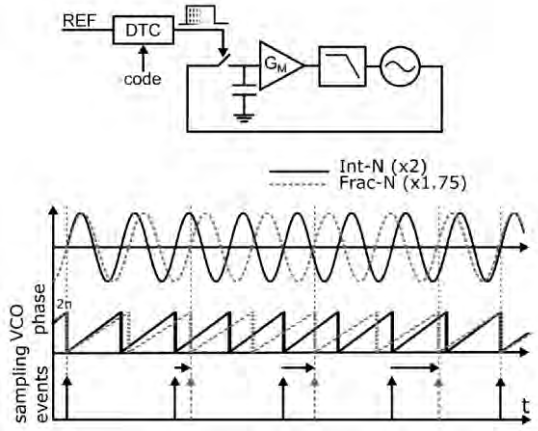


Figure 1: Fractional-N subsampling PLL operation.

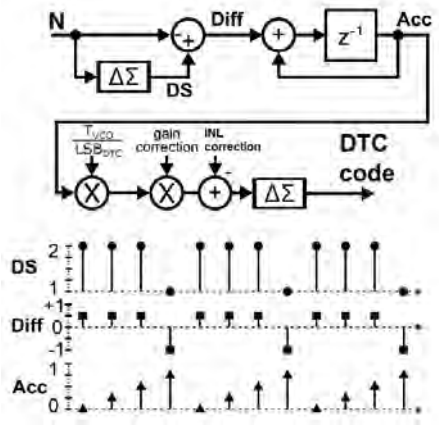


Figure 2: DTC input calculation path.

A. DTC quantization error

A DTC has finite resolution. To scale the output of the accumulated phase error to a digital tuning code, the output of the accumulator *Acc* in the Figure 2 needs to be multiplied by a factor $\frac{T_{REF}}{(LSB_{DTC} \cdot N_{FRAC})}$. The sampling moments hence occurs with accuracy limited by the LSB of the DTC and the resulting error is fed into the low pas filter (LPF), thereby instantaneously modulating the VCO, creating spurs.

System level simulations show that choosing a DTC LSB of 0.5 ps ensures that the quantization noise appears below other loop noise. Moreover, a second ΔΣ modulator (Figure 2) is used in front of the DTC to shape the associated quantization noise beyond the PLL bandwidth. Thanks to the fact that the stream is perfectly accurate on average, the average PLL frequency is also accurate, with no visible modulation.

Another modification to the basic system that helps to mitigate the problem of limited DTC resolution, is to use a MASH modulator in the beginning of the delay computation path (initial ΔΣ modulator in Figure 2). A MASH modulator provides better randomization of the generated code, which helps with reducing spurious content. Compared to a first-order ΔΣ, the generated codes have a larger range, which results in a larger delay range of the DTC. For this reason we choose a 10-bit DTC implementation. By generating delays larger than one VCO period, it is possible to effectively de-color the sampling data. Moreover, randomizing DTC codes provides an effect similar to dynamic element matching. Since e.g. four DTC codes are used in MASH 1-1-1 mode to generate the same effective sampling phase, the apparent DTC nonlinearity is randomized.

B. DTC gain error

DTC gain can be defined as the amount of delay per least-significant bit (LSB) code. The DTC is analog in its nature and susceptible to PVT variations, hence its absolute gain is unknown and varying with time and temperature. Gain error in the delay steps introduces problematic spurs in the spectrum of the PLL. Automatic background calibration which tracks the gain variations becomes therefore absolutely necessary.

An automatic DTC gain calibration can be designed similarly to the popular least-mean-square (LMS) based mechanisms used in digital PLLs [6] (Figure 3). Simply stated, it is possible to extract the sign of the sampled voltage and correlate it with a change in direction of the DTC word. Intuitively, if the modulator "tells" the DTC to sample later, but due to a gain error "early" samples get consecutively detected — it is possible to deduct that the DTC gain is too low. After accumulation, the correction word can be applied as a scaling factor to the computation path of Figure 2. When the correction loop converges, there is no penalty on phase noise. Figure 4 shows a simulation result where a 10% gain error was applied to the DTC. This error introduces a large ripple in the sampled voltage, which in turn results in large spurs at the output of the PLL. After the DTC gain is corrected, the sampled voltage converges back to zero.

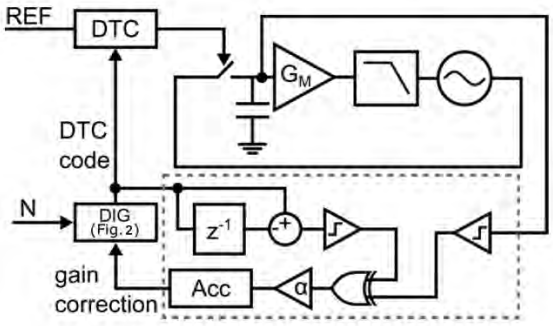


Figure 3: DTC gain digital background calibration.

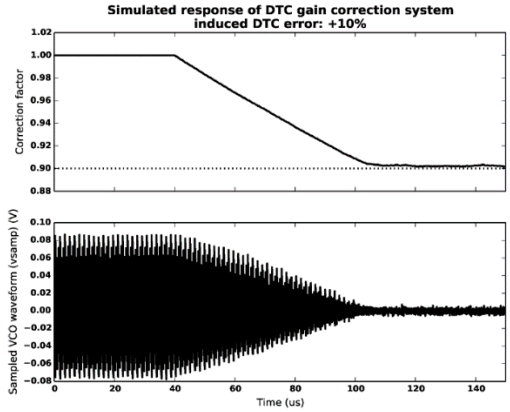


Figure 4: DTC gain background calibration simulation with 10% gain error on the DTC.

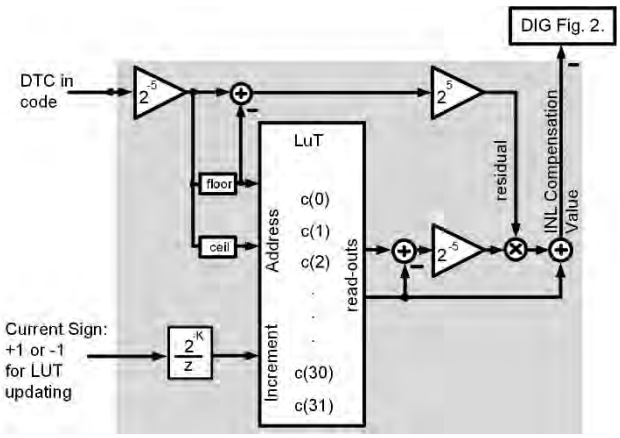


Figure 5: DTC INL calibration loop [7].

C. DTC nonlinearity

The DTC nonlinearity, will naturally increase spurious content at the output of the PLL. Many techniques for improving linearity which are present for digital-to-analog converters (DACs) also apply for the DTC. For example, careful layout of the tuning element is of highest priority. Advanced nanometer scale technologies offer a significant advantage in this regard, thanks to ever-improving lithography resolution.

Nevertheless, to ensure linearity of phase-error detection path, we employ DTC nonlinearity calibration loop. The calibration is based on Error Sign signal observation, which represents the sign of the instantaneous phase error. This 1-bit phase error is random and zero on average in a linear system, but in presence of DTC INL it becomes "colored" by its nonlinearity. Essentially, we exploit the correlation between the Error Sign signal and the particular DTC input code which induced it, to restore the INL curves and pre-compensate for them. A look-up-table (LuT) with a set of coefficients c(0:k-1) is used to approximate the 10-bit DTC INL curve (where k is 32). In every clock cycle, the input code addresses two neighboring LuT coefficients which piece-wise linearly approximate the expected, instantaneous INL error. The INL compensation value is simply subtracted from the original code, which ideally forces the DTC to produce no error (zero mean Error Sign) for the given code. The LuT correction coefficients are updated gradually, by integrating the scaled Error Sign value to the appropriate address (defined with the input code) in every cycle. When the calibration is initialized, the LuT is reset to zero. While the PLL operates, the coefficients c(0:k-1) slowly change towards their optimal calibrated values that cancel the INL error. At this moment, the coefficient updating can be disabled. The algorithm convergence speed is determined by the tap gain G, where a typical value of 2^{-13} results in 10 ms approximate calibration time. The DTC INL calibration loop is enabled after the gain error is corrected. An important detail is that the offset in the extracted error sign needs to be digitally compensated for the loop to converge.

D. DTC phase-noise

In this paper we propose a solution to enhance an integer-N subsampling PLL by placing a phase modulator (DTC) in the path of the reference. Unfortunately, the phase noise contribution of the DTC adds directly to the phase-noise of the reference. Ultimately, the in-band phase noise of the subsampling PLL is limited by the phase noise of both the reference and the DTC, since both pass the system in the same way. Therefore, great care must be taken to minimize the DTC's contribution to phase noise, otherwise the unique phase noise advantages of the subsampling architecture will be lost. Here, scaling of CMOS technology is again on our side, since transistors are getting faster with every node, reducing jitter and phase noise.

IV. CIRCUIT IMPLEMENTATION

The subsampling phase locked loop can only detect phase error, which makes it susceptible to false locking at any N. Therefore, a frequency acquisition loop is required in addition to the subsampling loop [1] (see Figure 6). A conventional PLL easily fulfills this requirement. It can be disabled once frequency has been acquired in order to save power.

Common to both frequency and phase acquisition loops are the low-pass filter (LPF) and the VCO. For the purpose of demonstrating the concept of the fractional-N subsampling PLL we have chosen the simplest LPF design—a passive third-order lead-lag filter. Tunable resistance in the LPF has been implemented to be able to change the bandwidth of the PLL. Such a simple filter can cause increase in reference spurs and is often avoided in classical charge-pump-based PLLs. Spurious content can

increase because the varying level of tuning voltage can introduce mismatches between the currents of the charge pump. In a subsampling PLL, however, any offset in currents of is compensated by a slight modification of the locking point Figure 7. A locked condition always means zero output current of the transconductor(G_M). If changes to the output level cause an input referred offset of the G_M, the PLL will adapt its phase to compensate for this offset.

A. Implementation of the Subsampling Loop

The subsampling loop consists of a VCO buffer, a sampler and an G_M. Additionally, the DTC provides the required phase modulation. Figure 8 shows the circuits along the subsampling path.

A VCO buffer is required in order to reduce the kickback effect from the sampler to the VCO [5] and to interface the signal levels between the blocks. In this test chip, to accommodate for changing phase noise requirements of a software-defined radio, we have implemented a low-noise VCO that can be operated from a variable supply as high as 1.8 V. Therefore, the input buffer needs to convert the level between the high voltage VCO domain (max. 1.8 V) and the core domain (0.9 V). The buffer is implemented with a tunable capacitive attenuator and a source follower pair (Figure 8). The tunable attenuator is built with metal-oxide-metal (MOM) capacitors and provides additional tuning of loop gain. The buffer is also the largest contributor to power consumption in this loop, as it needs to process a GHz-range signal.

The sampler is built around an NMOS switch and a small MOM capacitor. In total, taking into account the input capacitance of the G_M, the sampling capacitance is 20 fF. Thermal kT/C noise can be neglected because it is already suppressed by the large detection gain. The implemented sampler uses an auxiliary sampler operating in inverted phase to the primary sampler in order to reduce load variability of the VCO.

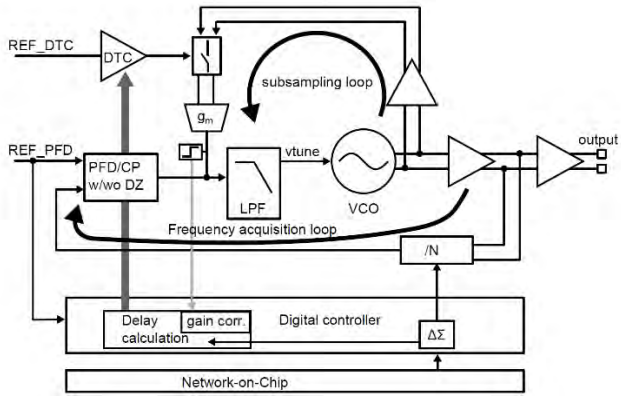


Figure 6: Complete system overview.

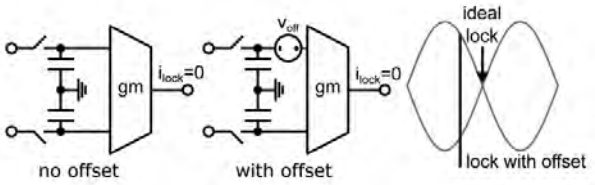


Figure 7: The subsampling PLL always locks into a state that guarantees zero output current, even in presence of offset and mismatch.

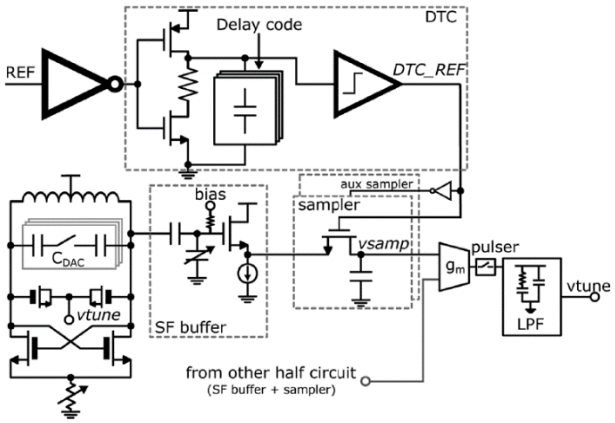


Figure 8: Simplified block diagram of the subsampling loop.

Since the implemented VCO can operate from the IO voltage (1.8V), the tuning voltage also has a range larger than the core voltage. Therefore, the output stage of the G_M needs to provide translation from the low voltage domain of the sampler to the high voltage domain of the LPF and the VCO. Identically to [1], the phase-error detection gain is so large that duty-cycling is required in the output stage of the G_M. Pulsing is done with a simple digital pulse generator that opens the output switches of the G_M.

An important part of the system is the background correction of DTC gain and nonlinearity. As said earlier, the error signal from within the PLL is present in the sign of the sampled voltage. However, this is true only if no mismatches are present in the system. If there are any mismatches in the phase detection circuitry the PLL will adjust the locking phase (and sampled voltage) so that the output current of the G_M is zeroed (Figure 7). Therefore, the gain correction mechanism requires detection of the sign of output current. Using a simple clocked comparator to detect the sign of the swing in relation to *Vtune* voltage is sufficient to obtain information about the sign of the output current.

B. Implementation of the Digital-to-Time Converter [8]

Since the DTC is at the input of the system, its phase noise is multiplied by a square of the PLL multiplication number when transferred to the output (here: 48 dB as N = 250). On top of that, any kind of non-linearity present in the phase error comparison path leads to potential noise folding or spurs [9].

From the PLL system perspective, we target a 10-bit DTC, with a 0.5 ps unit step. This delay range covers multiple VCO periods allowing operation with the third-order MASH 1-1-1 modulator. Because the 0.5 ps step is very small and we know from system simulations that the PLL is sensitive to its disturbance, we can suspect that the DTC needs a good isolation from any noise coming from the supply.

Implementation of the delay generator is shown in Figure 9 [8]. The first inverters in the chain serve as an input buffer towards the delay circuit loaded with a tunable MOM capacitance C_L.

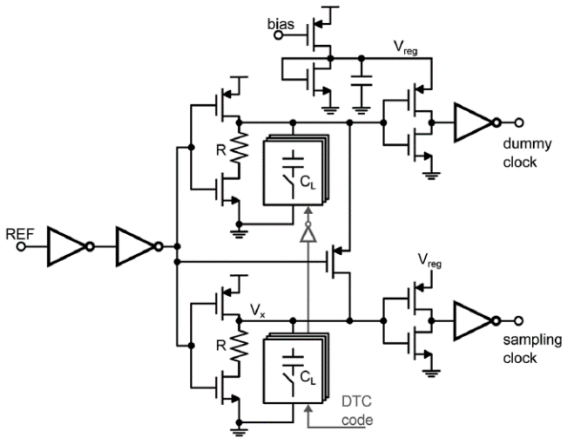


Figure 9: Implementation of the DTC.

To suppress mismatch-based errors for the chosen unit size, the capacitor array employs a 5 bit binary/thermometer segmentation. With the unit capacitor size of 3 fF, this ensures statistical DNL errors below 0.5 LSB. The 10-bit array is placed in a common centroid layout to avoid systematic nonlinearities. Only the high-to-low transition of the V_X voltage is important, because the subsampling loop reacts only on closing of the sampling switch. One could realize discharging of the load capacitance using a simple NMOS transistor, however, this would lead to an excessive $1/f$ noise contribution, which would dominate the PLL phase noise. To reduce this effect we introduce a resistor above the NMOS. The exponential discharging is determined then by the corresponding RC time constant. The delay is, however, a linear function of capacitance. A resistor sets the discharging slope and hence contributes to the output phase noise, however, it generates no $1/f$ noise. Furthermore, any supply ripple coming from the preceding buffer only modulates the NMOS switch resistance which is an order of magnitude smaller than the discharging resistor and does not affect delay. The phase noise level introduced by the delay generator can be derived as

$$\mathcal{L}_{white} \sim 10\log\left(f_{out}\,\frac{kT\,C_{Load}\,R^2}{V_{DD}^2}\right) \sim 10\log\left(f_{out}\,\frac{kT\,\tau_{delay}\,R}{V_{DD}^2}\right), \quad (1)$$

where *R* is the resistor value and *VDD* is the supply voltage of the delay element and *fout* is the output frequency. Based on (1) and the targeted minimal delay step of 0.5 ps we size the *R*=180 Ohm and *C* = 3 fF to lower the noise of this stage to 160 dBc/Hz for maximal delay. The RC delay control block is followed by a CMOS inverter serving as a comparator to restore steep slopes. This circuit toggling moment is unfortunately dependent on the input slope shape which degrades the linearity of a high range DTC. Care must also be given to the fact that regeneration of the RC-delayed slope is most vulnerable to supply modulation. A tunable regulated supply shown in Figure 9 is used to protect the supply of the comparator and the following buffer. The regulated supply consists of a constant current source biasing a diode-connected transistor. A capacitor of 4 pF is used for additional decoupling of the regulated supply node. At the moment of toggling, charge is instantaneously pulled from the capacitor, and not from the VDD. The dip in the regulated supply voltage is suppressed by the gain of the current source before reaching the top supply. The dynamic charge flow is in this way kept within the structure itself.

C. *Implementation of the VCO [10]*

The VCO is a thick-oxide NMOS cross-coupled core with current limiting realized using a tunable resistor. Simulations have shown that this architecture yields favorable 1/f noise compared to the traditional current-source-based architecture. The VCO has been designed to meet the stringent GSM900 specification for out-of-band phase noise. The inductor coil is created using a stack of two top copper metals (each <1 μm thick) and an aluminum redistribution layer. Digital tuning is realized using a bank of NMOS-only switched capacitor cells. The simulated Q of the tank reaches 18. The VCO is designed to operate with a low drop-out linear regulator (not present on chip) with a supply between 0.9V and 1.5 V, depending on the required phase noise performance and available power.

D. *Frequency acquisition loop*

The frequency acquisition loop has been implemented with a chain of divide-by-2/3 circuits, a traditional 3-state phase frequency detector (PFD), enhanced with a large deadzone [1] following and a very simple charge pump. The first stage of the divider is made with current mode logic, since the VCO frequency can reach 12 GHz, but the following stages of the divider are standard CMOS gates.

Once the frequency acquisition is complete, the loop automatically becomes inactive thanks to the increased dead-zone in the PFD and can be completely shut down, saving power. In general, the loop components for both the phase and the frequency acquisition loop can be made very simple and do not require neither good precision, nor good matching, nor low noise.

V. EXPERIMENTAL RESULTS

The prototype IC was fabricated in TSMC 28nm bulk digital CMOS technology, and its size is 0.77 mm2 (excluding IO ring). It operates on 0.9 V and 1.8 V supplies (IO interface and the Gm stage).

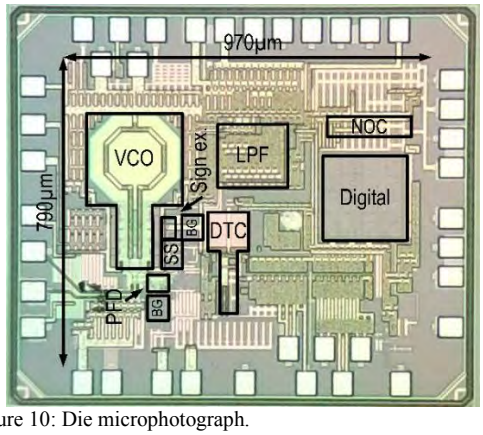


Figure 10: Die microphotograph.

The measured power consumption is 5.6mW in total, of which 1.8mW is for the loop components, 2.7mW for

the VCO, and 1.1mW for digital circuitry that all runs on reference clock of 40 MHz. The VCO tuning range is 10.1 – 12.4 GHz. The measured in-band phase noise around a close-to-integer fractional 11.72 GHz carrier is -107.9 dBc/Hz (Figure 11). The measured RMS jitter is 198 fs after calibration is enabled, with an integration range from 10 Hz to 40 MHz and all spurs (worst fractional) included. The measured spurious performance is shown in Figure 12. The worst fractional spur before calibration appears at -41 dBc but drops with 15.6 dB after calibration to -56.6 dBc. The integer spur is at -69 dBc. The PLL achieves one of the best reported FOMs in the recent state-of-the-art: -246.6 dB.

VI. Conclusion

A subsampling PLL is an architecture that offers extremely low phase noise, however, in its original form it is limited only to integer-N frequency multiplication. We presented a fractional-N subsampling PLL which operates based on a low-noise, low-quantization error DTC. The DTC is enhanced by digital background calibration which suppresses gain and nonlinearity issues. In this way we enhance the original phase noise performance of an integer-N subsampling PLL for fractional-N synthesis, without drawbacks of additional noise folding and spurs.

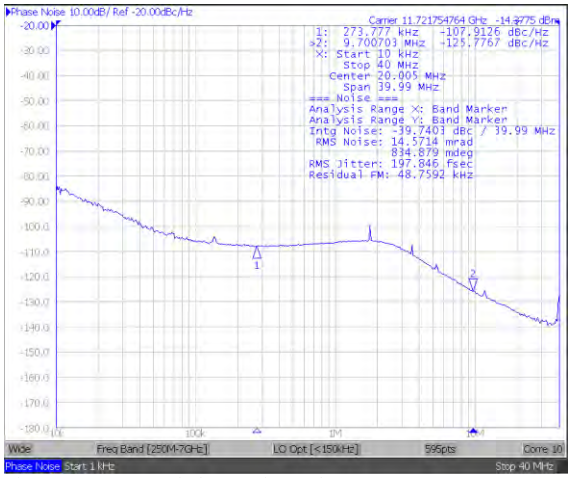


Figure 11: Measured phase noise at the PLL output.

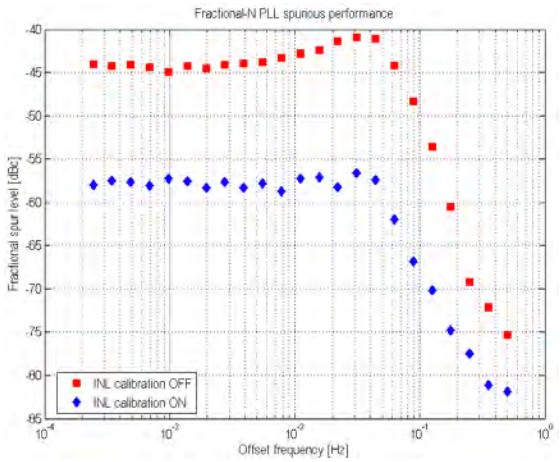


Figure 12: Measured fractional spur level at the PLL output.

Table 1: Performance summary and comparison to the state-of-art.

PLL Comparison	This work	Chen ISSCC'15	Huang ISSCC'14	Gao ISSCC'15
Architecture	Subsamp. Analog	Subsamp. Digital	Subsamp. Analog	TDC-based Digital
Reference (MHz)	40	49.15	48	40
Output (GHz)	10.1-12.4	2.6-3.9	2.2-2.9	5.82 GHz
Tuning range (%)	20.4	40	27.4	N.A.
Bandwidth (kHz)	1800	700	500	500[3]
In-band PN[1] (dBc/Hz)	-107.9	-97.2	-97.9	-105.9
Ref. spur (dBc)	-69	-60	-55	-76
Frac. spur (dBc)	-56.6	-62	-53	N.A.
RMS jitter (fs)	176/198 (10k-40M)	226/240 (1k-100M)	400/266 (10k-30M)	174 (10k-10M)
Power (mW)	5.6	11.5	17.3	9.5
FoM[2] (dB)	-247.6/-246.6	-241.8	-235.6/-239.1	-245.5
Area (mm²)	0.77	0.23	0.75	0.3
Process (nm)	28	65	180	28

[1] Scaled to 11.72 GHz.

[2] $\mathrm{FoM} = 10\log\left(\frac{\mathrm{jitter}^2}{1\mathrm{s}^2}\,\frac{\mathrm{power}}{1\mathrm{mW}}\right)$.

[1] X. Gao *et.al.*, "A Low Noise Sub-Sampling PLL in Which Divider Noise is Eliminated and PD/CP Noise is Not Multiplied by N^2," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 12, pp. 3253–3263, 2009.

[2] C.-W. Yao *et.al.*, "A low spur fractional-N digital PLL for 802.11 a/b/g/n/ac with 0.19 ps RMS jitter," in *VLSI Circuits (VLSIC), 2011 Symposium on*, 2011, pp. 110–111.

[3] C.-W. Yao *et.al.*, "A 2.8-3.2-GHz Fractional-Digital PLL With ADC-Assisted TDC and Inductively Coupled Fine-Tuning DCO," 2013.

[4] K. Raczkowski, *et.al.*, "A 9.2-12.7 GHz wideband fractional-N subsampling PLL in 28 nm CMOS with 280 fs RMS jitter," *Solid-State Circuits, IEEE Journal of*, vol. 50, no. 5, pp. 1203–1213, 2015.

[5] X. Gao *et.al.*, "Spur reduction techniques for phase-locked loops exploiting a sub-sampling phase detector," *Solid-State Circuits, IEEE Journal of*, vol. 45, no. 9, pp. 1809–1821, 2010.

[6] D. Tasca *et.al.*, "A 2.9-4.0-GHz Fractional-N Digital PLL With Bang-Bang Phase Detector and 560," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 12, pp. 2745–2758, 2011.

[7] N. Markulic *et.al.*, "9.7 A self-calibrated 10Mb/s phase modulator with -37.4dB EVM based on a 10.1-to-12.4GHz, -246.6dB-FOM, fractional-N subsampling PLL," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, 2016, pp. 176–177.

[8] N. Markulic *et.al.*, "A 10-bit, 550-fs step Digital-to-Time Converter in 28nm CMOS," in *European Solid State Circuits Conference (ESSCIRC), ESSCIRC 2014-40th*, 2014, pp. 79–82.

[9] S. Levantino, *et.al.*, "An adaptive pre-distortion technique to mitigate the DTC nonlinearity in digital PLLs," *Solid-State Circuits, IEEE Journal of*, vol. 49, no. 8, pp. 1762–1772, 2014.

[10] B. Hershberg *et.al.*, "A 9.1-12.7 GHz VCO in 28nm CMOS with a bottom-pinning bias technique for digital varactor stress reduction," in *European Solid State Circuits Conference (ESSCIRC), ESSCIRC 2014-40th*, 2014, pp. 83–86.

Infrared Protection System for High-Voltage Testing of SiC and GaN FETs used in DC-DC Converters

Filip Hormot, Josip Bačmaga, Adrijan Barić
University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
Tel: +385 (0)1 6129547, Fax: +385 (0)1 6129653, e-mail: filip.hormot@fer.hr

***Abstract*—This technical paper presents the design and testing of the protection system for evaluation of high-voltage devices and circuits, e.g. SiC and GaN devices or high-voltage switching DC-DC converters. The high-voltage testing area is protected from invading by an array of infrared (IR) emitters and detectors. Whenever an object passes through the IR protected space, a high-voltage DC source is disconnected from its power supply and harmful effects of high DC voltage are avoided. The functionality of the developed system is tested and its characteristic response timings are measured.**

***Index Terms*—hazardous DC voltage, high-voltage switch-mode power converter**

I. Introduction

The silicon FETs are being used for many years as power devices in switching DC-DC converters. However, power converters operating at high power levels and voltages of several hundreds of volts are not achievable using the existing silicon power devices [1]. Instead, the silicon carbide (SiC) power devices are being used in high-voltage power converters due to their high voltage breakdown and thermal conductivity [2]. Furthermore, gallium-nitride (GaN) FETs due to their extremely low gate charge and output capacitance can operate at higher switching frequencies [3] compared to the silicon FETs. Increase of a switching frequency reduces the size of external passive components which results in a higher power density and efficiency of the DC-DC converter.

High-voltage SiC and GaN FETs used in DC-DC converters are subjected to the voltage levels of several hundreds of volts during the evaluation of their operation [4], [5]. If a human body comes into contact with the device under test (DUT) during a high-voltage testing procedure, a harmful effects of high voltage can occur. The effect of a DC electric shock is determined by the amplitude of the current through the human body and the duration of the shock. The DC current amplitude of 300 mA is considered as a safety limit for a human body [6]. The DC voltage that will produce a current of a 300 mA through the human body depends on the human body resistance [7]. It can be seen in Fig. 1 that if 50% of human body serves as a current path, the DC voltage of approximately 400 V will cause a maximum allowable current of 300 mA through a human body. Therefore, the DC voltages typically used to test the operation of DC-DC converters based on SiC and GaN power devices are hazardous for a human.

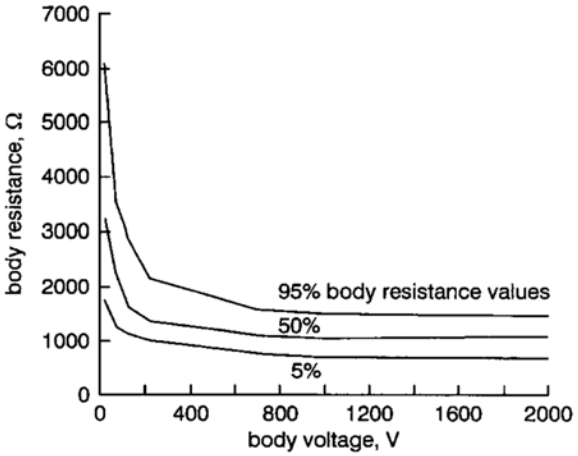


Fig. 1. Human body resistance as a function of body voltage [7].

In this paper, a protection system for high-voltage testing of SiC and GaN power switches in DC-DC converter applications is presented. The system is based on infrared (IR) detection of invasion into the high-voltage testing area. If the testing area is invaded, the protection system disconnects the high DC voltage supply from the DUT and the harmful effects of a DC voltage to a human body are avoided.

The overview of the developed infrared protection system is shown in Section II. Section III presents the evaluation of the protection system. Section IV concludes the paper.

II. Overview of the Protection System

A. Infrared Protection of the Testing Area

The top view of the high-voltage testing area is shown in Fig. 2. The testing area is enclosed by the 700-mm high fence attached to the 40-mm square tubes. The 820-mm wide entrance monitored by the IR protection system is left open in the fence for a user to arrange a measurement set-up. The mechanical sketch of the IR protected space is shown in Fig. 3. The 16 IR emitters [8] and 16 photodiodes [9] are uniformly distanced and placed in a "zigzag" order for effective monitoring of the entrance to the testing area. During the normal operation of the protection system, the photodiodes receive IR signal generated by the IR emitters. Once the IR protected space is invaded, the

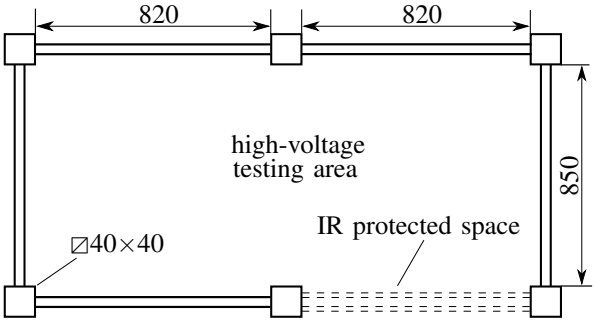


Fig. 2. Top view of the testing area. All dimensions are in milimeters.

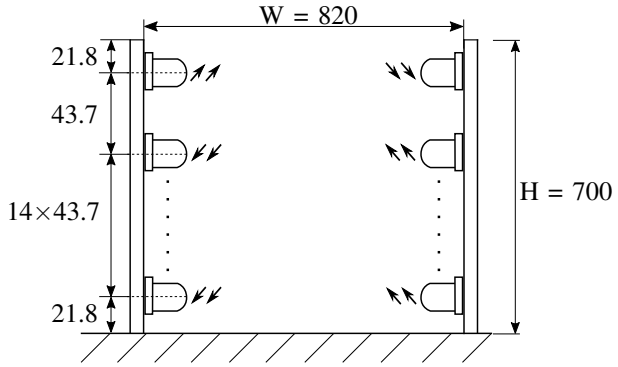


Fig. 3. Mechanical sketch of the space monitored by the IR protection system. All dimensions are in milimeters. The distances between the diodes are rounded to the first decimal.

transmission of the IR signal between the IR emitters and one or more photodiodes will be broken. This will cause a decrease of the current through the non-illuminated photodiodes large enough to produce a voltage difference that is detected by the input stage of the electronic control system.

B. Principle of Operation of the Protection System

The simplified schematic of the protection system shown in Fig. 4 can be divided into three function blocks:

- input stage that detects an invasion into the testing area,
- logic stage that sets the operation mode of the system and indicates the invasion of the testing area,
- output stage that disconnects the high-voltage DC supply from the DUT.

The input stage of the developed system consists of a voltage divider with variable resistors R_1 to R_{16} and reverse-biased photodiodes D_{PH_1} to D_{PH_16}. The current through each of 16 reverse-biased photodiodes I_{PH_i} $(i = 1, 2, ..., 16)$ is larger than zero when all the photodiodes are illuminated by the IR signal generated by the IR emitters. If any of the photodiodes looses the IR signal from the emitter, its current falls to zero. The voltage across each of 16 photodiodes is:

$$V_{PH_i} = V_{DD} - I_{PH_i} \cdot R_i, \quad i = 1, 2, ...16 \qquad (1)$$

where V_{DD} is a 5-V supply, V_{PH_i} is the voltage across the photodiode, I_{PH_i} is the current through the photodiode and R_i is the resistance of the corresponding variable resistor. When the photodiode is not illuminated by the IR signal, the voltage across it equals V_{DD}. The voltage acrosss the photodiode decreases as the photodiode gets more illuminated. The voltages V_{PH_i} are compared to the reference voltage V_{REF} using a 16 comparator circuits. The value of V_{REF} can be adjusted to achieve required sensitivity of the IR detection. The voltages at the comparator outputs V_{COMP_i} $(i = 1, 2, ..., 16)$ are:

$$V_{COMP_i} = \begin{cases} V_{DD}, & V_{PH_i} \leq V_{REF}, \\ 0\,\mathrm{V}, & V_{PH_i} > V_{REF}. \end{cases} \qquad (2)$$

Since invasion into the testing area has to be detected by any of 16 photodiodes, all 16 V_{COMP_i} signals are combined into one control signal using two 8-input NAND-gates and a 2-input OR-gate. The control signal V_{CTRL} is connected to the "set" input of the SR-latch as shown in Fig. 4. The relationship between the control signal V_{CTRL} and the V_{PH_i} is determined as:

$$V_{CTRL} = \begin{cases} 0\,\mathrm{V}, & V_{PH} \leq V_{REF},\ D_{PH_i}\ \text{illum.}, \\ V_{DD}, & V_{PH} > V_{REF},\ D_{PH_i}\ \text{not illum.} \end{cases} \qquad (3)$$

When the transmission of the IR signal to the either one of 16 photodiodes in interrupted, the control voltage V_{CTRL} rises to V_{DD}. This causes the turn-on of the MOSFET Q_1 and closing of the normally-open contact of the relay K_1 shown in Fig. 4. The relay K_1 Closing the contact of the K_1 will cause the closing of the normally-open contacts of the relay K_2 and disconnection of the high-voltage DC source V_{DC} from its power supply. The invasion into the testing area is indicated by the LED error indicator while disconnection of the V_{DC} from its power supply is indicated by the LED protection indicator shown in Fig. 4. The transistor Q_1 is turned off again by pressing the push-button S_{RST} and the V_{DC} is connected to its power supply.

The emergency stop switch in the output stage can be used independently of the part for the IR detection. Pressing the emergency stop switch will also cause the relay K_2 to disconnect the high-voltage DC source from its power supply.

The top view of the assembled electronic control circuit with indicated main blocks is shown in Fig. 5.

III. Evaluation of Infrared Protection System

The functionality of the control circuitry is verified by applying a short circuit between the inverting comparator input and a ground terminal to imitate a fully illuminated photodiode as shown in Fig. 6. When any of the comparator inputs are left open to imitate an invasion into the testing area, the control system triggers the on-state of the protection circuit and the voltage source that supplies the DUT is switched off.

In order to test the functionality of the complete protection system, the IR emitters and photodiodes are placed on the two 700-mm high tubes distanced by 800 mm as shown in Fig. 7. The system is calibrated by adjusting the variable resistors R_i $(i = 1, 2, ..., 16)$ shown in Fig. 4 in

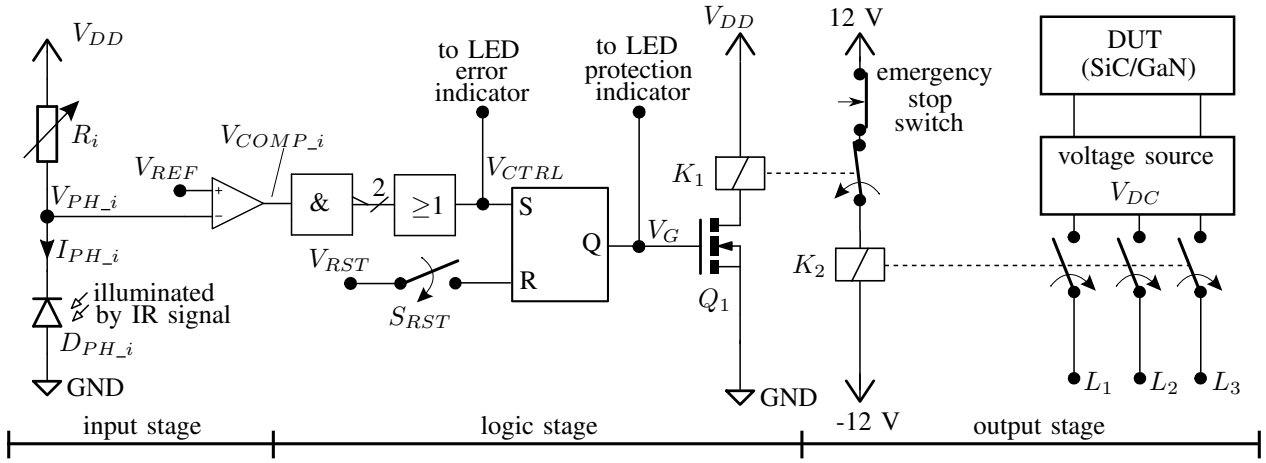


Fig. 4. Simplified schematic of the protection system. The system is divided into three function blocks: input, logic and output stage. The developed protection system contains 16 identical input stages ($i = 1, 2, ..., 16$). The array of 16 photodiodes is illuminated by IR signal generated by 16 IR emitters.

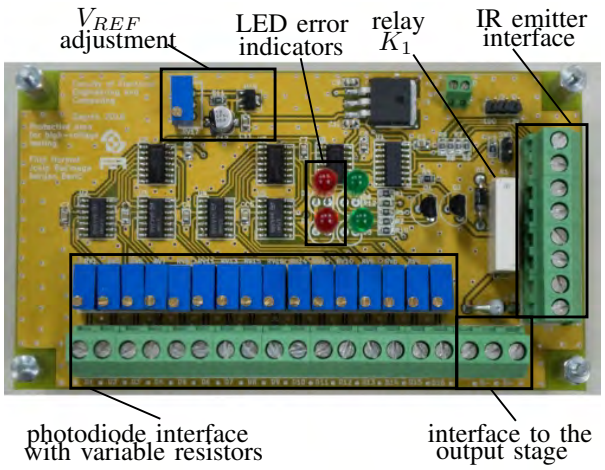


Fig. 5. The top view of the assembled control circuit.

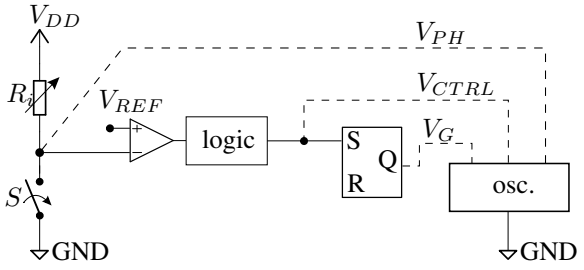


Fig. 6. Set-up to test the functionality of the control circuitry. The voltage waveforms are sensed using an oscilloscope (osc.).

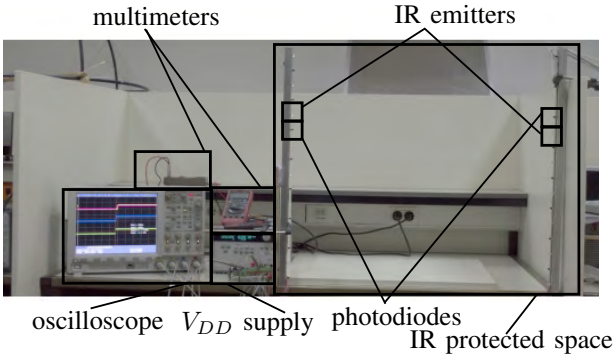


Fig. 7. Set-up to test the functionality of the complete protection system.

such a way that all the voltages across the photodiodes are the same during the normal operation of the system, i.e. when the testing area is not invaded. The reference voltage V_{REF} is adjusted to a value of 3.73 V that is high enough to avoid the impact of the noise on the illumination of the photodiodes caused by other sources of light such as daylight or room lightning.

The system is initially set in a normal mode of operation in which all the photodiodes are illuminated by the IR signal. The IR protected space is invaded and characteristic transient response timings of the protection system are measured. A pre-trigger of 1-5 ms is set at the oscilloscope to record the complete transition of the system response when the space protected by the IR emitters and photodiodes is invaded. All waveforms are recorded using the Agilent MSO7034B oscilloscope and Agilent 10073D passive probe configuration as follows: 1:1 ratio, DC coupling, bandwidth limit: OFF, input impedance: 10 MΩ.

The transient response waveforms of the control system when the IR protected space is invaded are shown in Fig. 8. The voltage across the photodiode V_{PH} rises to the value of V_{DD} after the invasion is detected. When V_{PH} crosses the value of V_{REF}, the voltages V_{CTRL} at the comparator output and V_G at the gate of the Q_1 instantaneously rise to the value of V_{DD}. The time-delay of the control circuitry, i.e. the time between the detection of the invasion into the testing area and start of the turn-on process of Q_1 is denoted as Δt_{CTRL1} in Fig. 8. The major part of Δt_{CTRL1} is the rise time of the voltage across the photodiode V_{PH} which depends on how fast an object passes through the IR protected space. If the object slowly invades into the testing area, it will take more time for the photodiode to become completely non-illuminated and for the V_{PH} to reach the V_{REF}. The time-delay of the digital

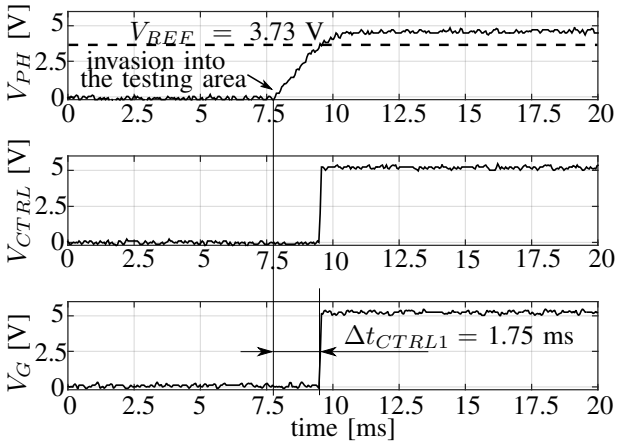


Fig. 8. Transient response of the control circuitry when IR protected space is invaded. Top to bottom: voltage across the photodiode, voltage at the output of the comparator and voltage at the gate of the Q_1.

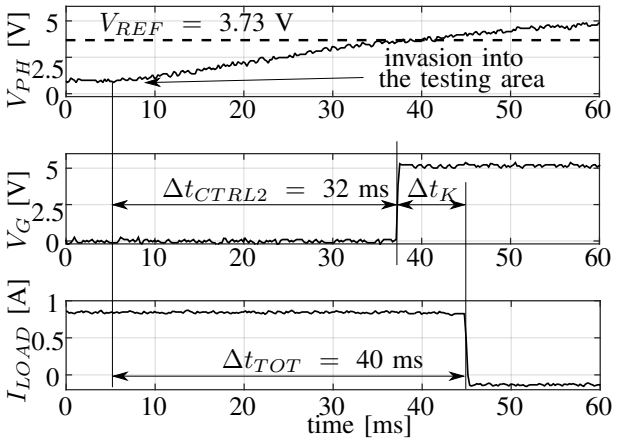


Fig. 9. Transient response of the complete system when IR protected space is invaded. Top to bottom: voltage across the photodiode, voltage at the output of the comparator and current through the dummy load.

logic parts and SR-latch are negligible.

A test case in which a dummy load used to imitate the output stage is being disconnected from the voltage source is designed to evaluate the response of the IR protection system. An invasion into the testing area is simulated and the voltage across the photodiode V_{PH}, voltage V_G at the gate of Q_1 and the current through the dummy load I_L are recorded and shown in Fig. 9. The total time-delay of the protection system Δt_{TOT} is the sum of the time-delay of the control circuitry Δt_{CTRL2} and the time Δt_K between the beginning of the turn-on process of Q_1 and fall of the I_L to 50% of its value. Since the simulated invasion into the testing area is much slower than for the test case shown in Fig. 8 in order to reproduce the worst-case-scenario, the time-delay of the control circuitry Δt_{CTRL2} is much larger than Δt_{CTRL1}. The value of Δt_K is independent of the speed of invasion into the testing area and equals 8 ms. The total time-delay Δt_{TOT} between the detection of the invasion into the testing area and break of the output high-voltage circuit is approximately 40 ms. The largest permissible duration of the electric shock with the most harmless effects on human body is 50 ms as specified in [7].

The power dissipation of all 16 IR emitters and 16 photodiodes when they are illuminated is 1 W while the dissipation of the control circuitry is approximately 35 mW.

IV. Conclusion

A protection system is developed to ensure safety of a user during a high-voltage testing of power devices such as SiC and GaN FETs used in switching DC-DC converters. The system is based on infrared detection of the invasion into the testing area in order to turn-off the high-voltage source that supplies the device under test. The functionality of the system is evaluated for different test cases and characteristic response timings are measured. The time-delay between the detection of an invasion and activation of the protection circuit is lower than the highest permissible duration of the electric shock that is specified by safety regulations.

Acknowledgment

This work was supported in part by the Croatian Science Foundation (HRZZ) within the project Advanced design methodology for switching dc–dc converters.

References

[1] J. Biela, M. Schweizer, S. Waffler, and J. W. Kolar, "SiC versus Si—Evaluation of Potentials for Performance Improvement of Inverter and DC–DC Converter Systems by SiC Power Semiconductors," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 7, pp. 2872–2882, July 2011.

[2] B. J. Baliga, *Fundamentals of Power Semiconductor Devices*, 1st ed. Springer, 2008.

[3] Narendra Mehta, *GaN FET module performance advantage over silicon*. Texas Instruments, March 2015, Application Note.

[4] O. Mostaghimi, N. Wright, and A. Horsfall, "Design and Performance Evaluation of SiC Based DC-DC Converters for PV Applications," in *IEEE Energy Conversion Congress and Exposition (ECCE)*, Sept 2012, pp. 3956–3963.

[5] R. Mitova, R. Gosh, U. Mhaskar, D. Klikic, M. Wang, and A. Dantella, "Investigations of 600-V GaN HEMT and GaN Diode for Power Converter Applications," in *IEEE Transactions on Power Electronics*, vol. 29, no. 5, October 2013, pp. 2441 – 2452.

[6] L. Gordon, "The physiological effects of electric shock in the pulsed power laboratory," in *Pulsed Power Conference, 1991. Digest of Technical Papers*, June 1991, pp. 377 – 380.

[7] C. H. Lee and A. P. S. Meliopoulos, "Comparison of touch and step voltages between IEEE Std 80 and IEC 479-1," *IEE Proceedings - Generation, Transmission and Distribution*, vol. 146, no. 6, pp. 593–601, Nov 1999.

[8] OSRAM. (2014) High Power Infrared Emitter SFH 4550 FA Product Details. [Online]. Available: http://www.osram-os.com/Graphics/XPic3/00116140_0.pdf

[9] ——. (2014) Silicon PIN Photodiode SFH 213 FA Product Details. [Online]. Available: http://www.osram-os.com/Graphics/XPic5/00101689_0.pdf

Optimal Conduction Angle of an E-pHEMT Harmonic Frequency Multiplier

Krunoslav Martinčić, Zagreb University of Applied Sciences, Electrical Engineering dept., Zagreb, Croatia
e-mail: krunoslav.martincic@tvz.hr

Abstract - This paper describes a classic method of an analog harmonic multiplier which uses a C class amplifier topology. The goal of this work is to propose the method of finding the optimal conduction angle of a E-pHEMT transistor so that a better spectral purity of the second or third harmonics can be obtained. MATLAB simulations are performed and the electronic circuit with an E-pHEMT transistor is designed. Measurements are made in both time and frequency domain and a comparsion of the results is presented.

I. INTRODUCTION

With the development of modern technologies, the characteristics of electronic components are greatly enhanced along with increasing operating frequency of electronic circuits. The rapid expansion of information and communication technology leads to the congestion of currently used frequency bands and also to a need for higher operating frequencies. Highly stable and in the harmonic-spectrum pure frequencies in higher GHz [1], [2] and THz [3] bands are practically impossible to generate directly in present time. Higher harmonic components of the fundamental frequency can be simply obtained by applying a sine voltage onto an electronic component having nonlinear current-voltage characteristics. Unfortunately, the byproduct of the wanted higher harmonics is also a large number of unwanted harmonic components of the fundamental frequency.

The paper describes a sine signal distortion method applied to finding the optimal waveform with the aim of gaining a single, specific higher harmonic component. Semiconductor electronic components possess highly non-linear current-voltage properties and as such are suitable for generating higher harmonics of fundamental frequency [4], [5]. The aim of this paper is to apply simulation tools and measurement of devices in order to obtain the defined harmonic component. The focus in this paper is not on the circuit topology, conversion efficiency, bandwidth or optimisation of high frequency matching under the non-linear operating conditions of semiconductor devices.

II. QUANTITATIVE MODEL

The quantitative model proposed here can be applied to any harmonically pumped nonlinear system. If we focus solely on the concept of multiplication of a harmonic frequency of a sine signal, then the objective is to obtain one particular, higher harmonic kω (k ∈ N, k > 1), with the power P(kω) derived from the initial signal. Nonlinear sine wave distortion generates a multitude of harmonic products nω (n = 1, 2, 3, ...). It is possible to extract the target component kω from such a spectrum by using a suitable filter. It is better if the side spectrum components: (k –2)ω, (k –1)ω, (k+1)ω and (k+2)ω are as small as possible. In this way, the unwanted components of the spectrum can be well suppressed by simpler, low order filters. Since the concept of spectral purity in literature is frequently linked to the notion of phase noise of an oscillator, the quantitative factor that describes the issues studied here can be named Harmonic Spectra Purity Factor (HSPF). Similarly to the distortion of linear systems, HSPF can be defined via the power of spectral components as:

$$HSPF(\omega_k) = \frac{P(\omega_k)}{\sum_{i=1}^{\infty} P(\omega_i)}; \quad i,k \in \mathrm{N}, k > 1, \qquad (1)$$

where P(kω) = P(ω_k) is the power of the k-th harmonic ω_k.

If the signal only contains the component ω_k, then HSPF(ω_k) = 1, i.e. the signal is 100% spectrally "clean". The same method of the quantitative description can be applied to other nonlinear systems, e.g. to mixers. Symmetrical (balanced) structures will suppress some of the unwanted products better than unbalanced structures. Also, with the same topology, a well balanced circuit will have higher HSPF than an unbalanced circuit.

The relation (1) is unsuitable for application, therefore, in practice, it can be used in a simplified form taking into account the spectrum of only two or four symmetric octaves with respect to ω_k. The components of the spectrum which are more distant from ω_k are more easily filtered out and generally have smaller amplitudes. HSPF2 and HSPF4, defined by two and four octaves, are expressed in Equations (2) and (3). For rational sums index, the nearest lower whole number (integer) is to be taken.

$$HSPF2(\omega_k) = \frac{P(\omega_k)}{\sum_{i=\frac{k}{2}}^{2k} P(\omega_i)}; \quad i,k \in \mathrm{N}, k > 1. \qquad (2)$$

$$HSPF4(\omega_k) = \frac{P(\omega_k)}{\sum_{i=\frac{k}{4}}^{4k} P(\omega_i)}; \quad i,k \in \mathrm{N}, k > 1. \qquad (3)$$

In a specific situation, HSPF will be represented by the function of a selected electronic component, static operating point, circuit topology, the amplitude of the driving signal and the temperature.

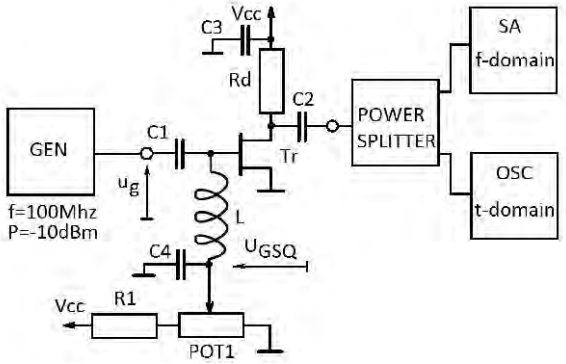


Fig.1, Measurement setup

III. PHYSICAL REALIZATION

Due to the simplicity in the setting of quiescent condition and theoretically simple transfer characteristic, the circuit is realized by using an enhanced pHEMT [6] transistor of high transconductance. The schematic of the circuit and the measurement setup is shown in Fig. 1. Firstly, the static transfer characteristic of the transistor at $U_{ds} = 4$V has been measured. HF and microwave transistors generally possess a narow gate and, consequentially, linear transfer characteristic in the saturation region. Due to nonlinearity around the threshold voltage, U_{GS0} (V_{TH}), the transfer characteristic is interpolated by a semi parabola:

$$I_d = 0.3\,(U_{gs} - 0.27)^2 \text{ A}, \quad U_{gs} > 0.27 \text{ V and} \quad (4)$$

$$I_d = 0, \quad U_{gs} \le 0.27 \text{ V} \quad (5).$$

The gate voltage U_{gs} is a sum of the DC bias U_{GSQ} and AC driving signal u_g (Fig. 1):

$$U_{gs} = U_{GSQ} + u_g \quad (6).$$

The measured and interpolated transfer functions are shown in Fig. 2. Some smaller differences are observed between the two curves, therefore, some deviations are

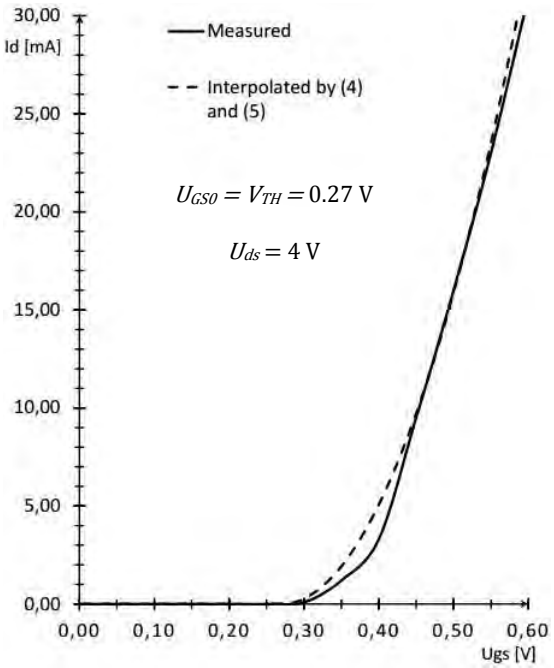


Fig. 2, E-PHEMT, Id(Ugs), Uds = const.

expected in the measurement results compared to the simulation results. In the realistic setting of the dynamic conditions, there are parasitic serial resistances, capacitances, and inductances in the active component which influence the shape of the signal in the time-domain as well as the spectrum. Therefore, one can expect a small difference between the data obtained from simulation and the results of measurement, especially at higher frequencies.

According to (4), (5) and (6), I_d will depend on the gate bias voltage U_{GSQ} and the amplitude of the driving signal u_g. The three fundamentally different operating modes are possible:

a) $U_{GSQ} \le U_{GS0}$, C-class, the transistor conducts for less than half of a period, with low side clipping,
b) $U_{GSQ} > U_{GS0}$, non-linear operation, the transistor conducts for more than half of a period, depending on the amplitude of the driving signal, clipping is possible both on the bottom and the top side,
c) $U_{GSQ} >> U_{GS0}$, depending on the amplitude of the driving signal, mainly saturation, and high side clipping.

In the frequency multiplication circuits with transistors, the most commonly used mode is a C-class since such a waveform contains a multitude of higher harmonic components of the spectral distribution [7]:

$$A(\omega_k, t_0) \sim \omega t_0 \left| \frac{\cos\left(\frac{k\omega t_0}{2}\right)}{1 - (k\omega t_0)^2} \right|, \quad (7)$$

$$\omega = 2\pi f, \;\; T = \frac{1}{f}, \;\; t_0 \le \frac{T}{2}, \;\; \theta = 2\pi \frac{t_0}{T}, \;\; k \ge 1,$$

$$t_0 - Conduction\ Time, \;\; \theta - Conduction\ Angle\ .$$

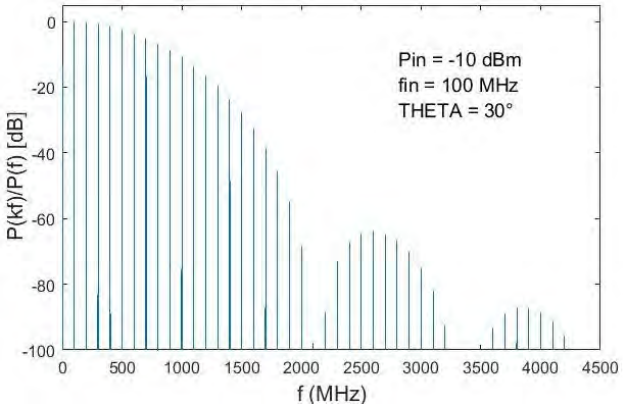


Fig. 3, Spectrum normalized to $P(f_1)$ for $\theta = 30°$

MATLAB simulation (FFT) of spectrum for narrow half-sinusoidal pulse train ($\theta << 2\pi$) is shown in Fig. 3 and for large conduction angle ($\theta \approx \pi$) in Fig. 4. The waveform and the spectrum depend on the gate bias voltage U_{GSQ} and the amplitude of the driving signal u_g. According to (4), (5) and (6) both of these variables unambiguously determine the conduction angle θ of the transistor. If the amplitude of the driving signal u_g is held constant and the gate bias voltage U_{GSQ} is being changed (Fig. 1), the clipping level and

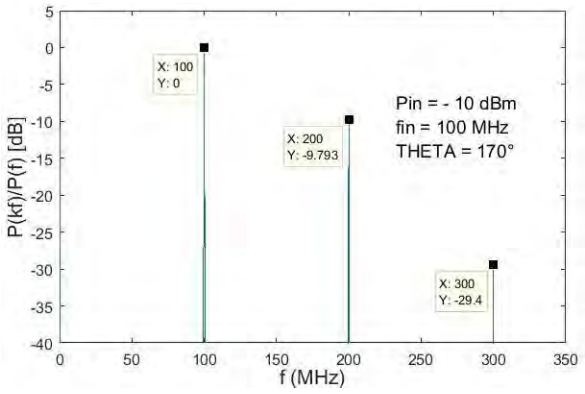


Fig. 4, Spectrum normalized to $P(f_1)$ for $\theta = 170°$

consequently the conduction angle can be adjusted. For this reason, the simulation and the spectrum analysis by MATLAB have been done at a conduction angle θ as an independent variable.

IV. Simulations and Measurements

With the constant amplitude u_g, and by changing the gate bias voltage U_{GSQ}, with the aid of a resistive divider R_1 - POT_1, according to Fig. 1, it is possible to adjust the

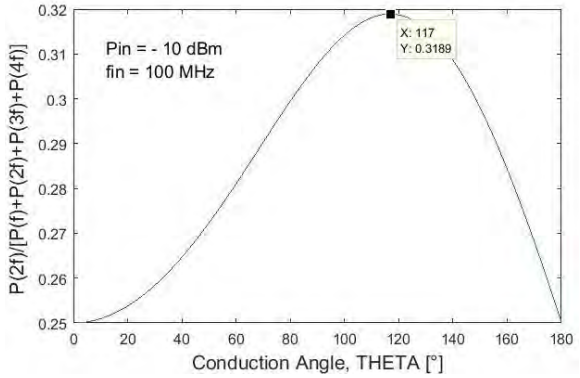


Fig. 5, $HSPF2(\omega_2)$ as a function of θ

conduction angle from 0º to 360º. With smaller conduction angles, the total power of the spectral components is smaller, too. The aim of this work is not the optimisation of power or efficiency, but the optimisation of the spectral composition. Once the wanted component has been separated from the spectrum and the unwanted components have been sufficiently suppressed, the signal level can be amplified to the defined level by a linear amplifier. In Fig. 4 ($\theta = 170°$), the fundamental harmonic dominance in the spectrum can be observed. With a frequency doubler, it would be desirable that the amplitudes of the fundamental and third harmonics are as low as possible. The simulation of the conduction angle ranging from 0º to 180º has been performed in MATLAB. Fig. 5 shows $HSPF2(\omega_2)$ as the function of the conduction angle. The optimal ratio of the amplitudes of spectral components in terms of generating the second harmonic occurs at a conduction angle of $\theta \approx 117°$, i.e. one third of the period. This case is shown for time domain in Fig. 6, and for frequency domain in Fig. 7. There is a relative increase of the amplitude of the second harmonic with respect to the amplitude of the fundamental component

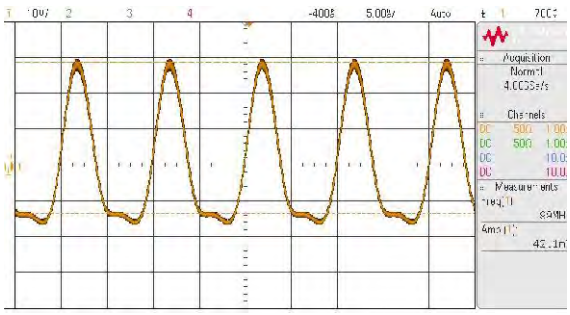


Fig. 6, $I_d(t)$, $\theta = 120°$, $U_{GSQ} = 0.12$ V, $P_{in} = -10$dBm, $f_{in} = 100$MHz

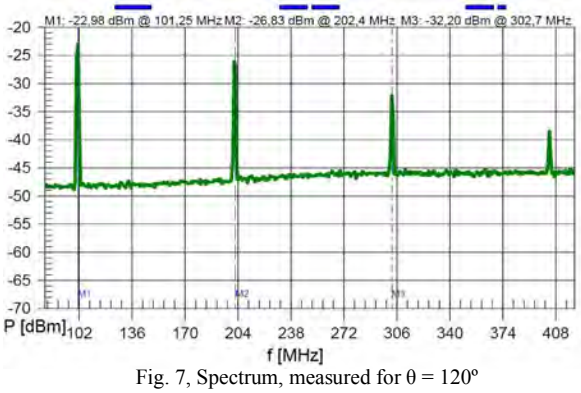


Fig. 7, Spectrum, measured for $\theta = 120°$

compared to the prior case of the conduction angle $\theta = 170°$, Fig. 4.

The same procedure has been performed in the case of a frequency tripler, and Fig. 8 shows the obtained

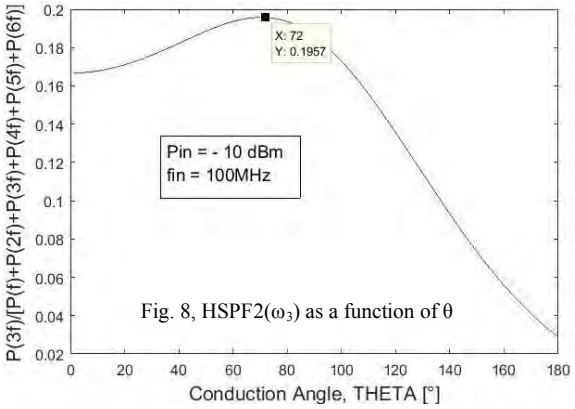


Fig. 8, $HSPF2(\omega_3)$ as a function of θ

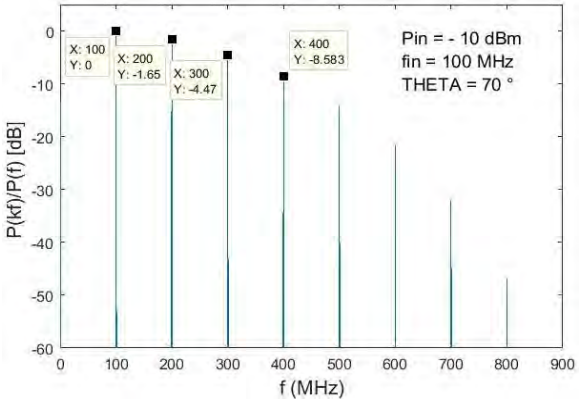


Fig. 9, Spectrum normalized to $P(f_1)$ for $\theta = 70°$

dependence of $HSPF2(\omega_3)$ on the conduction angle. The graph shows the maximum for a conduction angle of 71°,

i.e. ≈ 1/5 of the period. Fig. 9 shows the spectrum obtained by computer simulation for the referred case. From the graphs a relative increase can be observed in the power of the third harmonic compared to the first and second harmonic compared to the prior case of the frequency doubler and conduction angle $\theta = 120°$.

V. CONCLUSION

This paper presents a model for quantitative description of harmonically pumped nonlinear systems. With the aid of MATLAB, an analysis is conducted, including the calculation of the proposed factors for the case of a frequency multiplier implemented by using E-pHEMT in C-class amplifier. The measurement results have confirmed the computer simulation results and have also justified the use of the proposed factors. In the future work, the intention is to carry out an analysis of circuits with an exponential current-voltage characteristic, as well as of symmetrical topologies with suppression of the fundamental component.

REFERENCES

[1] Yuan Chun Li, "20–40 GHz dual-gate frequency doubler using 0.5 μm GaAs pHEMT technology", IEEE, Electronics Letters, Volume: 50, Issue: 10, 2014.

[2] Rauscher, C., "High Frequency Doubler Operation GaAs Field Effect Transistors", IEEE Trans. On Microwave Theory and Techniques, Vol. 31, No. 8, June 1983,

[3] T.W. Crowe, J.L. Hesler, S.A. Retzloff, C. Pouzou, G.S. Schoenthal, "Solid-State LO Sources for Greater than 2THz", Virginia Diodes Inc., Charlottesville, VA, 2011 ISSTT Digest, 22nd Symposium on Space Terahertz Technology, April 26th-28th, 2011, Tucson Arizona, USA, www.vadiodes.com

[4] S. A. Maas, "The RF and Microwave Circuit Design Cookbook", Artech House, Inc., Norwood, MA, 1998.

[5] Edmar Camargo, "Design of FET Frequency Multipliers and Harmonic Oscillators", Artech House, Norwood, MA, 1998.

[6] Samuel Y. Liao, "Microwave Devices and Circuits", Third Edition, Prentice Hall, Englewood Cliffs, NJ, 1996.

[7] S. A. Maas, "Nonlinear Microwave and RF Circuits", Artech House, Inc., Norwood, MA, 2003.

Ultra-Wideband Transmitter Based on Integral Pulse Frequency Modulator

T. Matić*, M. Herceg*, J. Job* and L. Šneler**
* Faculty of Electrical Engineering Osijek, Osijek, Croatia
** Supracontrol, Zagreb, Croatia
tmatic@etfos.hr

Abstract - The paper presents a novel short-range wireless sensor node architecture, based on Integral Pulse Frequency Modulator (IPFM) and Ultra-Wideband (UWB) pulse generator. Due to the lack of internal clock signal source and multi user implementation without application of a microprocessor, the architecture is simple and energy efficient. Multi-user coding is performed using delay elements, where each user has unique delay time value. The output of the IPFM is fed to the delay element. Delayed and original signal from the IPFM are feeding UWB pulse generator. The paper presents the transmitter architecture and measurements of the transmitter signals in time and frequency domain.

I. INTRODUCTION

Wireless sensor networks have gained great attention in last two decades. Following development of the semiconductor technology and modern CMOS mixed signal circuit design, the wireless sensor networks are becoming the core part of the various modern communication systems [1-4]. Particularly important is its application in the systems where the node size and the power consumption is the critical demand, like Wireless Body Area Networks (WBAN) [5-8]. In such networks, the size of the wireless node and its power consumption are limited and it is important to achieve minimization and lowest possible consumption [9].

Ultra-wideband (UWB) communication systems have shown great promise in low power communication systems. Since the UWB communication systems are based on the narrow pulses, it provides energy consumption below 1 nJ per transmitted pulse. Depending on the application areas, for slowly varying signal measurements, where low duty cycle operation is enabled, wireless sensor node can provide extremely low power operation. In such a system, there is a promising application of energy harvesting power supplies that would enable battery-free operation.

Wireless sensor nodes aimed to acquire a signal from an analog sensor are relaxing in terms on computational complexity. Due to analog input signal, they could consist of the simple analog-to-digital converter followed by an UWB generator employing suitable modulation for multi-user coding. Such a sensor could employ simple 1-bit analog-to-time conversion and delay based multiuser coding, as it is proposed in [10 and 11]. The proposed multi-user coding is based on transmitted reference ultra-wideband modulation. The major difference is that it employs asynchronous transmission and no clock nor microprocessor is required at the transmitter side.

Clock-less analog to time conversion can easily be implemented with Time Encoding Machine (TEM) [12] at the analog sensor output. The analog input signal is transformed to output pulse train that contains the information on analog input signal in time domain, like output pulse frequency, distance or duty cycle, depending on the type of TEM [11].

For energy efficient time-encoding, the best choice is pulse-based TEM, like Integrate and Fire (IAF) or Integral Pulse Frequency Modulator (IPFM) that transform analog input signal to pulse train with information in pulse distance/frequency. The maximum output signal value is present only during short term pulse duration, while the rest of the period output signal is equal to zero. Such an operation enables low duty-cycling and therefore low energy consumption.

In next chapter, the paper presents the TEM theory, focusing on IPFM application for UWB wireless sensor application. The UWB IPFM transmitter is implemented as discrete-type laboratory prototype which can be used as a wireless sensor node at the analog sensor output. The third chapter present the measurement results.

II. IPFM UWB TRANSMITTER

A. Transmitter architecture

The presented wireless transmitter is aimed for the application at the analog sensor output for remote signal acquisition. As it can be seen form Fig. 1, it transforms analog input signal to time information using TEM as a modulator. The pulse train $y(t)$ is fed to the delay line that is used for multi user coding and optionally fed back to IPFM. Each sensor has unique delay value that is used for user identification at the receiver side. The delayed and original pulse train ($y'(t)$ and $y(t)$ signals) are fed to the UWB pulse generator where the pulse pair is formed for direct wireless transmission. The ultra-wideband generator can optionally be fed to the transmission line or antenna to achieve wireless or wired communication.

In case of wireless transmission, the pulses are in high frequency bands over 3 GHz frequencies, while in wired mode communication, the frequency should be significantly lower to minimize losses and reflections.

B. Time Encoding Machine

The Time Encoding Machines transform the analog input signal to information at the output. The basic TEM circuits, according to [12] are Asynchronous Sigma-Delta Modulator (ASDM) and Integrate and Fire (IAF) modulator. Beside ASDM and IAF, Integral Pulse Frequency Modulator can also be considered as a TEM [11], since its consecutive output pulses distance is proportional to analog input voltage value.

For ASDM application in TEM, the information on the analog input voltage is transformed to output pulse train duty cycle [13]. Therefore, the information on rising and falling edge is required for analog information recovery. Due to more complex transmitter and receiver architecture for transmission of positive and negative pulses, IPFM is more attractive for transmitter implementation. Since information is carried only in pulse distances, only positive UWB pulses can be transmitted for reliable analog signal demodulation at the receiver side.

The block scheme of the Integral Pulse Frequency Modulator is depicted in Fig.2. The analog input signal defines the comparator threshold voltage. To enable unipolar supply for bipolar input voltages, the input voltage span is transformed from [$-V_{CC}$, V_{CC}] to [0, V_{CC}], where $c = 0.5$, while S_{ref} is equal to $V_{CC}/2$.

The integrator $L(s)$ from the Fig. 2 constantly integrates until the integrator output $l(t)$ reaches comparator threshold. The transition of the comparator output $y(t)$ occurs and feedback loop signal $y'(t)$ resets the integrator to integrate from zero voltage till the next crossing occurs. The delay block can be implemented outside the feedback loop or inside, like in Fig.2. The consecutive pulse distance is proportional to the comparator threshold $u_{TH}(t) = x(t)/2 + V_{CC}/2$ for ideal integrator $L(s)$ with time constant τ_i. The integrator output signal in time domain can be expressed by the following equation:

$$l(t_k) = \int_{t_{k-1}}^{t_k} \frac{V_{CC}}{\tau_i} dt + C_{k-1}. \quad (1)$$

At the time t_k when integrator triggers comparator, under consumption that at the beginning of the integration slope t_{k-1}, the integrator output was equal to 0 ($l(t_{k-1}) = C_{k-1} = 0$) the following equation holds:

$$l(t_k) = \int_{t_{k-1}}^{t_k} \frac{V_{CC}}{\tau_i} dt = u_{TH} = \frac{x(t_k)}{2} + \frac{V_{CC}}{2}, \quad (2)$$

if $\tau_i << (t_k - t_{k-1})$. According to (2), the distance between consecutive output pulses T_k and T_{k-1} of the signal $y(t)$, if $\tau_n = 0$ is equal to:

$$T_k - T_{k-1} = t_k - t_{k-1} + T_{ON} + T_{OFF} = \\ = \frac{\tau_i}{2V_{CC}} x(t_k) + \frac{\tau_i}{2} + T_{ON} + T_{OFF}, \quad (3)$$

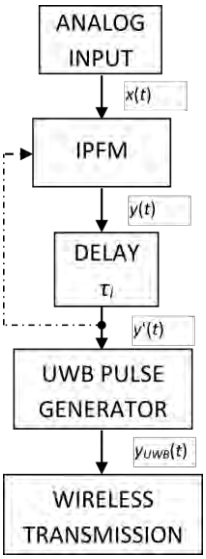


Figure 1. Block scheme of the IPFM UWB transmitter

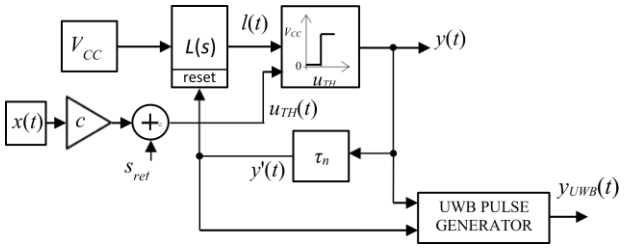


Figure 2. Block scheme of the IPFM modulator

where T_{ON} and T_{OFF} are on and off switching times of the integrator $L(s)$ and their sum is equal to the output pulse width ($T_P = T_{ON} + T_{OFF}$). For $\tau_n \neq 0$, the equation (3) becomes:

$$T_k - T_{k-1} = \frac{\tau_i}{2V_{CC}} x(t_k) + \frac{\tau_i}{2} + 2\tau_n + T_{ON} + T_{OFF}. \quad (4)$$

If integrator time constant τ_i, multi user delay τ_n and switching delay times T_{ON} and T_{OFF} are known at the receiver side, according to (4), it's easy do demodulate pulse train and to obtain analog input voltage value $x(t_k)$. The difference in two multi user delays τ_n - τ_{n-1} should be high enough that variation in integrator time constant τ_i, delay times T_{ON} and T_{OFF} does not affect detection at the receiver side.

C. Circuit implementation

The pulse train $y(t)$ is applied to the delay line which generates delayed replica $y'(t)$. In that way, two pulse trains are formed to feed the UWB pulse generator for the wireless pulse transmission. The discrete type prototype is implemented, based on modified version of UWB generator published in [14]. Modification allows impulse pair generation, independently on the rectangular pulse width.

The implemented circuit is depicted on Fig.3. Instead of the operational amplifier (OPAMP) implementation, the simple RC-filter is used as integrator. One OPAMP is used as comparator within IPFM, while the other one is used as a comparator within delay circuit. Both

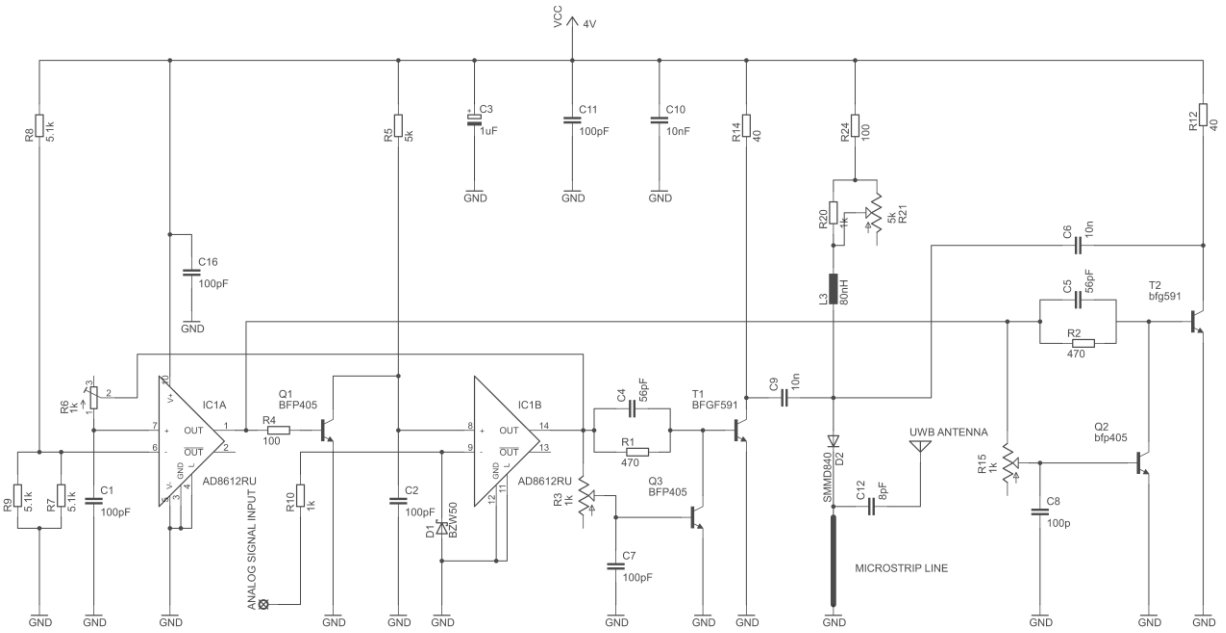


Figure 3. The circuit implementation of the IPFM UWB transmitter

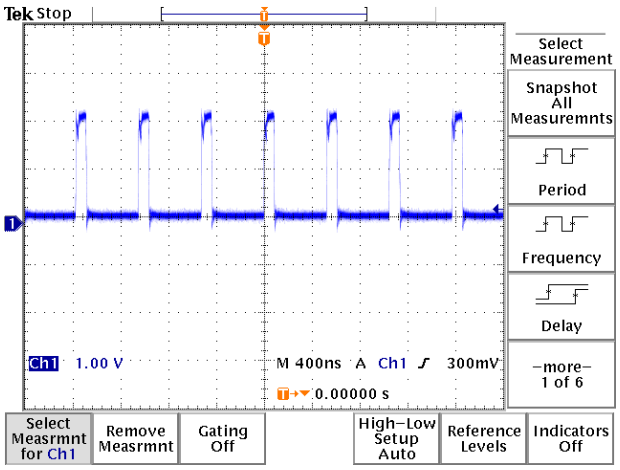


Figure 4. The output of the comparator $y'(t)$

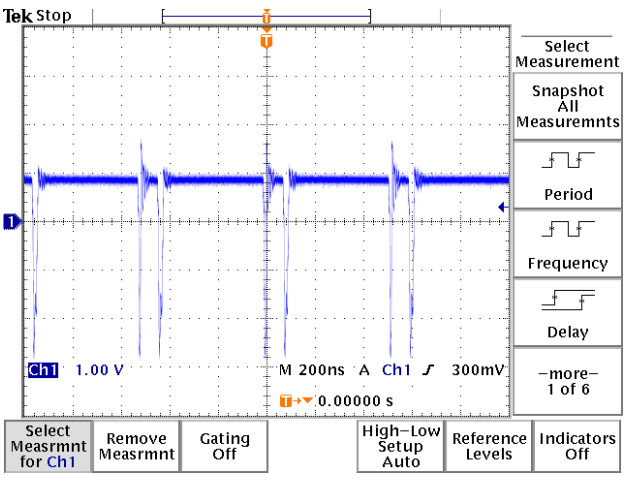


Figure 5. The step recovery diode voltage

comparators trigger UWB generator inputs that are implemented with high frequency transistors that trigger step recovery diode connected to microstrip line. At each rising edge of both original and delayed pulse trains, the UWB generator will form an ultra-wideband pulse which are summed and form signal $y_{UWB}(t)$ to be sent wirelessly. The microstrip line is connected to antenna via capacitor. Since analog input voltage is unipolar signal, constant c = 1 and S_{ref} is not connected.

III. RESULTS

The laboratory prototype from Fig.3 was implemented in discrete components implementation and for UWB pulse generation step recovery diode has been used. Due to constant forward biasing and constant current consumption this implementation is not energy efficient. To achieve low power operation IC based pulse generation techniques have to be used to ensure low duty cycling and zero output current in excess of UWB pulse [15 and 16]. The delayed comparator output pulse train $y'(t)$ is depicted in Fig 4. Information on the input signal level is contained in the pulse distance $T_k - T_{k-1}$, which is equal to 500 ns in the present setup.

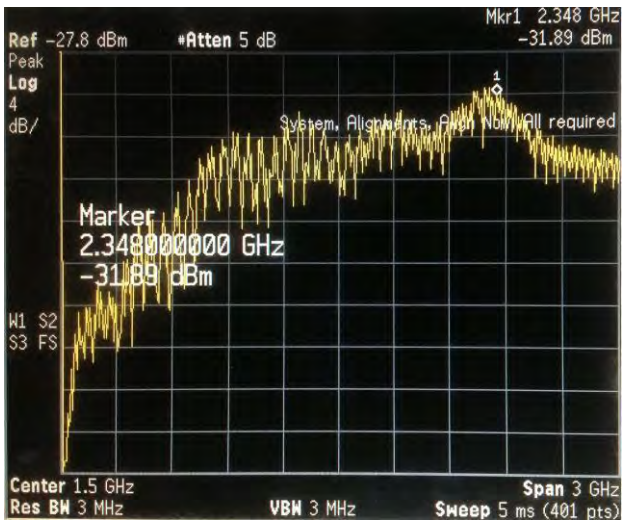


Figure 6. The output spectrum of the UWB pulse train

Fig.5. presents the step recovery diode voltage. It is visible that the pulse pair reversely polarizes diode. At each reverse polarization pulse step recovery diode will

form the narrow Gaussian pulse. The spectrum of the transmitted pulse is measured with receiving antenna at the 30 cm distance without LNA. Received power spectrum is depicted on Fig. 6. Due to limitations of the step recovery diode based UWB pulse shaper, the pulse in not positioned entirely in unlicensed spectrum.

IV. CONLUSION

The paper presents novel UWB IPFM transmitter for short range wireless communications. Architecture of the transmitting system is achieved without internal clock application nor microprocessor application for multi-user coding. Unique delay enables simple and efficient multi-user coding and UWB pulses provide low energy consumption per transmitted pulse. To achieve low power operation, future work is directed to CMOS IC implementation of the system which will enable low power operation in periods of non-activity between two consecutive UWB pulses. Downscaling the pulse width to sub 10 ps range, the integrator time constant could be lowered to enable integration of the capacitors on silicon as well. Also, IC design will provide more efficient spectral shaping and spectral positioning in unlicensed spectrum over 3 GHz due to delay-based UWB pulse generation and more accurate pulse shaping comparing to SRD implementation.

ACKNOWLEDGMENT

This work has been supported in part by Croatian Science Foundation under the project UIP-2014-09-6219 Energy efficient asynchronous wireless transmission.

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “A survey on sensor networks,” IEEE Commun. Mag., vol. 40, no. 8, pp.102–114, Aug. 2002.

[2] Li-Yuan Chang; Pei-Yin Chen; Tsang-Yi Wang; Ching-Sung Chen, "A Low-Cost VLSI Architecture for Robust Distributed Estimation in Wireless Sensor Networks," in Circuits and Systems I: Regular Papers, IEEE Transactions on , vol.58, no.6, pp.1277-1286, June 2011.

[3] Bellasi, D.E.; Benini, L., "Energy-Efficiency Analysis of Analog and Digital Compressive Sensing in Wireless Sensors," in Circuits and Systems I: Regular Papers, IEEE Transactions on , vol.62, no.11, pp.2718-2729, Nov. 2015.

[4] Bellasi, D.E.; Rovatti, R.; Benini, L.; Setti, G., "A Low-Power Architecture for Punctured Compressed Sensing and Estimation in Wireless Sensor-Nodes," in Circuits and Systems I: Regular Papers, IEEE Transactions on , vol.62, no.5, pp.1296-1305, May 2015.

[5] S. Diao, Y. Zheng, and C.-H. Heng, “A CMOS ultra low-power and highly efficient UWB-IR transmitter for WPAN applications,” IEEE.Trans. Circuits Syst. II, Exp. Briefs, vol. 56, no. 3, pp. 200–204, 2009.

[6] Poon, C.C.Y.; Lo, B.P.L.; Yuce, M.R.; Alomainy, A.; Yang Hao, "Body Sensor Networks: In the Era of Big Data and Beyond," in Biomedical Engineering, IEEE Reviews in , vol.8, no., pp.4-16, 2015.

[7] Sarkar, S.; Misra, S., "From Micro to Nano: The Evolution of Wireless Sensor-Based Health Care," in Pulse, IEEE , vol.7, no.1, pp.21-25, Jan.-Feb. 2016.

[8] Ozols, K., "Implementation of reception and real-time decoding of ASDM encoded and wirelessly transmitted signals," in Radioelektronika (RADIOELEKTRONIKA), 2015 25th International Conference , vol., no., pp.236-239, 21-22 April 2015.

[9] Sarpeshkar, R., "Universal Principles for Ultra Low Power and Energy Efficient Design," in Circuits and Systems II: Express Briefs, IEEE Transactions on , vol.59, no.4, pp.193-198, April 2012.

[10] T. Matic, M. Herceg, J. Job, “Energy-efficient system for distant measurement of analogue signals”, WO/2014/195739, 11.12.2014.

[11] T. Matic, M. Herceg, J. Job, “Energy-efficient system for distant measurement of analogue signals”, WO/2014/195744, 11.12.2014.

[12] A. A. Lazar and L. T. Toth, “Time encoding and decoding of a signal”, US 7,573,956 B2, 11 August, 2009.

[13] Ouzounov, S.; Engel Roza; Hegt, J.A.; van der Weide, G.; van Roermund, A.H.M., "Analysis and design of high-performance asynchronous sigma-delta Modulators with a binary quantizer," in Solid-State Circuits, IEEE Journal of , vol.41, no.3, pp.588-596, March 2006.

[14] P. Protiva, J. Mrkvica, J. Macháč, Universal Generator of Ultra-Wideband Pulses, Radioengineering, vol. 17, no. 4, December 2008.

[15] Xie, H.L.; Wang, X.; Wang, A.; Qin, B.; Chen, H.; Zhao, B., "An ultra low-power low-cost gaussian impulse generator for uwb applications," in Solid-State and Integrated Circuit Technology, 2006. ICSICT '06. 8th International Conference on , vol., no., pp.1817-1820, 23-26 Oct. 2006.

[16] Radic, J.B.; Djugova, A.M.; Videnovic-Misic, M.S., "A 3.1–10.6 GHz impulse-radio UWB pulse generator in 0.18μm," in Intelligent Systems and Informatics (SISY), 2011 IEEE 9th International Symposium on , vol., no., pp.335-338, 8-10 Sept. 2011.

Design of a transmitter for high-speed serial interfaces in automotive micro-controller

A. Bandiziol*, W. Grollitsch**, F. Brandonisio**, R. Nonis**, P. Palestri*
* Università degli Studi di Udine, Udine, Italy
** Infineon Technologies, Villach, Austria
palestri@uniud.it

Abstract - This work reports about the system level design of a *transmitter* for the next generation of High-Speed Serial Interfaces (HSSI) to be implemented in a micro-controller for automotive Electronic Control Unit (ECU) applications, pushing the transmission speed up to 10 Gbps over a 10cm long cable. A *voltage mode* architecture is selected for *low power* considerations. We focus our analysis here on the system-level implementation of Feed-Forward Equalization as an FIR filter consisting of different transmitter slices driven by different bits in the data sequence. We consider different data rates and number of taps and analyze how the performance of the equalizer is affected by the *quantization* of the values of the taps in the practical implementation of the FIR filter.

I. INTRODUCTION

High-speed digital I/O is becoming increasingly popular and emerged as one of the hot research topics in microelectronics. In this context, high-speed serial links are widely used in a variety of different fields [1]. Moreover, in the last twenty years the speed of these links has been constantly increasing, leading to the adoption of a broad variety of standards [2]. Since the baud rates at which these links work today are extremely high, inter-symbol interference (ISI) has become the most limiting factor. In order to mitigate ISI, the most effective solution is channel equalization [3]. Equalization can be done either at the transmitter side just before the channel (and it is called Feed-Forward Equalization, FFE) or at the receiver (for example in the form of Decision-Feedback Equalization, DFE). Figure 1 shows the general structure of a transmitter with feed-forward equalization: delayed versions of the bit stream are fed to drivers with different strengths, thus implementing an FIR filter.

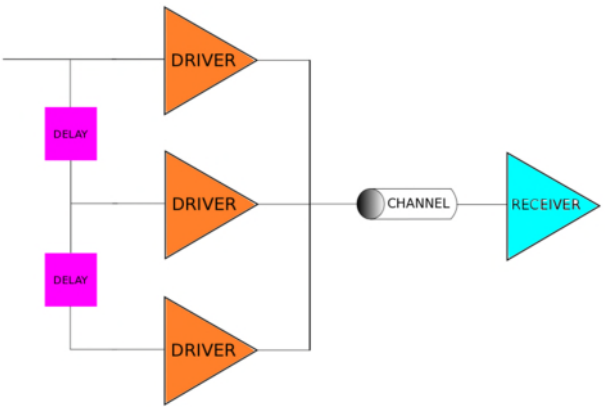


Figure 1. General scheme of a serial link implementing FFE using driver slices

II. DRIVER ARCHITECTURE

For the transmitter, a voltage mode architecture has been chosen, because it consumes less power than a current mode with the same output swing [4]. Figure 2 shows the general structure of a voltage-mode differential driver: two inverters are driven by bit b_i and $\overline{b_i}$ to produce a differential voltage v_o. The MOSFETs are large enough to have a negligible voltage drop when on. Impedance matching is implemented by the resistances R_{Di}.

Figure 2 is also the starting point to explain the basic principle of FFE. A single driver can be split into many slices, and the same bit b_i of the serial data input can drive many of these slices, i.e. the different drivers in Figure1 are each formed by many slices with the schematics of Figure 2. This can be better understood in Figure 3. The index "i" of a bit indicates the position in the bit stream.

When dividing the driver into slices, one must still guarantee impedance matching. In fact, when FFE is not implemented, only one slice might be used, and its output resistance should match R_0=50Ω. If more slices are used, it has to be considered that the total resistance of all the slices in parallel should be 50Ω, leading therefore to bigger values for R_{Di} of each slice. This results in the following equation

$$\sum_i 1/R_{D_i} = 1/R_0 \qquad (1)$$

Under this assumption, one can then write a closed form expression for the output voltage v_o, which reads

$$v_0 = \frac{V_{DD}}{2} \cdot \sum_i (n_i/M) \cdot b_i \cdot \mathrm{sgn}_i \qquad (2)$$

where (see Figure 3) n_i is the number of slides connected to the bit b_i, M is the number of total slices and sgn_i is the sign of b_i (i.e. putting b_i at the left or right part of Figure 2 making the driver inverting or non-inverting). For the b_i, '1' corresponds to b_i =1 and '0' means b_i =-1.

Eq. (2) can be rewritten as

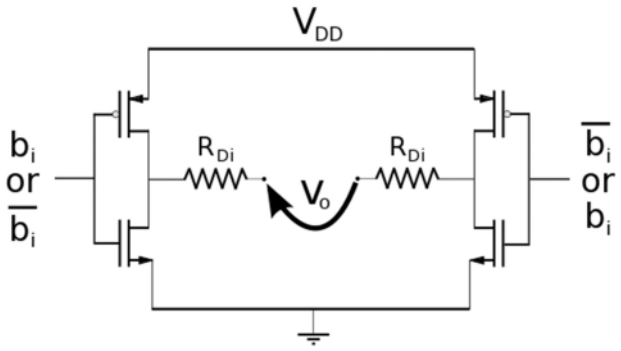


Figure 2. General scheme of a voltage-mode transmitter.

$$v_0[j] = \frac{V_{DD}}{2}\sum_i w_i \cdot b_i[j] = \frac{V_{DD}}{2}\sum_i w_i \cdot data[j-i] \quad (3)$$

where j indicates a bit period and w_i is the strength of the slices connected to the i-th bit normalized to the full driver strength. From now on, we will refer to w_i as to the "weight" for the i-th tap. From Eq. (3) we see that the structure realizes an FIR filter, which implements a convolution between the bit stream (*data[j]*) and the tap vector w_i.

It should be noted that in order to obtain fine impedance tuning and compensate PVT (Process, Voltage and Temperature) variations, one does not strictly follow Eq. (1): the driver (divided in slices) is sized to obtain a resistance much larger than 50Ω; then many replicas of the structure are duplicated and the number of such replicas put in parallel is adjusted to match the 50Ω target [5].

III. CHOICE OF EQUALIZATION TAPS

In this paragraph, we will cover the basic steps that are required to determine the weights w_i for FFE.

The approach that we follow in our equalization process is the zero-forcing method [6]. Here the discussion will be at a tutorial level to bridge the gap between the theory in [6] and the actual hardware implementation (Figs. 2, 3). The goal is to minimize the distance between the desired response of the transmitter+channel (i.e. a signal without ISI) and that actually received; this is done via a Least Squares Minimization problem that reads

$$\min_{w_{ZFE}} \| z_{DES} - H_{CH} w_{ZFE} \|^2 \quad (4)$$

where w_{ZFE} is the weights vector $[w_0,w_1,\ldots,w_i,\ldots w_N]$ to be determined by the minimization problem and z_{DES} is the desired output response. In other words, if we consider a bit stream '10000', the transmitter will generate a sequence of pulses of height $[w_0,w_1,\ldots,w_i,\ldots w_N]$, each one stimulating the channel. We want to set w_i in such a way that the receiver samples the original sequence '10000' without ISI, meaning that z_{DES} =[1,0,0,…,0.]. H_{CH} is a matrix that rearranges the channel pulse response (h) in order to transform the convolution with the different pulses of height w_i into a matrix product. For example, if the channel pulse response is not null only for the first three samples and we decide to equalize with 5 post-cursor taps, then we have

$$H_{CH} = \begin{bmatrix} h_0 & 0 & 0 & 0 & 0 \\ h_1 & h_0 & 0 & 0 & 0 \\ h_2 & h_1 & h_0 & 0 & 0 \\ 0 & h_2 & h_1 & h_0 & 0 \\ 0 & 0 & h_2 & h_1 & h_0 \\ 0 & 0 & 0 & h_2 & h_1 \\ 0 & 0 & 0 & 0 & h_2 \end{bmatrix} \quad (5)$$

The solution of the minimization problem introduced by Eq. (4) requires extracting the channel pulse response. Therefore the simulation setup shown in Figure 4 is used. This setup is implemented in Ansys Electronic Desktop [7] and consists of a differential link with a pulse generator (with ideally steep rise and fall times) and an S-parameter block. This block changes from system to system and represents all the elements that compose the system after the transmitter. In this work we analyze two different systems, which we will refer as "BGA system" and "Leadframe system". The Ball Grid Array (BGA) system is composed by a via, approximately 5 mm long, which connects the transmitter output signal to the package output, a BGA package, a Printed Circuit Board (PCB) and a cable (10 cm long). At the receiver end, the

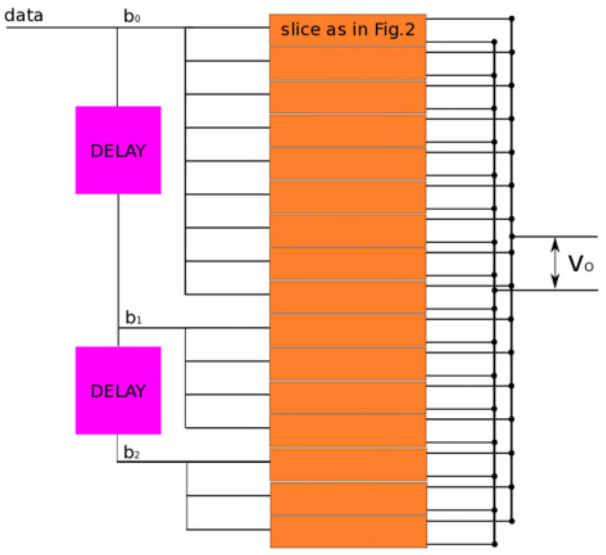


Figure 3. Example of FFE implementation using slices (with schematics as in Figure 2) put in parallel.

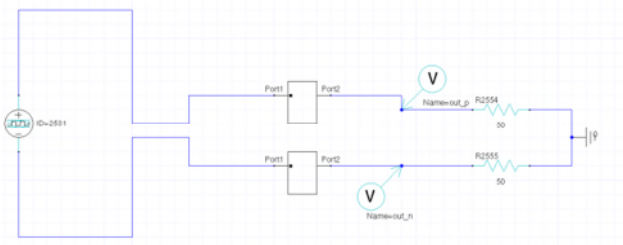


Figure 4. Setup implemented in Ansys Electronic Desktop in order to extract the channel pulse response

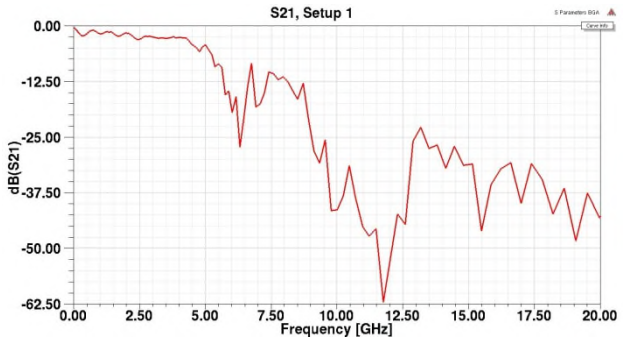


Figure 5. S_{21} of the BGA System

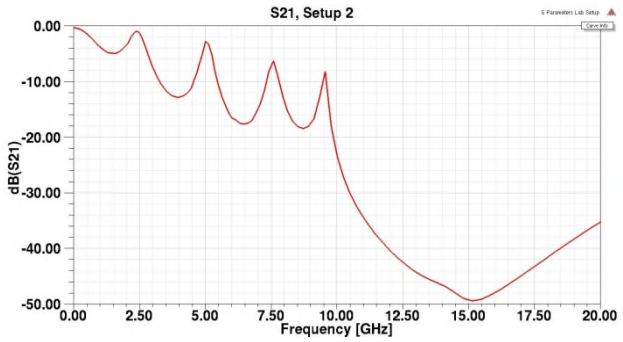


Figure 6. S_{21} of the Leadframe System

impedance is matched at 100Ω differential and the differential output is measured via voltage probes. The Leadframe System is similar to the BGA one, but instead of a BGA package it uses a leadframe one and the cable low-pass characteristic has been mimicked by lumped elements. Figs. 5 and 6 show the S_{21} of both systems.

The solution of the Eq. (4) reads

$$w_{ZFE} = (H_{CH}^{T} \cdot H_{CH})^{-1} \cdot H_{CH}^{T} \cdot z_{DES} \qquad (6)$$

which is the well-known solution for the LS objective function of Eq.4. Once the weights w_{ZFE} have been obtained, one can evaluate how close the overall transmitter+channel response fits the wanted z_{DES} by checking it against the product $H_{CH} \cdot w_{ZFE}$ (i.e. the overall response of the FFE+channel). This is shown in Figure 7, which compares the effect that equalizations with different number of taps has on the overall pulse response of the channel of Figure 6. In this particular case (consistent with the description above), only post-cursor taps were used, therefore the desired impulse channel response would be a 1 followed by a number of zeroes equal to the post-cursor taps used. So, for a 4-taps equalization (main and three post-cursors) the desired channel pulse response would be

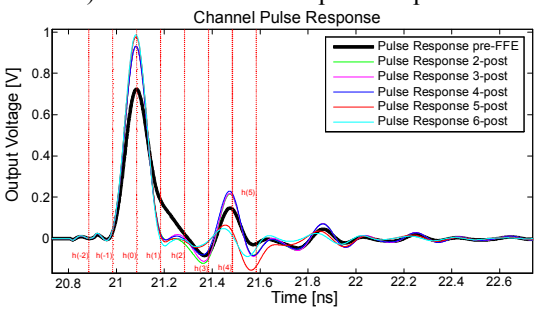


Figure 7. Response of channel of Figures 4-6 operating at 10Gbps along with responses after equalization for various numbers of taps

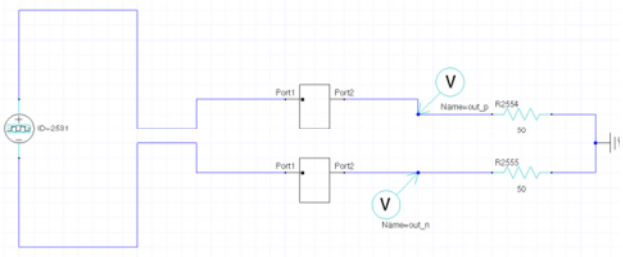


Figure 8. Setup implemented in Ansys Electronic Desktop in order to extract the obtain the eye diagram of the system

z_{DES} = [1,0,0,0]. If pre-cursor taps are used, then the z_{DES} elements are shifted to the right by a number of places equal to the number of pre-cursor taps.

IV. Example with Realistic Channels

Once the optimal weights have been found, one can then evaluate the effectiveness of FFE not only by looking at the channel pulse response (as in Figure 7), but also by simulating a structure equivalent to the system composed by transmitter, package, chip and channel and evaluating the improvements obtained in the eye diagram. The software used to this end is Ansys Electronic Desktop. Figure 8 illustrates a typical simulation setup: the only difference to the structure presented in Figure 4 is the use of a PRBS (Pseudo Random Bit Sequence) generator instead of a pulse generator. The PRBS generator offers the possibility to adapt its output based on equalization weights inserted by the user: in this case the vector w_{ZFE} obtained with Eq. (6). The low and the high voltage levels have been set to -450mV and 450 mV, respectively. Thermal noise typical of MOSFETs in the TX has not been modeled and will be one of next steps of our work.

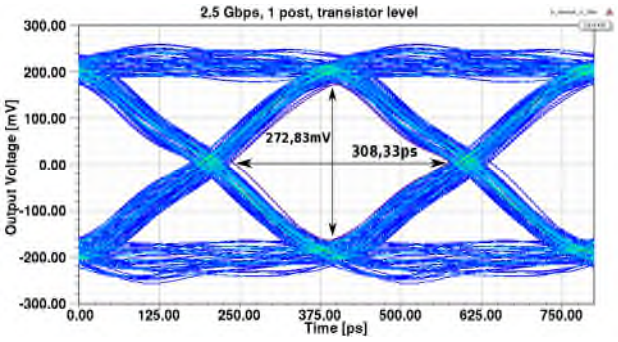


Figure 9. Eye diagram at 2.5 Gbps for BGA System obtained with a transsitor level simulation and one tap de-emphasis. The transmitter is described in [8].

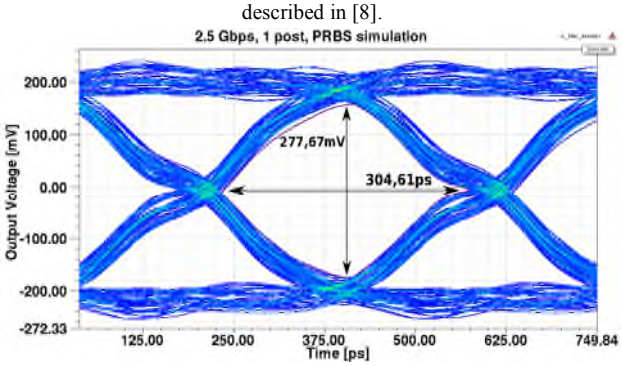


Figure 10. Eye diagram at 2.5 Gbps for BGA System obtained with a PRBS generator and one tap de-emphasis.

The channel is terminated with a 50 Ohm resistance.

In order to confirm the validity of our simplified approach, in Figs. 9 and 10 two eye diagrams of the same system (at 2.5 Gbps) are shown. One obtained with a PRBS generator and the other one with a transistor level model of the transmitter: the eye diagram parameters are very similar. Since the system level design is performed before actually designing the TX at transistor level, in the following, for the 5 Gbps and the 10 Gbps cases only the PRBS source is used. Figs. 11, 12, 13 and 14 report the eye diagrams in such cases with and without equalization. For the 5 Gbps situation, the improvement due to FFE is marginal, whereas to work at 10 Gbps with the Leadframe System, FFE is mandatory. Note that, at given V_{DD}, the inclusion of FFE lowers the high and low levels of the eye (e.g. 400mV vs. 300mV in Figs.9 and 10): this is because when FFE is implemented, some slices will be driven by bits having opposite sign with respect to the main one, and this implies that the driver is not working at full strength.

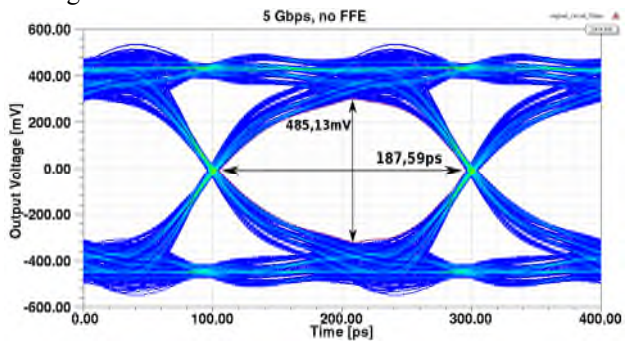


Figure 11. Eye diagram at 5 Gbps for Leadframe System without Feed-Forward Equalization.

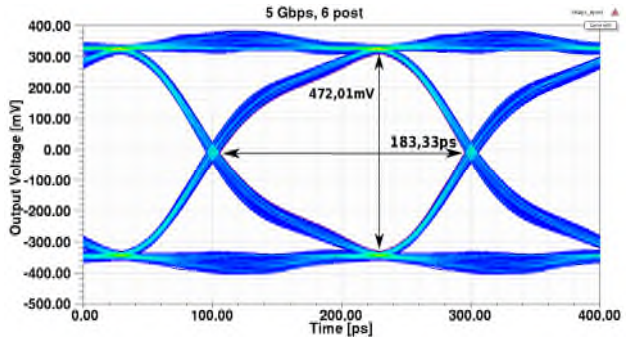


Figure 12. Eye diagram at 5 Gbps for Leadframe System with 6 post-cursor taps. The weights that generate this eye are w_0=0.8009, w_1=0.0403, w_2=-0.1018, w_3=0.0395, w_4=-0.0045, w_5=-0.013.

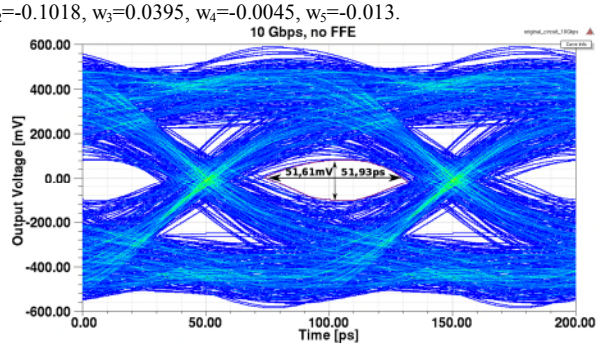


Figure 13. Eye diagram at 10 Gbps for Leadframe System without Feed-Forward Equalization.

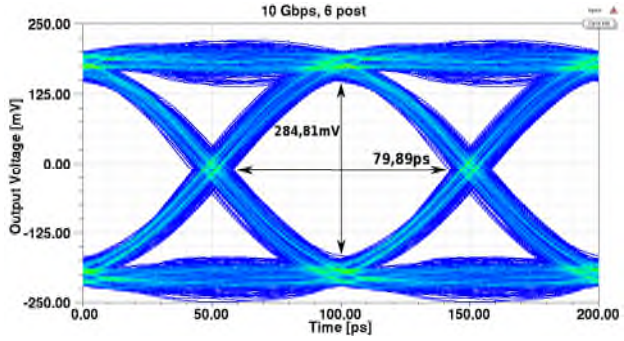


Figure 14. Eye diagram at 10 Gbps for Leadframe System with 6 post-cursor taps. The weights that generate this eye are w_0=0.577, w_1=-0.1526, w_2=0.0148, w_3= 0.0579, w_4=-0.1392 and w_5=0.0585.

V. EFFECT OF TAP QUANTIZATION

Eq. (6) provides optimum tap weights, but one must also think at a real world implementation, which obviously implies quantizing these obtained weights since each bit will be connected to a finite number of slices. This problem is peculiar to the voltage-mode transmitter divided in slices. In fact previous works already introduced equalization implemented with sliced drivers, but mainly in current mode logic [9]-[10], which makes equalization easier to implement with high granularity.

For this reason, we analyzed the effect of quantization with two different granularities, 8 and 16 levels (i.e. M=8 or M=16 as in Figure 3). These two different granularities

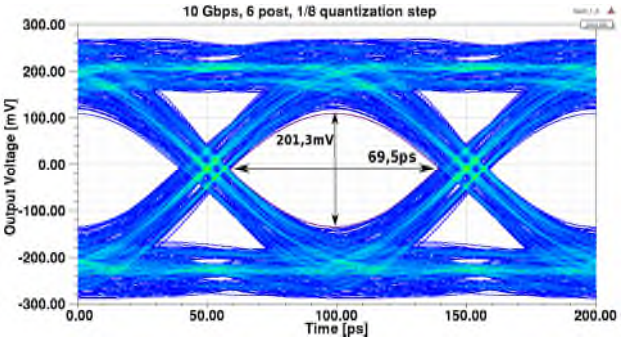


Figure 15. Eye diagram at 10 Gbps for Leadframe System with 6-post cursor taps and quantization step of 1/8. The weigths are w_0=0.625, w_1=-0.125, w_2=0, w_3=0, w_4=-0.125 and w_5=0.125, which correspond to 5 slices connected to b_0 while $\overline{b_1}$, $\overline{b_4}$ and b_5 require one slice each.

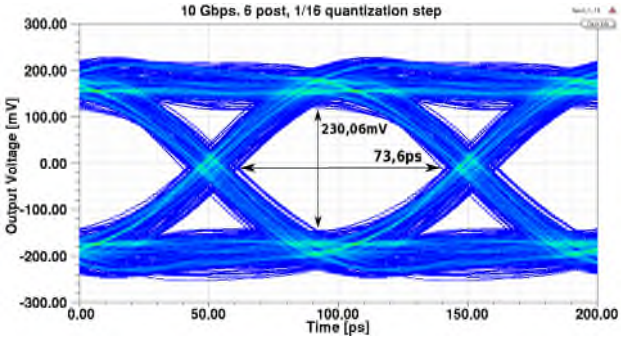


Figure 16. Eye diagram at 10 Gbps for Leadframe System with 6-post cursor taps and quantization step of 1/16. The weights are w_0=0.5625, w_1=-0.1875, w_2=0, w_3=0.0625, w_4=-0.125 and w_5=0.0625, which correspond to 9 slices connected to b_0, three connected to $\overline{b_1}$, one slice connected to b_3, one to b_5 and two to $\overline{b_4}$.

offer quantization steps of 0.0625 and 0.125 respectively.

Figs. 15 and 16 show the effect of quantization on the operation at 10 Gbps of Leadframe System when equalized with 6 post-cursor, which without quantization has already been shown in Figure 14. With 16 slices we obtain eye parameters very close to what is obtained from Eq. (6), whereas with 8 slices there is a degradation of the eye.

VI. CONCLUSIONS

We have reported on the design of a high speed transmitter at 10 Gbps, focusing on the system level planning of the feed-forward-equalization. It has been shown that for a realistic channel typical of automotive ECU applications, communication at 10 Gbps requires 6 taps. The driver has to be partitioned in at least 16 slices in order to have the requested granularity of the taps for the FIR filter. If the transmission speed gets even higher, then a higher number of taps is needed and the effect of quantization becomes more and more relevant. The eye parameters with and without FFE and including different granularity in the tap quantization are summarized in Figs. 17-20. These will be checked against experimental data once the test chip will be available. Transistor level design with an advanced CMOS technology is ongoing and the results will be published later on. We include in these figures also a 15 Gbps situation that requires 9 taps for equalization when considering the Leadframe System (7 for the BGA System). In Figs. 17-20 we see that FFE improves the eye height and width, although part of the improvement is lost if a too coarse granularity is used for the w_i.

ACKNOWLEDGMENT

The authors would like to thank Prof. L. Selmi (University of Udine) for support and for many helpful discussions.

REFERENCES

[1] Carusone, Tony Chan. "Introduction to Digital I/O: Constraining I/O Power Consumption in High-Performance Systems." *Solid-State Circuits Magazine, IEEE*, vol. 7, no. 4 (2015): 14-22.

[2] Chang, Ken, Geoff Zhang, and Christopher Borrelli. "Evolution of Wireline Transceiver Standards: Various, Most-Used Standards for the Bandwidth Demand." *Solid-State Circuits Magazine, IEEE*, vol. 7, no. 4 (2015): 47-52.

[3] Bulzacchelli, John F. "Equalization for Electrical Links: Current Design Techniques and Future Directions." *Solid-State Circuits Magazine, IEEE,* vol. 7, no. 4 (2015): 23-31.

[4] Razavi, Behzad. "Historical Trends in Wireline Communications: 60X Improvement in Speed in 20 Years." *Solid-State Circuits Magazine, IEEE*, vol. 7, no. 4 (2015): 42-46.

[5] Kossel, Marcel, et al. "A T-coil-enhanced 8.5 Gb/s high-swing SST transmitter in 65 nm bulk CMOS with 16 dB return loss over 10 GHz bandwidth." *Solid-State Circuits, IEEE Journal of,* vol. 43. no. 12 (2008): 2905-2920.

[6] Proakis, John G. Intersymbol interference in digital communication systems. John Wiley & Sons, Inc., 2001.

[7] Ansys Inc. PDF Documentation for Release 15.0

[8] Cossettini, A., et al. "Design, characterization and signal integrity analysis of a 2.5 Gb/s High-Speed Serial Interface for automotive applications overarching the chip/PCB wall." *Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI), 2015 IEEE 1st International Forum on* IEEE, 2015.

[9] Bulzacchelli, J. F., et al. "4.1 A 10Gb/s 5-Tap-DFE/4-Tap-FFE Transceiver in 90nm CMOS." *Solid-State Circuits, IEEE Journal of,* vol. 41. no. 12 (2006): 2885-2900.

[10] Beukema, Troy, et al. "A 6.4-Gb/s CMOS SerDes core with feed-forward and decision-feedback equalization." *Solid-State Circuits, IEEE Journal of* 40.12 (2005): 2633-2645.

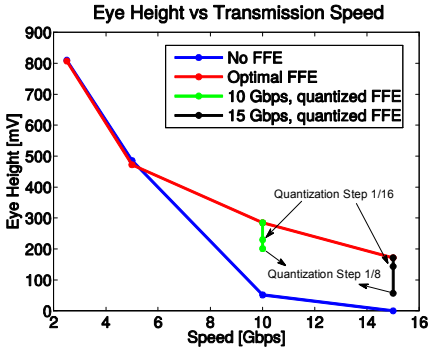


Figure 17. Eye height versus transmission speed for Leadframe System when not equalized, optimally equalized and when quantization is applied to w_i.

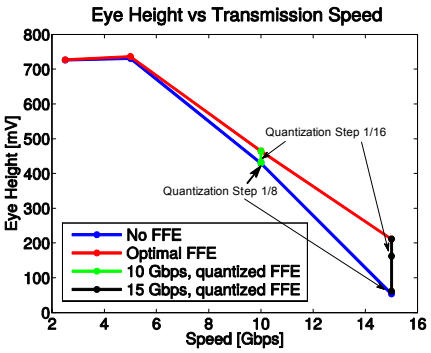


Figure 19. Eye height versus transmission speed for BGA System when not equalized, optimally equalized and when quantization is applied to w_i.

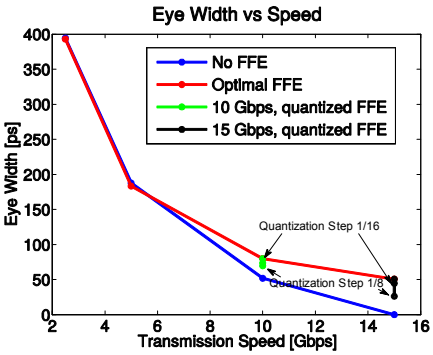


Figure 18. Eye width versus transmission speed for Leadframe System when not equalized, optimally equalized and when quantization is applied to w_i.

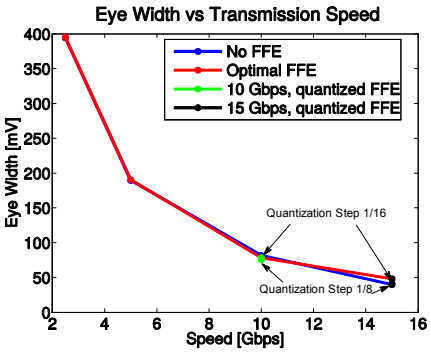


Figure 20. Eye width versus transmission speed for BGA System when not equalized, optimally equalized and when quantization is applied to w_i.

Application of the calculation-experimental method in the design of microwave filters

A.S. Geraskin*, A.N. Savin*, I.A. Nakrap**, V.P. Meshchanov***
* National Research Saratov State University/ Faculty of Computer Science and Information Technologies, Saratov, Russian Federation
** National Research Saratov State University/ Faculty of Physics, Saratov, Russian Federation
*** Yuri Gagarin State Technical University of Saratov/ Institute of electronic engineering and mechanical engineering, Saratov, Russian Federation
Gerascinas @mail.ru, savinan@info.sgu.ru, nakrapia@info.sgu.ru, nika373@bk.ru

The estimation of the applicability of calculation-experimental method to optimize the design of microstrip microwave filters on the example of the band-pass filter on the half-wave resonators with Chebyshev characteristic. This method is based on an iterative process of the correction of the result designed device parameters synthesis in his experimental output characteristics.

It is shown that the method allows to consider the influence of various factors relating to the manufacturing process and features of the materials used. However, an additive method of accounting in the target function of deviations from given output characteristics does not provide a rapid convergence process at large deviations.

In order to increase the efficiency of calculation - experimental optimization method is offered its modification, consisting in additional correction parameters of materials and mathematical models of components of the experimental output characteristics of the device.

Accounting irregularities of the relative dielectric permittivity and the thickness of the microstrip filter components in its design with the help of a modified method has allowed for a single iteration to reduce the unevenness of the reflection coefficient module of the filter with 7.1 dB to 1.7 dB.

I. INTRODUCTION

At the present stage of science development and technology are widely used various microstrip microwave devices. The development of such devices assumes the numerical or physical modeling of the processes occurring in them, and following optimization of the device in order to achieve the required values of its output parameters [1].

Design of microwave devices, as a rule, is carried out with the use of specialized computer-aided design [2]. However, the real parameters of the microstrip microwave devices can be significantly different from designed parameters. This can be due to various factors: inhomogeneity of parameters of the materials used, deviations microstripes size and quality of their surface, and inaccurate mathematical models of the device components.

In this connection arises accounting task of indicated factors during development. In [3] propose calculation-experimental method of parametric optimization (CEMO), allowing to solve the given task based on the use of experimental output characteristics of microwave devices.

Purpose of the given work is to study the possibility of applying CEMO at designing of microstrip microwave filters.

II. THE ALGORITHM OF THE CALCULATION-EXPERIMENTAL METHOD BASED ON CHEBYSHEV APPROXIMATION OF THE OUTPUT CHARACTERISTICS

In the process of designing microwave devices, as a rule, solves the problem of parametric synthesis of the vector v variable parameters of the device in order to minimize the deviation of the output characteristic of the $f(v,\theta)$ from the given characteristics of $F(\theta)$ in the frequency domain $\theta \in [\theta_1, \theta_2]$:

$$\min_{v} \max_{\theta\in[\theta_1,\theta_2]} |f(v,\theta) - F(\theta)| \quad (1)$$

The implementation process is carried out usually with the use of specialized tools computer-aided design (for example, NI AWR Microwave Office [2]) using various approximate mathematical models of components for microwave devices design. Accordingly, the resulting output characteristic $f(v^*,\theta)$ may differ from the experimental characteristics $f^e(v^*,\theta)$ of a real device (v^* is the optimal vector of parameters in the device synthesized in the solution of the task (1)). Proposed in [3] CEMO is based on correction results of the synthesis using the experimental device characteristics. Thus the physical implementation of the device is considered as an intermediate step of the iterative process of its development.

You can select the following steps of the algorithm of this method [3, 4]:

1. Synthesizing optimal vector v_0^* of the microwave device parameters for a given frequency domain $\theta \in [\theta_1, \theta_2]$ the characteristic $F(\theta)$ and the calculation of the output characteristics $f_0(v_0^*, \theta)$ (i.e. the solution of task (1)) with the use of computer-aided design.
2. Making of microwave devices designed for the model and measuring the real characteristics $f_0^e(v_0^*, \theta)$ at the m points of the frequency range $[\theta_1, \theta_2]$, corresponding to the output characteristics $f_0(v_0^*, \theta)$ of the Chebyshev approximation. The use of Chebyshev approximation allows the use in the synthesis process of the vector v_0^* the minimum number of points for a given accuracy of the approximation and correspondingly the minimum number of experimental data.
3. The construction new target characteristics $F_1(\theta)$, takes into account the difference $\Delta(\theta) = f_0(v_0^*, \theta) - f_0^e(v_0^*, \theta)$ between the calculated $f_0(v_0^*, \theta)$ and experimental $f_0^e(v_0^*, \theta)$ output characteristics is carried out by the specular reflection amendment $\Delta(\theta)$ relative to the origin given characteristics $F(\theta)$:

$$F_1(\theta) = F(\theta) + \Delta(\theta) \quad (2)$$

   The use of output characteristics $f_0^e(v_0^*, \theta)$ for a given correction characteristics $F(\theta)$ allows to take into account inaccuracies of the mathematical model of the output characteristics $f_0(v_0^*, \theta)$ and the features of the materials used and the manufacturing process of the microwave devices.
4. Synthesis of new optimal vector v_1^* parameters of microwave devices by new a given characteristic $F_1(\theta)$ and the calculation of the output characteristics $f_1(v_1^*, \theta)$ (i.e. solution of task (1) with a modified target function). Thus as the initial approximation used the earlier resulting vector v_0^*.
5. Making the new device with the parameter vector v_1^*, and to measure its output characteristics $f_1^e(v_1^*, \theta)$.
6. If the obtained output characteristic $f_1^e(v_1^*, \theta)$ the required accuracy coincides with a given output characteristic $F(\theta)$, the process ends with the development of the device, otherwise returns to step 3 with the use of (2) a correction value $\Delta(\theta) = f_0(v_0^*, \theta) - f_1^e(v_1^*, \theta)$ in the formation of new characteristics and a given vector v_1^* as a first approximation for synthesizing the device settings at step 4.

In [3] it is shown that in most cases only one of the calculation-experimental cycle for the satisfactory coincidence of the functions $f_1^e(v_1^*, \theta)$ with a given $F(\theta)$.

In order to assess the applicability of this method in the designing microstrip microwave filters was carried out to develop a band pass filter fourth order on half-wave resonators by the given algorithm.

III. Development of Microstrip Microwave Filter Using Computational Experimental Method

Developed microstrip microwave filter should have the following characteristics:

- type of filter – band-pass;
- center frequency $\theta_0 = 2.35$ GHz;
- the boundary frequencies of passband: $\theta_1 = 2.2$ GHz, $\theta_2 = 2.5$ GHz;
- the non-uniformity of modulus of transmission coefficient $|S_{21}|$ of the filter in a passband of not more than 0.1 dB;
- the module of the reflection coefficient $|S_{11}|$ filter in the passband must have the minimum possible deviation from the theoretically calculated and the minimum non-uniformity, measured by the maximum in bandwidth;
- the filter should be connected to microstrip lines with a characteristic impedance $\rho = 50$ Om;
- the required transfer function must be the Chebyshev polynomial of the fourth order;
- mathematical models of the components of the filter must have the ultimate shape;
- the substrate material of the filter – two-sided coated fibreglass FR4-2 c parameters: $H = 1.5\,mm$, $T = 0.035\,mm$, $\sigma = 5.88 \cdot 10^7\ S/m$, $\varepsilon_r = 4.4$.

Filter topology on half-wave resonators, working principle of which is discussed in detail in [5], is shown in Fig. 1.

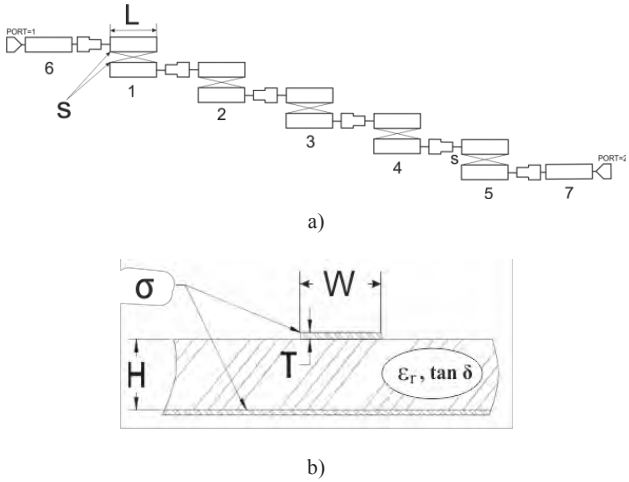


Figure 1. Topology of the developed filter

In accordance with steps 1 and 2 CEMO in the program NI AWR Microwave Office (Trial version) was synthesized microstrip band pass filter fourth order on half-wave resonators with Chebyshev transfer function, satisfying the above requirements and made its prototype. The vector of optimal parameters v_0^* of the filter are given in section a) Table. 1. Calculated and experimental S –parameters of the filter shown in Fig. 2.

A sufficiently large difference between the experimental characteristics of the filter calculated from "Fig. 2" may be related to characteristics of the technology used in the manufacture of filter inhomogeneity relative permittivity $\varepsilon_r$of the substrate and the thickness T of microstrips "Fig. 1b", the quality of the surface, and inaccurate mathematical model of the filter, taking into account in the calculation of T -wave.

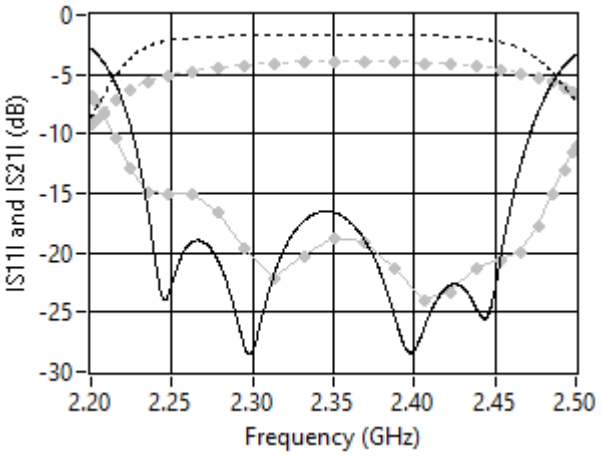


Figure 2. Calculated (——, - - - -) and experimental (—♦—, - -♦- -) S -parameters of the filter parameters v_0^* synthesized in NI AWR according to specified requirements

Further optimization of the filter design for the purpose of reduction of reflection coefficient module $|S_{11}|$ layout of the experimental to the calculated mean was carried out in accordance with steps 3 – 5 CEMO.

In accordance with the requirements for the formation of a new specified $F_1(\theta)$ in the expression (2) as the output $f_0(v_0^*, \theta)$ and set $F(\theta)$ characteristics of the considered module of the reflection coefficient $|S_{11}|$ calculated filter, and the output characteristics of

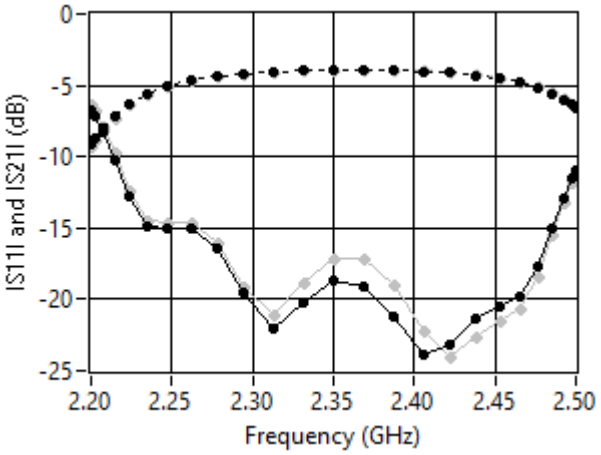


Figure 3. Experimental S -parameters of the filter with the parameters of v_0^* (—•—, - -•- -) and v_1^* (—♦—, - -♦- -), obtained CEMO

experimental $f_0^e(v_0^*, \theta)$ – $|S_{11}|$ of experimental filter, obtained in steps 1 and 2. Thus the calculated and experimental $|S_{11}|$ was set to $m = 25$ points of the frequency range $[\theta_1, \theta_2]$ used Chebyshev approximation $|S_{11}|$ "Fig. 2".

On the basis of the detailed analysis of the influence of the dimensions of the filter to half-wave resonators in its output characteristics, is given in [5], were defined varying parameters in the vector $v: L$ - length of microstrip all components of the filter and S – gaps in segments of the associated lines "Fig. 1".

New vector v_1^*, the optimal filter parameters, obtained by synthesis on a given new feature $F_1(\theta)$ given in item b) of the table. 1. Output characteristics of the layout of the filters with the parameters defined by the vectors v_0^* v_1^*, is shown in Fig. 3.

TABLE I. PARAMETERS OF THE SYNTHESIZED FILTERS

№ component of filter	H, mm	W, mm	S, mm	L, mm	T, mm	ε_r
a) The filter parameters v_0^*, synthesized by a given characteristic $F(\theta) = (f_0(v_0^*, \theta)$						
1, 5	1.5	1.9	0.326	17.24	0.035	4.4
2, 4	1.5	2.7	1.383	16.73	0.035	4.4
3	1.5	2.7	1.883	17.08	0.035	4.4
6, 7	1.5	2.856	-	8.606	0.035	4.4
b) The filter parameters v_1^*, synthesized by a given characteristic $F_1(\theta) = f_0(v_0^*, \theta) + (f_0(v_0^*, \theta) - f_0^e(v_0^*, \theta))$						
1, 5	1.5	1.9	0.326	17.25	0.035	4.4
2, 4	1.5	2.7	1.386	16.72	0.035	4.4
3	1.5	2.7	1.914	17.09	0.035	4.4
6, 7	1.5	2.856	-	8.83	0.035	4.4
c) The filter parameters $v_{0,c}^*$, synthesized by a given characteristic $F(\theta) = f_0^e(v_0^*, \theta)$ in view of inhomogeneity ε_r and T						
1, 5	1.5	1.9	0.326	17.24	0.039	4.283
2, 4	1.5	2.7	1.383	16.73	0.033	4.507
3	1.5	2.7	1.883	17.08	0.040	4.348
6, 7	1.5	2.856	-	8.606	0.035	4.727
d) The filter parameters $v_{1,c}^*$, synthesized by a given characteristic $F(\theta) = f_0(v_0^*, \theta)$ in view of inhomogeneity ε_rи T						
1, 5	1.5	1.9	0.345	17.351	0.039	4.283
2, 4	1.5	2.7	1.371	16.605	0.033	4.507
3	1.5	2.7	1.880	17.113	0.040	4.348
6, 7	1.5	2.856	-	6.900	0.035	4.727

The difference between the outer lateral maximums characteristic filter with parameter vector v_0^* is 5.7 dB. The difference for the characteristic with the vector parameters v_1^* is equal 7.1 dB "Fig. 3". As you can see, using one iteration, CEMO, to improve the characteristics of the filter failed. This may be explained by the wrong choice at step 4 variable filter parameters, do not provide in this case a consideration of the factors discussed above affecting the output characteristic.

For the purpose of solving this task was proposed to modify steps 3, 4 CEMO by using prior synthesis to adjust the parameters of the materials used and, if necessary, the study of mathematical models of components of the device according to the experiment proposed in [6] for the synthesis of waveguide microwave systems [7].

This idea has been used to optimize the filter developed by adjusting the technological parameters of

the experimental layout: the relative dielectric constant ε_r of the substrate and the thickness T of microstrips.

Modified CEMO step 3 (step 3') consists of the solutions of problem (1) synthesis of the optimal correcting vector $v_{0,c}^*$ in which the varied parameters are only ε_r and T elements, and the remainder obtained in step 1 correspond to the vector v_0^*. At the same time as the given characteristics of $F(\theta)$ is used characteristic output experimental $f_0^e(v_0^*, \theta) - |S_{11}|$, measured in step 2.

Modified CEMO step 4 (step 4') solves the problem (1) synthesis of new optimal vector $v_{1,c}^*$, where as the original CEMO, varied parameters are L – length of microstrip all components of the filter and S – gaps in segments of the associated lines [5], but ε_r and T elements obtained in step 3', correspond to the vector $v_{0,c}^*$. As specified $F(\theta)$ is used as output estimated characteristics $f_0(v_0^*, \theta) - |S_{11}|$, calculated in step 1.

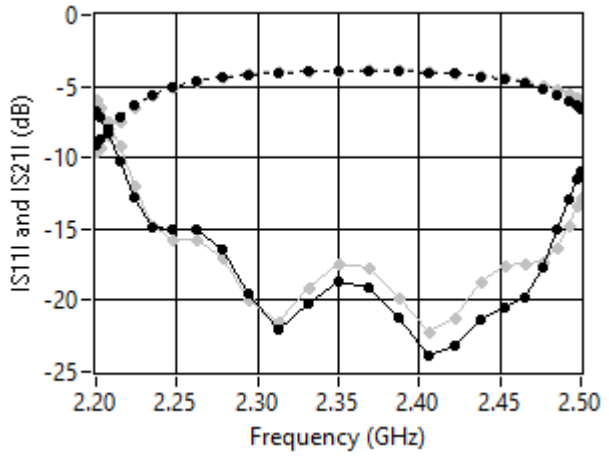


Figure 4. Experimental S -parameters of the filter with the parameters of v_0^* (—•—, - -•- -) and $v_{1,c}^*$ (—♦—, - -♦- -), the obtained modified CEMO

Further in accordance with steps 5 and 6 CEMO the manufacture of the new layout of the filter with parameter vector $v_{1,c}^*$ and to measure its output characteristics $f_{1,c}^e(v_{1,c}^*, \theta)$.

Obtained in steps 3', 4' corrective $v_{0,c}^*$ and new $v_{1,c}^*$ vector of optimal parameters of filter are given in item c) and d) of the table. 1. Output characteristics of the layout of the filters with the parameters defined by the vectors v_0^* and $v_{1,c}^*$, is shown in Fig. 4.

TABLE II. DEVIATIONS OF SUBSTRATE PARAMETERS FROM THE NOMINAL

№ component of filter	ε_r, %	T, %
1, 5	-2.7	11.4
2, 4	2.4	-5.7
3	-1.2	14.3
6, 7	7.3	0

The difference between the outer lateral maximums characteristic filter with parameter vector v_0^* is 5.7 dB. The difference for the characteristic with the vector parameters v_1^* is equal 1.7 dB "Fig. 4". As can be seen, the use of a mathematical model of the filter to adjust the parameters of the substrate according to the experimental data allowed us to align $|S_{11}|$ filter. The relative deviations of parameters from nominal substrate for filter components are small "Table. 2" and correspond to the tolerances on the characteristics of the material used. Thus, the proposed modification based on the correct mathematical model of the filter on its real parameters, greatly improves the efficiency of CEMO.

IV. CONLUSION

For example, the development of bandpass microstrip microwave fourth-order filter on the half-wave resonators with Chebyshev characteristic evaluated the effectiveness calculation-experimental optimization method, founded on an iterative process of correction of the results of the developed filter synthesis using his experimental output characteristics.

It is shown that the method allows to take into account the influence of various factors relating to process of manufacturing and features of the materials used. However, additive methods of accounting in the target function of deviations from given output characteristics do not provide a rapid convergence process at large deviations. To reduce the unevenness of the reflection coefficient module of the filter for one iteration failed.

With the aim of increasing the effectiveness of calculation-experimental method of optimization is offered his modification, consisting in an additional adjustment of materials parameters and mathematical models of components of experimental output characteristics of the device. The proposed modification of the method can be used for a wide class of microstrip devices.

Accounting irregularities of the relative dielectric permittivity and the thickness of the microstrip filter components in its design with the help of a modified method has allowed for a single iteration to reduce the unevenness of the reflection coefficient module of the filter with 7.1 dB to 1.7 dB.

ACKNOWLEDGMENT

Work is performed with financial support of Ministry of Education and Science of Russian Federation within the project part of state assignment in the field of scientific activity #3.1155.2014/K.

REFERENCES

[1] T. C. Edwards, M. B. Steer, "Foundations for Microstrip Circuit Design, Fourth Edition", Hardcover - Wiley & Sons Ltd, 2016.

[2] http://www.awrcorp.com/products/microwave-office.

[3] B. M. Kac, V. P. Meshchanov, A. L. Feldshtein, "Optimalnyj sintez ustrojstv svch s T volnami./ Pod red. V. P. Meshchanov, M.:Radio i svyaz, 1984. 288 s.

[4] A. M. Bogdanov, M. V. Davidovich, B. M. Kac i dr. “Sintez sverhshirokopolosnyh mikrovolnovyh struktur./ Pod red. A. P. Krenickogo i V. P. Meshchanova, M.: Radio i svyaz, 2005, 514 s.

[5] M. I. Nikitina, “Sistema proektirovaniya mikropoloskovyh polosno-propuskayushchih filtrov”, dis. kand. tekh. nauk, Krasnoyarsk, Krasnoyarskij gos. tech.un-t, 1998, 151 s.

[6] A. N. Savin, “Issledovanie ehlektrodinamicheskih harakteristik struktur vakuumnoj ehlektroniki i magnitoehlektroniki svch na osnove regressionnyh modelej”, dis. kand. fiz. – mat. nauk, Saratov, Saratovskij gos. un-t, 2003, 184 s.

[7] I. A. Nakrap, A. N. Savin, and Yu. P. Sharaevskii, “Modeling of Wideband Slow-Wave Structures of Coupled-Cavities-Chain Type by the Impedance Design of Experiments”, Journal of Communications Technology and Electronics, 2006, Vol. 51, No. 3, pp. 316–323.

Minimax Design of Multiplierless Sharpened CIC Filters Based on Interval Analysis

Goran Molnar[1], Aljosa Dudarin[1], and Mladen Vucic[2]

[1]Ericsson Nikola Tesla d. d.
Krapinska 45, 10000 Zagreb, Croatia

[2]University of Zagreb, Faculty of Electrical Engineering and Computing
Department of Electronic Systems and Information Processing
Unska 3, 10000 Zagreb, Croatia
E-mails: goran.molnar@ericsson.com, aljosa.dudarin@ericsson.com, mladen.vucic@fer.hr

***Abstract*—Polynomial sharpening of amplitude responses of cascaded-integrator-comb (CIC) filters is often used to improve folding-band attenuations. The design of sharpened CIC (SCIC) filters is based on searching the polynomial coefficients ensuring required magnitude response. Several design methods have been developed, resulting in real, integer, and sum-of-powers-of-two (SPT) coefficients. The latter are preferable since they result in multiplierless structures. In this paper, we present a method for the design of SCIC filters with SPT polynomial coefficients by using the minimax error criterion in folding bands. To obtain the coefficients, we use a global optimization technique based on the interval analysis. The features of the presented method are illustrated with the design of wideband SCIC filters.**

I. Introduction

Cascaded-integrator-comb (CIC) filter [1] is common building block in digital-down converters. However, in processing of wideband signals, the CIC filter is often incapable of meeting the requirement for high folding-band attenuations. To improve the CIC-filter folding-band response, various structures have been developed. An efficient structure arises from the polynomial sharpening of CIC response [2]. This structure implements so called sharpened CIC (SCIC) filter.

The design of sharpened CIC filters is based on searching the polynomial coefficients ensuring required magnitude response. Several design methods have been developed, resulting in real, integer, and sum-of-powers-of-two (SPT) coefficients. The latter are preferable since they result in multiplierless structures.

Well-established sharpening method was developed by Kaiser and Hamming [3]. It gives integer coefficients using analytic expression. The incorporation of the Kaiser-Hamming polynomial in SCIC filter was first presented in [4]. This structure is further improved in [5]−[8]. Recently, the Chebyshev polynomials have been used in the design of SCIC filters with very high folding-band attenuations [9]. Furthermore, closed-form methods for the design of SCIC filters using the weighted least-squares [10] and minimax [11] error criterion in the passband and folding bands have been proposed. These methods provide SCIC filters ensuring small passband deviations and rather high folding-band attenuations. However, the coefficients obtained take real values, what results in the structure employing multipliers. In [12], the particle swarm optimization has been used to calculate SPT polynomial coefficients in order to achieve a given passband deviation. In [13], partially sharpened CIC filters have been developed, employing the polynomials in Bernstein's form [14], [15]. The proposed filters are multiplierless, but they only support power-of-two decimation factors. In [16] and [17], the sharpening has been combined with passband compensation.

In this paper, we present a method for the design of multiplierless SCIC filters by using the minimax error criterion in folding bands only. The design is performed over the SPT coefficient space. To obtain the coefficients, we use a global optimization technique based on the interval analysis [18]. The presented method brings significant reduction in the filter complexity compared to the Chebyshev SCIC filters [9].

The paper is organized as follows. Section II briefly describes the sharpening of CIC filters. The method for the minimax design of multiplierless SCIC filters is presented in Section III. Section IV contains examples illustrating the features of the proposed filters.

II. Sharpened CIC Filter

The amplitude response of the CIC filter of the Nth order is given by [1]

$$H_{CIC}(\omega)=\left[\frac{1}{R}\frac{\sin\left(\frac{\omega R}{2}\right)}{\sin\left(\frac{\omega}{2}\right)}\right]^{N} \tag{1}$$

where R denotes the decimation factor. The sharpening polynomial of the Mth order is given by

This paper was supported in part by Ericsson Nikola Tesla d.d. and University of Zagreb, Faculty of Electrical Engineering and Computing under the project Improvement of Competences for LTE Radio Access Equipment (ILTERA), and in part by Croatian Science Foundation under the project Beyond Nyquist Limit, grant no. IP-2014-09-2625.

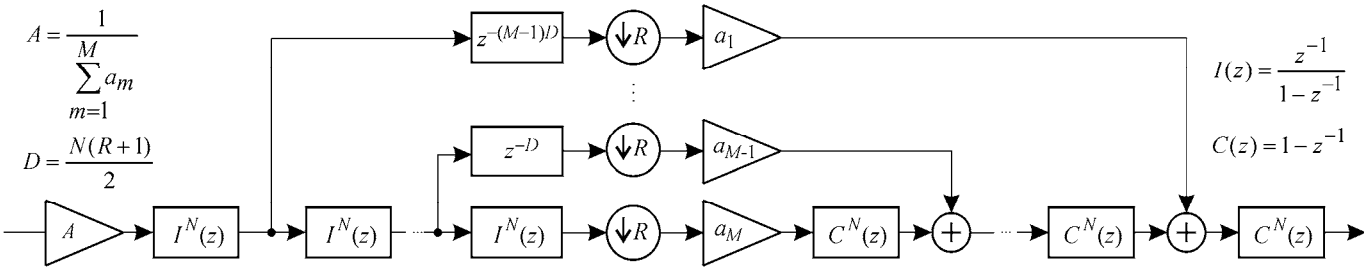


Figure 1. Structure of sharpened CIC decimation filter.

$$f(x) = \sum_{m=0}^{M} a_m x^m \tag{2}$$

By substituting $x = H_{CIC}(\omega)$ into (2), we obtain the amplitude response of the sharpened CIC filter. It is given by

$$H_{SCIC}(\omega) = \sum_{m=0}^{M} a_m H_{CIC}^m(\omega) \tag{3}$$

It is well known that the amplitude response of a CIC filter has zeros placed at the central frequencies of the folding bands. This property is ensured if $a_0 = 0$. Therefore, we deal with the amplitude response

$$H_{SCIC}(\omega) = \sum_{m=1}^{M} a_m H_{CIC}^m(\omega) \tag{4}$$

The structure of SCIC decimation filter is shown in Figure 1 [2]. It is clear from the figure that one multiplier is necessary for each non-zero polynomial coefficient. It makes the structure complex, especially for polynomials of high orders. However, by using the SPT coefficients, the multipliers are replaced by adders, resulting in an efficient structure. It should be noted that the structure is suitable if the delay elements introduce integer delays. It is achieved if $N(R+1)$ is an even number. The structure can be further simplified such that elements with delays greater than R are split into two blocks by applying the noble identity. Consequently, one element operates at high, whereas the other one operates at low sampling rate.

III. Minimax Design of Multiplierless Sharpened CIC Filters

A. Minimax Approximation

Two approaches to the design of sharpened CIC filters have been considered in literature. The first approach simultaneously sharpens the passband and the folding-band responses. The sharpening polynomial is obtained using the maximally flat [4], least squares [10], or minimax [11] approximation. The second approach includes sharpening only within the folding bands. Such an approach includes the design based on the Chebyshev polynomials [9]. To obtain high folding-band attenuations, the second approach is preferable.

Here, we perform the minimax design of SCIC filters within the folding bands. We start the design with weighted absolute error

$$\varepsilon(\mathbf{a}) = \max_{\omega \in \Omega} w(\omega) \left| H_{SCIC}(\omega, \mathbf{a}) - H_d(\omega) \right| \tag{5}$$

where $\mathbf{a}$ is the vector of polynomial coefficients given by

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_M \end{bmatrix}^T \tag{6}$$

$w(\omega)$ is a positive weighting function, and $H_d(\omega)$ is the desired amplitude response defined within the band Ω. The error function in (5) should take into account *only* the folding bands. They are given by

$$\Omega = \begin{cases} \frac{2n\pi}{R} - \omega_p \le \omega \le \frac{2n\pi}{R} + \omega_p & ; \quad n = 1, \ldots, \frac{R}{2} - 1 \\ \pi - \omega_p \le \omega \le \pi \end{cases} \tag{7}$$

for an even R, and by

$$\Omega = \frac{2n\pi}{R} - \omega_p \le \omega \le \frac{2n\pi}{R} + \omega_p \quad ; \quad n = 1, \ldots, \frac{R-1}{2} \tag{8}$$

for an odd R, where ω_p is the edge of filter's passband.

In folding bands, the desired response is zero. In addition, we assume the unity weighting function. By substituting $H_d(\omega) = 0$ and $w(\omega) = 1$ into error function in (5), we arrive at

$$\varepsilon(\mathbf{a}) = \max_{\omega \in \Omega} \left| H_{SCIC}(\omega, \mathbf{a}) \right| \tag{9}$$

Note that function in (9) does not take into account the passband gain of the filter. Therefore, we normalized it to constant passband gain at one frequency. Here, we choose the unity gain at $\omega = 0$. The error function thus takes the form

$$\varepsilon(\mathbf{a}) = \max_{\omega \in \Omega} \left| \frac{H_{SCIC}(\omega, \mathbf{a})}{H_{SCIC}(0, \mathbf{a})} \right| \tag{10}$$

The expression for $H_{SCIC}(0,\mathbf{a})$ is easily obtained by substituting $\omega = 0$ into (4), resulting in

$$H_{SCIC}(0, \mathbf{a}) = \sum_{m=1}^{M} a_m \tag{11}$$

Finally, by substituting (11) into (10), we arrive at

$$\varepsilon(\mathbf{a}) = \max_{\omega\in\Omega} \left| \frac{\sum_{m=1}^{M} a_m H_{CIC}^m(\omega)}{\sum_{m=1}^{M} a_m} \right| \tag{12}$$

In the design, we calculate $\varepsilon(\mathbf{a})$ using the CIC response evaluated on uniformly spaced frequency grid $Q = \{\omega_k; k=0, \ldots, K-1\}$ defined within the folding bands, Ω. Hence, the objective function is obtained as

$$\varepsilon(\mathbf{a}) = \max_{\omega_k\in Q} \left| \frac{\sum_{m=1}^{M} a_m H_{CIC}^m(\omega_k)}{\sum_{m=1}^{M} a_m} \right| \tag{13}$$

B. Problem Formulation

Our goal is to find the optimum SPT polynomial coefficients of the sharpened CIC filter in the minimax sense. Such a design is described by the optimization problem

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \left[\varepsilon(\mathbf{a})\right] \quad \text{subject to: } \mathbf{a} \text{ is SPT representable} \tag{14}$$

According to the real polynomial coefficients obtained with the methods in [9] and [11], it is reasonable to assume the SPT coefficients have integer and fractional parts. Therefore, we deal with a_m expressed as

$$a_m = \sum_{k=-F}^{S-1} b_{m,k} 2^k \quad ; \quad m = 1,\ldots,M \tag{15}$$

where $b_{m,k} \in \{-1,0,1\}$, and S and F are the wordlengths of the integer and fractional part. Each $b_{m,k} \neq 0$ is called term. Generally, it occupies one adder in implementation. Therefore, the number of terms per polynomial or the number of terms per coefficient is limited to a prescribed value, P. For simplicity, here we use the latter approach.

From (13) it is clear that $\varepsilon(\mathbf{a}) = \varepsilon(-\mathbf{a})$. Apparently, two global minimizers with opposite signs exist. However, we need the minimizer that provides positive gain of the filter. It is achieved by adding the constraint $H_{SCIC}(0,\mathbf{a}) > 0$ to the objective function. It simplifies the search because the optimization deals with only a half of the overall SPT coefficient space.

To obtain the optimum SPT coefficients, we employ the global optimization technique based on the interval analysis. Recently, such a technique has been used in the minimax design of symmetric non-recursive filters [18]. According to the paper referred to, solving the problem in (13)−(15) is based on the interval extension of the objective function. In our design of SCIC filter, the interval extension of the objective function in (13) can be easily obtained by using the extensions of elementary operations and functions.

IV. Design Examples

To illustrate the features of the proposed multiplierless SCIC filters, two examples of wideband filters are described. The first example presents simple sharpened CIC filters, whereas the second example describes a filter offering very high folding-band attenuations.

The optimum SPT polynomial coefficients of the SCIC filters are given in Table I, together with their maximum passband deviations and minimum folding-band attenuations. The coefficients are tabulated for various orders of sharpening polynomials, passband edge frequencies, and filter complexities, assuming $R = 10$, $K = 900$, and $S+F = 20$.

It is well known that for a given order of the CIC filter, the amplitude response negligibly changes the shape within the passband and folding bands for $R \geq 10$ [1]. In that sense, the tabulated SPT coefficients can be used for sharpening of any CIC filter with $R \geq 10$.

A. Simple SCIC Filters

Here, we describe the multiplierless sharpening of the CIC response with $N = 2$ and $R = 10$, by using the polynomial with $M = 3$ and $P = 1$, and the passband edge frequencies given in Table I. Figure 2 shows the magnitude responses of the SCIC filters obtained for $\omega_p = 0.02\pi$ and $\omega_p = 0.05\pi$. The filter with $\omega_p = 0.02\pi$ ensures the minimum folding-band attenuation of 81 dB, whereas the filter with $\omega_p = 0.05\pi$ exhibits the minimum attenuation of 106 dB and, as expected, higher passband droop. Other simple SCIC filters have the passband and folding-band responses placed somewhere between the described responses. From Table I, it is clear that provided filters exhibit rather high folding-band attenuations. However, they have the structure in which three general-purpose multipliers are replaced by only one adder. In addition, they have similar responses within $|\omega| \leq 0.5\pi/R$, what makes them suitable for uniform passband compensation.

B. SCIC Filter Offering High Alias Rejection

In this example, we illustrate the design of multiplierless SCIC filters with high alias rejection. Such a filter can be obtained by the sharpening of the CIC response with $N = 2$ and $R = 10$, by using the polynomial with $M = 4$ and $P = 2$. The passband edge frequency $\omega_p = 0.04\pi$ is chosen.

The optimum multiplierless filter is compared with the filter obtained by the Chebyshev sharpening of the same CIC response, but with real coefficients of the sharpening polynomial of the same order [9]. Figure 3 shows the magnitude responses of both filters. Generally, filters with SPT coefficients contain a real gain constant, which is usually not implemented in practice. Therefore, only for comparison purposes, our response is normalized to 0 dB at $\omega = 0$.

It is clear from Figure 3 that the responses are very similar. The Chebyshev response ensures the minimum folding-band attenuation of 141 dB, whereas the multiplierless filter exhibits somewhat smaller attenuation

TABLE I
SPT POLYNOMIAL COEFFICIENTS, PASSBAND DROOP (A_P), AND MINIMUM FOLDING-BAND ATTENUATION (A_S) OF VARIOUS MINIMAX SHARPENED CIC FILTERS WITH $N = 2$ AND $R \geq 10$

***M*=3**										
a_m	$\omega_p=0.2\pi/R$		$\omega_p=0.25\pi/R$		$\omega_p=\pi/(3R)$		$\omega_p=0.4\pi/R$		$\omega_p=0.5\pi/R$	
	P=1	P=2	P=1	P=2	P=1	P=2	P=1	P=2	P=1	P=2
a_1	2^{-14}	$2^{-14}+2^{-19}$	2^{-12}	$2^{-12}-2^{-14}$	2^{-10}	$2^{-10}-2^{-12}$	2^{-11}	$2^{-9}+2^{-14}$	2^{-8}	$2^{-8}-2^{-12}$
a_2	-2^{-6}	$-2^{-6}-2^{-10}$	-2^{-5}	$-2^{-5}+2^{-11}$	-2^{-4}	$-2^{-4}-2^{-9}$	-2^{-4}	$-2^{-3}+2^{-7}$	-2^{-3}	$-2^{-3}-2^{-8}$
a_3	2^{0}	$2^{0}-2^{-5}$	2^{0}	$2^{0}+2^{-3}$	2^{0}	$2^{0}+2^{-2}$	2^{0}	$2^{1}-2^{-1}$	2^{0}	$2^{0}+2^{-7}$
A_P, dB	0.86	0.86	1.35	1.35	2.43	2.42	3.52	3.54	5.69	5.70
A_S, dB	132	142	125	129	106	113	94.9	102	81.0	89.3
***M*=4**										
a_m	$\omega_p=\pi/(3R)$		$\omega_p=0.4\pi/R$		$\omega_p=0.5\pi/R$		$\omega_p=0.6\pi/R$		$\omega_p=2\pi/(3R)$	
	P=1	P=2	P=1	P=2	P=1	P=2	P=1	P=2	P=1	P=2
a_1	0	$-2^{-16}+2^{-19}$	2^{-15}	$-2^{-15}-2^{-17}$	2^{-13}	$-2^{-12}+2^{-18}$	-2^{-14}	$-2^{-11}-2^{-14}$	-2^{-12}	$-2^{-9}+2^{-14}$
a_2	2^{-10}	$2^{-9}-2^{-14}$	-2^{-14}	$2^{-8}+2^{-14}$	2^{-9}	$2^{-6}+2^{-10}$	2^{-6}	$2^{-5}-2^{-7}$	2^{-6}	$2^{-4}+2^{-10}$
a_3	-2^{-4}	$-2^{-4}-2^{-6}$	-2^{-4}	$-2^{-3}+2^{-8}$	-2^{-3}	$-2^{-2}-2^{-4}$	-2^{-2}	$-2^{-2}-2^{-5}$	-2^{-2}	$-2^{-1}-2^{-3}$
a_4	2^{0}	$2^{0}+2^{-8}$	2^{0}	$2^{0}+2^{-3}$	2^{0}	$2^{1}-2^{-2}$	2^{0}	$2^{0}+2^{-6}$	2^{0}	$2^{1}-2^{-3}$
A_P, dB	3.23	3.24	4.67	4.73	7.50	7.62	11.4	11.5	14.3	14.7
A_S, dB	144	150	128	139	110	120	96.4	105	87.7	97.6
***M*=5**										
a_m	$\omega_p=0.5\pi/R$		$\omega_p=0.6\pi/R$		$\omega_p=2\pi/(3R)$		$\omega_p=0.75\pi/R$		$\omega_p=0.8\pi/R$	
	P=1	P=2	P=1	P=2	P=1	P=2	P=1	P=2	P=1	P=2
a_1	-2^{-18}	2^{-18}	-2^{-16}	2^{-14}	-2^{-14}	$2^{-13}+2^{-18}$	2^{-12}	$2^{-11}-2^{-13}$	-2^{-10}	$2^{-12}-2^{-17}$
a_2	2^{-12}	$-2^{-11}+2^{-17}$	2^{-13}	-2^{-8}	2^{-9}	$-2^{-7}+2^{-12}$	2^{-11}	$-2^{-6}+2^{-10}$	2^{-8}	$-2^{-6}-2^{-10}$
a_3	2^{-10}	$2^{-6}+2^{-10}$	2^{-6}	$2^{-4}+2^{-6}$	2^{-8}	$2^{-3}+2^{-6}$	-2^{-13}	$2^{-2}-2^{-4}$	2^{-3}	$2^{-2}-2^{-7}$
a_4	-2^{-3}	$-2^{-2}+2^{-5}$	-2^{-2}	$-2^{-1}-2^{-3}$	-2^{-2}	$-2^{0}-2^{-6}$	-2^{-2}	$-2^{0}+2^{-5}$	-2^{0}	$-2^{0}-2^{-2}$
a_5	2^{0}	$2^{0}-2^{-6}$	2^{0}	$2^{1}-2^{-2}$	2^{0}	$2^{1}+2^{-1}$	2^{0}	$2^{1}-2^{-2}$	2^{1}	$2^{1}+2^{-3}$
A_P, dB	9.32	9.52	14.0	14.4	17.7	18.3	22.9	24.7	28.7	29.2
A_S, dB	139	150	122	132	109	123	93.6	111	91.7	102

of 139 dB. In the passband, the filters introduce nearly the same droop of 4.73 dB.

From the complexity point of view, the Chebyshev SCIC filter needs five general-purpose multipliers to incorporate sharpening polynomial in the structure. However, our filter employs only six adders instead.

V. CONCLUSION

The design of multiplierless sharpened CIC filters based on minimax approximation in folding bands was presented. The design was formulated as an unconstrained optimization problem. The optimum polynomial coefficients are obtained by using the global optimization technique based on the interval analysis. The proposed filters exhibit similar amplitude behavior as the Chebyshev sharpened CIC filters. From the complexity point of view, the multiplierless filters are favorable, because they bring significant reduction in the structure.

VI. REFERENCES

[1] E. B. Hogenauer, "An economical class of digital filters for decimation and interpolation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 2, pp. 155–162, Apr. 1981.

[2] T. Saramäki and T. Ritoniemi, "A modified comb filter structure for decimation," in *Proc. IEEE Int. Symp. Circuits Syst.*, vol. 4, Hong Kong, Jun. 1997, pp. 2353–2356.

[3] J. Kaiser and R. Hamming, "Sharpening the response of a symmetric nonrecursive filter by multiple use of the same filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-25, no. 5, pp. 415–422, Oct. 1977.

[4] A. Y. Kwentus, Z. Jiang, and A. N. Willson Jr., "Application of filter sharpening to cascaded integrator-comb decimation filters," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 457–467, Feb. 1997.

[5] G. Stephen and R. W. Stewart, "High-speed sharpening of decimating CIC filter," *Electron. Lett.*, vol. 40, no. 21, pp. 1383–1384, Oct. 2004.

[6] G. Jovanovic Dolecek and S. K. Mitra, "A new two-stage sharpened comb decimator," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 52, no. 7, pp. 1414–1420, Jul. 2005.

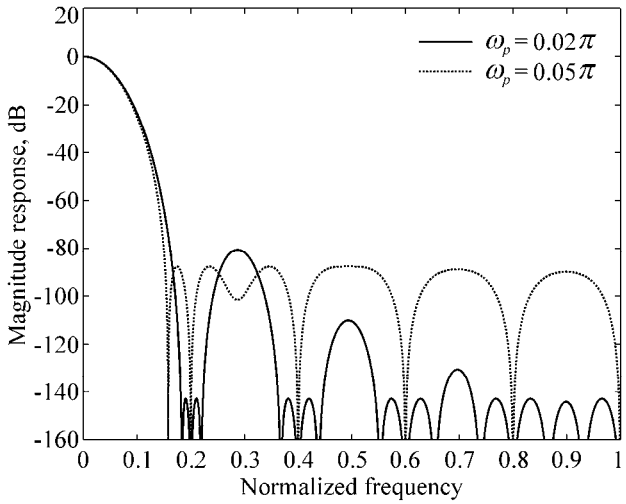


Figure 2. Magnitude responses of sharpened CIC filters with $N = 2$ and $R = 10$, obtained by minimax SPT sharpening with $M = 3$ and $P = 1$, assuming passbands are $|\omega| \le 0.02\pi$ and $|\omega| \le 0.05\pi$.

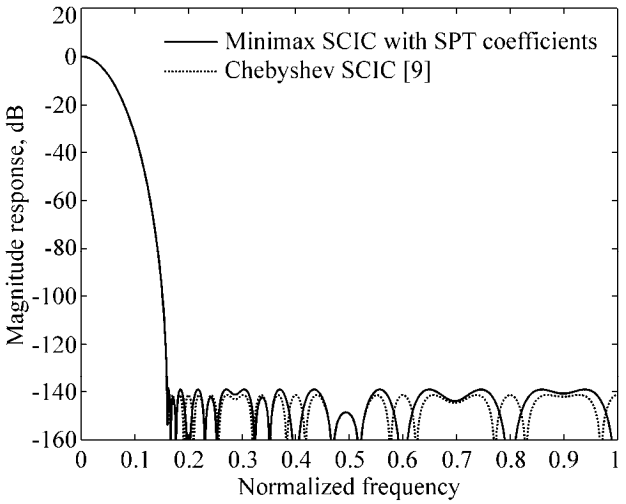


Figure 3. Magnitude responses of sharpened CIC filters with $N = 2$ and $R = 10$, obtained by minimax SPT sharpening with $M = 4$ and $P = 2$, and by Chebyshev sharpening of fourth order. Passband is $|\omega| \le 0.04\pi$.

[7] Q. Liu and J. Gao, "Efficient comb decimation filter with sharpened magnitude response," in *Proc. 5th Int. Conf. WICOM*, 2009, pp. 1–4.

[8] H. Zaimin, H. Yonghui, K. Wang, J. Wu, J. Hou, and L. Ma, "A novel CIC decimation filter for GNSS receiver based on software defined radio," in *Proc. 7th Int. Conf. WICOM*, 2011, pp. 1–4.

[9] J. O. Coleman, "Chebyshev stopbands for CIC decimation filters and CIC-implemented array tapers in 1D and 2D," *IEEE Trans. Circuits Syst. I: Reg. Papers*, vol. 59, no. 12, pp. 2956–2968, Dec. 2012.

[10] G. Molnar, M. Glavinic Pecotic, and M. Vucic, "Weighted least-squares design of sharpened CIC filters," in *Proc. 36th Int. Conf. MIPRO*, vol. MEET, Opatija, Croatia, May 2013, pp. 104–108.

[11] G. Molnar and M. Vucic, "Weighted minimax design of sharpened CIC filters," in *Proc. IEEE 20th ICECS*, Abu Dhabi, UAE, Dec. 2013, pp. 869–872.

[12] M. Laddomada, D. E. Troncoso, and G. Jovanovic Dolecek, "Improved sharpening of comb-based decimation filters: Analysis and design," in *Proc. IEEE 11th Int. Conf. CCNC*, 2014, pp. 11–16.

[13] M. G. C. Jiménez, V. C. Reyes, and G. Jovanovic Dolecek, "Sharpening of non-recursive comb decimation structure," in *Proc. 13th ISCIT*, 2013, pp. 458–463.

[14] R. J. Hartnett and G. F. Boudreaux-Bartels, "Improved filter sharpening," *IEEE Trans. Signal Process.*, vol. 43, no.12, pp. 2805–2810, Dec. 1995.

[15] S. Samadi, "Explicit formula for improved filter sharpening polynomial," *IEEE Trans. Signal Process.*, vol. 48, no. 10, pp. 2957–2959, Nov. 2000.

[16] D. E. Troncoso Romero, M. Laddomada, and G. Jovanovic Dolecek, "Optimal sharpening of compensated comb decimation filters: analysis and design," *Sci. World J.*, Hindawi, vol. 2014, ID 950860, 9 pages.

[17] M. G. C. Jiménez, D. E. Troncoso Romero and G. Jovanovic Dolecek, "On simple comb decimation structure based on Chebyshev sharpening," in *Proc. IEEE 6th LASCAS*, Feb. 2015, pp. 1–4.

[18] M. Vucic, G. Molnar, and T. Zgaljic, "Design of FIR filters based on interval analysis," in *Proc. 33rd Int. Conf. MIPRO*, vol. MEET & GVS, Opatija, Croatia, May 2010, pp. 197–202.

Minimization of Maximum Electric Field in High-Voltage Parallel-Plate Capacitor

Raul Blecic*†, Quentin Diduck‡, Adrijan Baric*
*University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
Tel: +385 (0)1 6129547, fax: +385 (0)1 6129653, e-mail: raul.blecic@fer.hr
†KU Leuven, ESAT-TELEMIC, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
‡Ballistic Devices Inc, 904 Madison St, Santa Clara, CA, 95050 United States

***Abstract*—Minimization of maximum electric field of a parallel-plate capacitor for high-voltage and temperature stable applications is presented. Cubic zirconia is used as a dielectric material because of its high relative permittivity, high dielectric strength and high temperature stability. The maximum electric field present in the structure limits the maximum achievable capacitance of the capacitor structure. Reducing the maximum electric field of the capacitor allows to reduce the thickness of the dielectric material, which increases its capacitance. The impact of geometrical and electrical parameters of the parallel-plate capacitor on the maximum electric field is analyzed by a 2D multiphysics solver. The guidelines for the minimization of the maximum electric field are given.**

***Index Terms*—cubic zirconia, dielectric strength, edge effects, electrostatics, fringing fields.**

I. Introduction

High relative permittivity, high dielectric strength and high temperature stability of cubic zirconia [1], [2], [3] make it a candidate for a dielectric material for high-voltage, temperature stable capacitors.

Most of the capacitance of a parallel-plate capacitor is contained in the central part of the structure. Yet, the electric field at the edges of the structure can be several times higher than the field in its center [4], which reduces the maximum achievable capacitance of the capacitor. Reducing the maximum electric field present in the capacitor allows to reduce the thickness of the dielectric material which increases the capacitance.

The objective of this paper is to analyze the impact of geometrical and electrical parameters of a parallel-plate capacitor on the maximum electric field in order to minimize the maximum electric field present in the capacitor and to consequently maximize its capacitance. The analysis is performed by a 2D multiphysics solver.

This paper is structured as follows. Section II introduces a parallel-plate capacitor. Section III presents a numerical analysis of the impact of electrical and geometrical parameters of the structure on the maximum electric field. Section IV concludes the paper.

II. Capacitance of a Parallel-Plate Capacitor

Capacitance of a parallel-plate capacitor, neglecting the contribution of the fringing fields, can be expressed as:

$$C_{pp} = \epsilon \cdot \frac{S}{d}, \tag{1}$$

where ϵ is the permittivity of the material between the plates, S is the area of the plates and d is the separation between the plates.

The total capacitance of a parallel-plate capacitor, taking the fringing fields into account (assuming that the plates are completely surrounded by the same material), can be expressed by the Palmer's equation [5]:

$$C_{total} = \epsilon \cdot \frac{S}{d} \cdot K_a \cdot K_b, \tag{2}$$

$$K_a = \left(1 + \frac{d}{\pi \cdot a}\left(1 + \ln\left(\frac{2 \cdot \pi \cdot a}{d}\right)\right)\right), \tag{3}$$

$$K_b = \left(1 + \frac{d}{\pi \cdot b}\left(1 + \ln\left(\frac{2 \cdot \pi \cdot b}{d}\right)\right)\right), \tag{4}$$

where ϵ is the permittivity of the material which surrounds the plates, while a and b are the width and the length of the plates ($S = a \cdot b$).

The relative contribution of the fringing fields to the total capacitance of a square parallel-plate capacitor calculated as $C_{ff} = C_{total} - C_{pp}$ is shown in Fig. 1. The contribution of the fringing field capacitance to the total capacitance of the parallel-plate capacitor is negligible if the size of the plates is large relative to the separation of the plates ($a >> d$ and $b >> d$), which is typically true for high-capacitance capacitors.

Although the contribution of the fringing fields to the capacitance of a parallel-plate capacitor is negligible for large sizes of plates, the fringing fields can be several times larger than the field in the center of the capacitor [4]. The fringing fields limit the minimum separation of the plates and the maximum achievable capacitance of the capacitor. Reducing the maximum electric field enables the reduction of the separation between the plates, which increases the capacitance.

III. 2D Simulations

A. Simulation Domain

The simulation domain of a parallel-plate capacitor in a 2D multiphysics solver is shown in Fig. 2. It consists of two 0.1-mm thick copper plates. Material between the copper plates is a 0.15-mm thick cubic zirconia which has a high relative permittivity ($\epsilon_{r,CZ} = 17$), high dielectric strength and high temperature stability. The capacitor is enclosed by a surrounding material. The surrounding material has to be able to sustain large electric fields

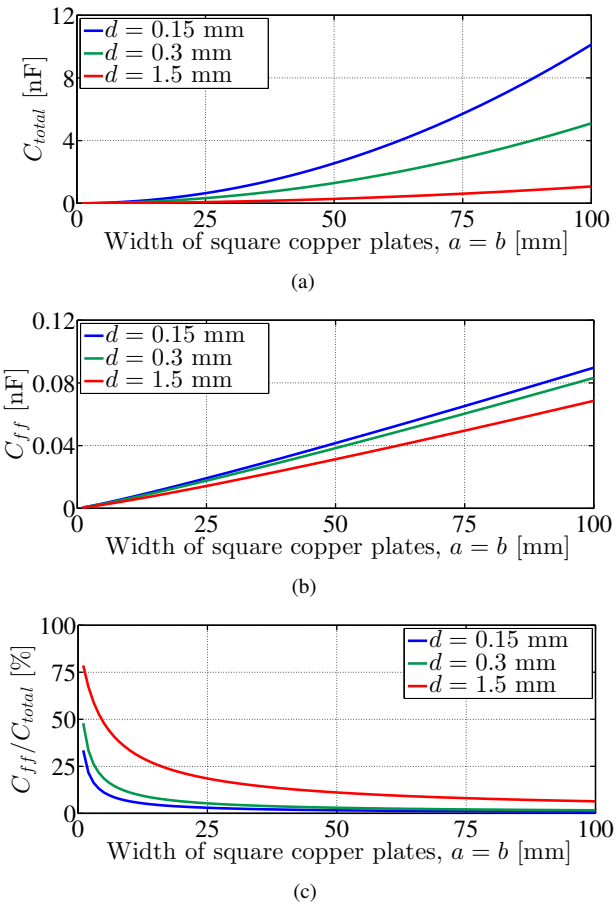


Fig. 1. Capacitance of a square parallel-plate capacitor as a function of the size of the plates: (a) total capacitance, (b) fringing-field capacitance ($C_{ff} = C_{total} - C_{pp}$), and (c) relative contribution of the fringing fields ($C_{ff} = C_{total} - C_{pp}$) to the total capacitance.

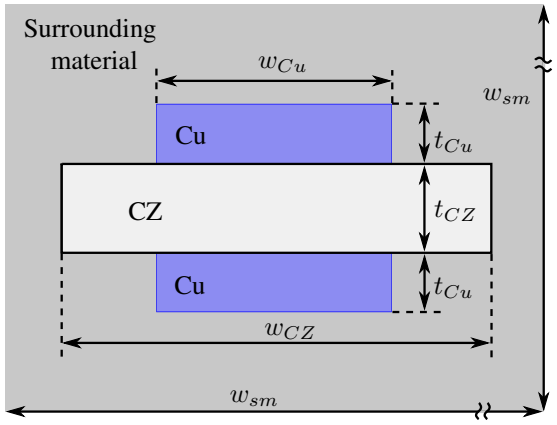


Fig. 2. Simulation domain in a 2D multiphysics solver.

present at the edges of the copper plates (has to have a high dielectric strength), however, its relative permittivity and temperature stability do not significantly contribute to the overall capacitance and temperature stability of the capacitor.

B. Methodology

Stationary electrostatic analysis is performed. Boundary condition is set to perfect insulator (modelled as the zero charge boundary condition [6]) and it is placed sufficiently far away from the parallel-plate capacitor to have a negligible impact on the simulation results. The top copper plate is set to 1 V, while the bottom copper plate is set to 0 V. This gives the electric field in the center of the structure equal to $E_{center} = (1\text{ V})/t_{CZ} = 6.67$ kV/m.

TABLE I
PARAMETERS OF THE SIMULATED STRUCTURE.

Parameter	Value
Width of the copper plates, w_{Cu} [mm]	1
Width of the surrounding material, w_{sm} [mm]	5
Thickness of the cubic zirconia, t_{CZ} [mm]	0.15
Thickness of the copper plates, t_{Cu} [mm]	0.1
Relative permittivity of the cubic zirconia, $\epsilon_{r,CZ}$	17
Losses of the cubic zirconia, $\tan\delta_{CZ}$	0
Losses of the surrounding material, $\tan\delta_{sm}$	0
Conductivity of copper, σ_{Cu} [S/m]	6e7

TABLE II
VARIABLES OF THE SIMULATED STRUCTURE.

Variable	Values
Width of the cubic zirconia, w_{CZ} [mm]	0.8–1.6
Relative permittivity of the surrounding material, $\epsilon_{r,sm}$	1, 4, 20, 100

The parameters and variables of the simulated structure are given in Table I and Table II, respectively. The width of the copper plates is set to a small value ($w_{Cu} = 1$ mm) to reduce the simulation domain. In actual application, it is much larger to provide a sufficiently large capacitance.

C. Impact of the Width of Cubic Zirconia and the Dielectric Constant of the Surrounding Material

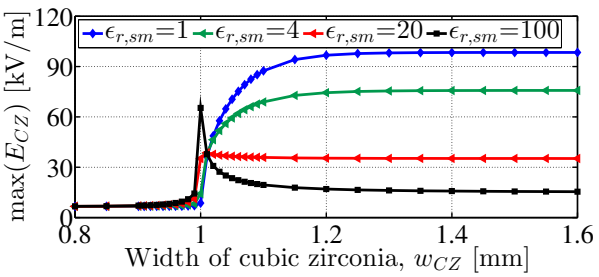
Parametric simulations for the different widths of cubic zirconia w_{CZ} and for four values of the dielectric constant of the surrounding material $\epsilon_{r,sm}$ is performed. The maximum electric fields in the cubic zirconia and in the surrounding material are shown in Fig. 3.

The results can be divided into three regions, the first region in which the width of the cubic zirconia is smaller than the width of the copper plates ($w_{CZ} < 0.95$ mm), the second region in which the width of the cubic zirconia is larger than the width of the copper plates ($w_{CZ} > 1.1$ mm), and the third region in which the two sizes are comparable ($0.95 < w_{CZ} < 1.1$ mm).

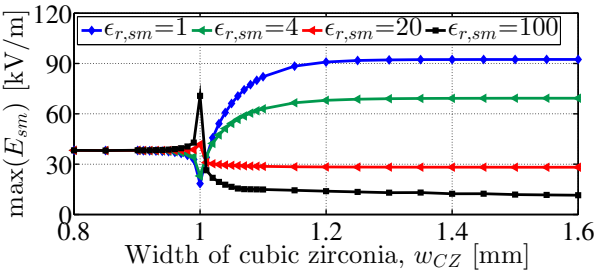
The results in the first region show that the maximum electric field in the cubic zirconia decreases as its width decreases. Also, it shows a weak dependence on the relative permittivity of the surrounding material. The maximum field in the surrounding material does not depend on its relative permittivity nor on the width of the cubic zirconia.

In the second region, the maximum electric fields in the cubic zirconia and in the surrounding material depend on the relative permittivity of the surrounding material. Increasing it decreases the maximum electric field in both materials.

In the third region, the maximum electric fields in



(a)



(b)

Fig. 3. Maximum electric field as a function of the width of the cubic zirconia and the relative permittivity of the surrounding material: (a) in the cubic zirconia and (b) in the surrounding material.

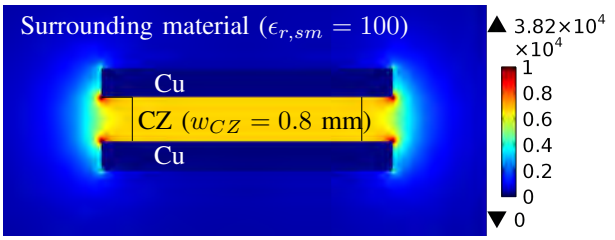
both materials depend on the relative permittivity of the surrounding material. Increasing it increases the maximum electric field in both materials. For relative permittivity of the surrounding material larger than that of the cubic zirconia $\epsilon_{r,sm} > \epsilon_{r,CZ}$, the peak maximum electric field in both materials is obtained for $w_{CZ} = w_{Cu}$.

Comparing all three regions, the first region shows the smallest maximum electric field in the cubic zirconia. The second region, for the relative permittivity of the surrounding material larger than that of the cubic zirconia $\epsilon_{r,sm} > \epsilon_{r,CZ}$, shows the smallest maximum field in the complete structure (in the dielectric material and in the surrounding material). Also, the impact of the surrounding material on the temperature stability of the capacitor is expected to be the smallest in the second region since the surrounding material is not present between the plates. Therefore, the second region (for relative permittivity of the surrounding material larger than that of the cubic zirconia $\epsilon_{r,sm} > \epsilon_{r,CZ}$) is considered to be the best solution for the temperature stable applications. The third region does not provide any significant advantage over the first and the second region.

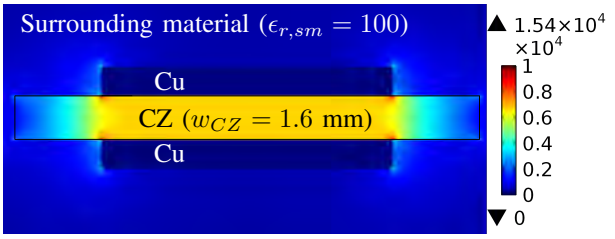
The electric field distributions in the simulated structure for three characteristic cases ($w_{CZ} = 0.8$ mm, $\epsilon_{r,sm} = 100$; $w_{CZ} = 1.6$ mm, $\epsilon_{r,sm} = 100$ and $w_{CZ} = 1.6$ mm, $\epsilon_{r,sm} = 1$) are shown in Fig. 4. The maximum electric fields in the cubic zirconia and in the surrounding material for $w_{CZ} = 0.8$ mm and for $w_{CZ} = 1.6$ mm are given in Table III.

D. Impact of the Shape of the Copper Plates

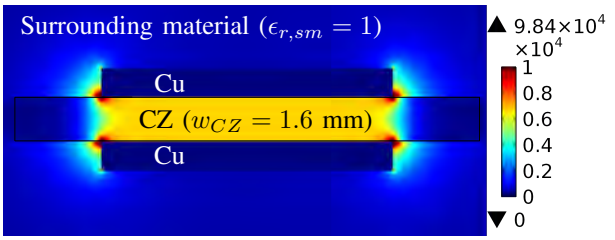
The impact of the shape of the copper plates on the maximum electric field is analyzed as follows. Four structures are analyzed and compared. The first one is



(a)



(b)



(c)

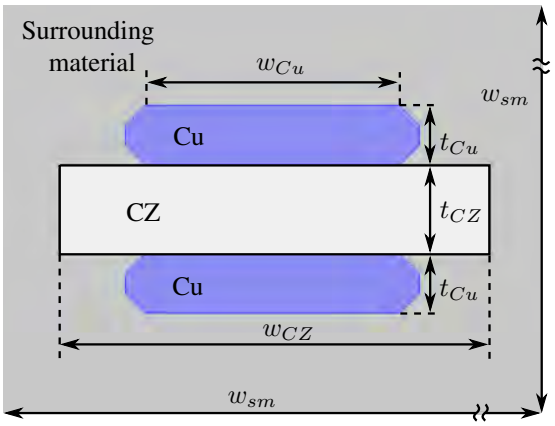
Fig. 4. Electric field in the simulated structure for: (a) $w_{CZ} = 0.8$ mm, $\epsilon_{r,sm} = 100$, (b) $w_{CZ} = 1.6$ mm, $\epsilon_{r,sm} = 100$ and (c) $w_{CZ} = 1.6$ mm, $\epsilon_{r,sm} = 1$. The maximum electric field can be observed at the edges of the copper plates.

TABLE III
MAXIMUM ELECTRIC FIELD IN THE CUBIC ZIRCONIA AND IN THE SURROUNDING MATERIAL FOR THE WIDTH OF THE CUBIC ZIRCONIA $w_{CZ} = 0.8$ MM AND $w_{CZ} = 1.6$ MM, AND FOR THE RELATIVE PERMITTIVITY OF THE SURROUNDING MATERIAL $\epsilon_{r,sm} = 1, 4, 20, 100$.

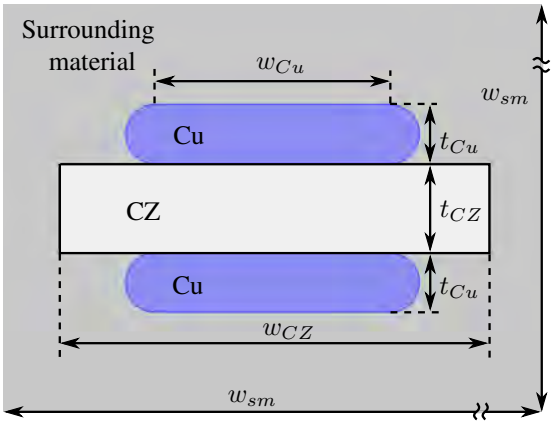
$\epsilon_{r,sm}$	w_{CZ} [mm]	max(E_{CZ}) [kV/m]	max(E_{sm}) [kV/m]
1	0.8	6.67	38.24
	1.6	98.42	92.41
4	0.8	6.67	38.24
	1.6	75.79	69.31
20	0.8	6.70	38.24
	1.6	35.21	28.17
100	0.8	6.72	38.24
	1.6	15.43	11.48

the structure with rectangular edges as shown in Fig. 2. The second structure has triangular edges (cut at 45°) (Fig. 5(a)), the third has round edges (with the radius $r = t_{Cu}/2 = 0.05$ mm) (Fig. 5(b)), while the fourth has Rogowski profile edges ($\psi = \pi/2$) terminated with a circular section [7] (with the radius $r = 0.031$ mm) (Fig. 5(c)). The parameters of the simulated structures are given in Table I. The width of the cubic zirconia is set to $w_{CZ} = 1.6$ mm.

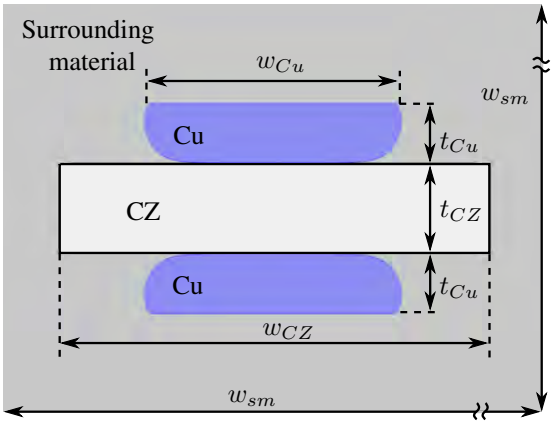
The maximum electric fields in the cubic zirconia and in



(a)



(b)



(c)

Fig. 5. The shapes of copper plates of the parallel-plate capacitor: (a) triangular edges, (b) round edges ($r = t_{Cu}/2 = 0.05$ mm) and (c) Rogowski profile edges ($\psi = \pi/2$) terminated with a circular section ($r = 0.031$ mm).

the surrounding material for the four simulated structures are given in Table IV. The results show that the Rogowski profile edges provide the smallest maximum electric field in the cubic zirconia among the four simulated structures for any value of the relative permittivity of the surrounding material. The maximum electric field in the surrounding material for the structure with the triangular, round or Rogowski profile edges is significantly reduced only if the relative permittivity of the surrounding material is larger than that of the cubic zirconia ($\epsilon_{r,sm} > \epsilon_{r,CZ}$).

TABLE IV
MAXIMUM ELECTRIC FIELD IN THE CUBIC ZIRCONIA AND IN THE SURROUNDING MATERIAL FOR THE RECTANGULAR, TRIANGULAR AND ROUND EDGES OF THE COPPER PLATES, AND FOR THE RELATIVE PERMITTIVITY OF THE SURROUNDING MATERIAL $\epsilon_{r,sm} = 1, 4, 20, 100$ ($w_{CZ} = 1.6$ MM).

$\epsilon_{r,sm}$	shape of edges	max(E_{CZ}) [kV/m]	max(E_{sm}) [kV/m]
1	rectangular	98.42	92.41
	triangular	91.67	115.72
	round	20.07	332.40
	Rogowski profile	7.37	122.08
4	rectangular	75.79	69.31
	triangular	54.56	64.04
	round	12.25	51.03
	Rogowski profile	6.81	29.03
20	rectangular	35.21	28.17
	triangular	19.26	13.63
	round	8.54	7.65
	Rogowski profile	6.67	9.04
100	rectangular	15.43	11.48
	triangular	9.34	7.08
	round	7.45	3.17
	Rogowski profile	6.72	4.63

Although the structure with the Rogowski profile edges shows the smallest maximum electric field, the design of such structure with solid dielectric materials presents an additional challenge from a mechanical point of view. The maximum electric field in the cubic zirconia for the triangular edges and for the relative permittivity of the surrounding material equal to $\epsilon_{r,sm} = 100$ is 40% larger than the electric field in the center of the structure, while for the round edges and for the relative permittivity of the surrounding material equal to $\epsilon_{r,sm} = 100$ is only 12% larger than the electric field in the center of the structure which provides a significant improvement when compared to the structure with the rectangular edges.

IV. CONCLUSION

The impact of geometrical and electrical parameters of the parallel-plate capacitor, with the cubic zirconia used as a dielectric material, on the maximum electric field present in the capacitor structure is analyzed by a 2D multiphysics solver. The maximum electric field present in the capacitor can be significantly reduced if the width of the dielectric is designed to be larger than the width of the copper plates and if the capacitor is enclosed by a material with a large relative permittivity, i.e. larger than the relative permittivity of the dielectric material. The maximum electric field can be further reduced by patterning the edges of the copper plates in a triangular, round or Rogowski profile shape.

REFERENCES

[1] C.-H. Chang, B. Hsia, J. P. Alper, S. Wang, L. E. Luna, C. Carraro, S.-Y. Lu, and R. Maboudian, "High-temperature All Solid-state Microsupercapacitors Based on SiC Nanowires Electrode and YSZ Electrolyte," *ACS Appl. Mater. Interfaces*, vol. 7, pp. 26 658–26 665, 2015.

[2] O. Jongprateep, V. Petrosky, and F. Dogan, "Yttria Stabilized Zirconia as a Candidate for High Energy Density Capacitor," *Kasetsart Journal*, vol. 42, pp. 373–377, 2008.

[3] ——, "Effects of Yttria Concentration and Microstructure on Electric Breakdown of Yttria Stabilized Zirconia," *Journal of Metal, Materials and Minerals*, vol. 18, no. 1, pp. 9–14, 2008.

[4] K. P. P. Pillai, "Fringing field of finite parallel-plate capacitors," *Proceedings of the Institution of Electrical Engineers*, vol. 117, no. 6, pp. 1201–1204, June 1970.

[5] M. Hosseini, G. Zhu, and Y.-A. Peter, "A new formulation of fringing capacitance and its application to the control of parallel-plate electrostatic micro actuators," in *Analog Integr. Circ. Sig. Process*, 2007, pp. 119–128.

[6] COMSOL, *AC/DC Module User's Guide*. COMSOL, Nov. 2013.

[7] N. G. Trinh, "Electrode Design for Testing in Uniform Field Gaps," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-99, no. 3, pp. 1235–1242, May 1980.

Modelling SMD Capacitors by Measurements

Roko Mišlov*, Marko Magerl*, Sandra Fratte-Sumper†, Bernhard Weiss†, Christian Stockreiter† and Adrijan Barić*
*University of Zagreb Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
Email: roko.mislov@fer.hr
†ams AG, Tobelbader Strasse 30, Premstaetten 8141, Austria
Email: christian.stockreiter@ams.com

***Abstract*—The lumped models of several capacitors in SMD-0603 package are extracted from S-parameter measurements. The S-parameters of the SMD components are obtained: (i) by modelling the RF connector and transmission lines and de-embedding, and (ii) by directly calibrating the reference plane to the SMD component using on-board calibration standards. The extracted parasitics related to the SMD component package are compared for capacitors of different nominal values and for the two calibration methods in the frequency range up to 8 GHz.**

***Index Terms*—SMD capacitor modelling, SMA connector modelling, CPW-CB modelling, S-parameter measurements**

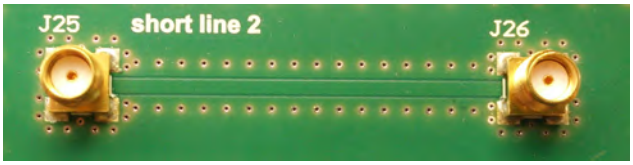
I. Introduction

Surface mount (SMD) capacitors are used in many electronics applications, e.g. decoupling in power delivery networks (PDN) [1], injection of the RF disturbance in electromagnetic immunity measurements, such as direct power injection (DPI) [2], or in the output filters for DC/DC converters [3]. A high-frequency capacitor model available in the design phase can significantly improve the performance of the first-pass design.

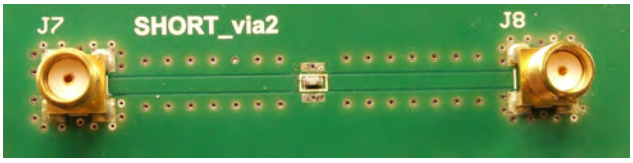
Several modelling methods for passive SMD components are reported in literature [4], [5], [6], where the lumped or lumped-distributed models are built using S-parameters of the SMD components. The S-parameters are measured using a vector network analyzer (VNA), primarily because of its high dynamic range and broad frequency range. Alternatively, the S-parameters can be obtained by applying the inverse Fourier transform on the time-domain reflectometry (TDR) measurements [7].

The reference plane of the VNA measurements is calibrated to the device-under-test (DUT) by measuring well-defined short, open, load and through (SOLT) calibration standards. The VNA calibration kit enables calibration up to the end of the coaxial cables. In [4] on-board calibration standards are used to calibrate the reference up to the DUT on the PCB. The on-board open calibration standard is characterized using VNA port extension, based on the assumption that the remaining on-board calibration standards are well-defined.

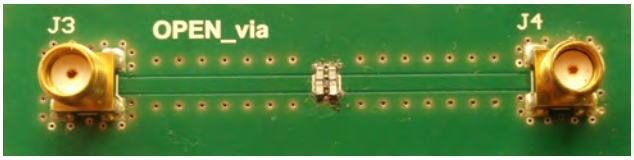
An alternative calibration approach is to model the fixture between the coaxial cables and the DUT: the radio-frequency (RF) connectors and the transmission lines. The model can be used to de-embed the fixture from the measurements obtained by calibrating the reference plane using the well-defined calibration standards in the VNA calibration kit.



(a) Test structure (i): SMA connectors connected by CPWG lines.



(b) Test structure (ii): SMA connectors, CPWG lines and the DUT.



(c) Test structure (ii): SMA connectors, CPWG lines and the OPEN standard.

Fig. 1: Manufactured test structures.

In this paper, both of these calibration methods are used and compared. Instead of using port extension, the on-board calibration standards are characterized by de-embedding their S-parameters using the fixture model. Additional test structures are manufactured and measured, in order to characterize the substrate, the RF connectors and the transmission lines. The transmission lines are designed as conductor-backed coplanar waveguide (CPWG) according to the design guidelines given in [5]. Three capacitors of different nominal values, but in identical SMD-0603 package are modelled.

This paper is organised as follows. The measurement setup is presented in Section II, and the measurement results are shown in Section III. The modelling methodology and results are presented in Section IV, and discussed in Section V. The paper is concluded in Section VI.

II. Measurement Setup

A. Test structures

Two types of test structures are manufactured on printed circuit boards (PCB): (i) structures for FR4 substrate char-

TABLE I: The nominal substrate values and the 50 Ω CPWG line geometry calculated using [11].

parameter	nominal value
ε_r	4.5
$\tan\delta$	0.02
substrate height [mm]	1.5
Cu thickness [µm]	35
trace width [mm]	1.548
clearence to GND [mm]	0.3

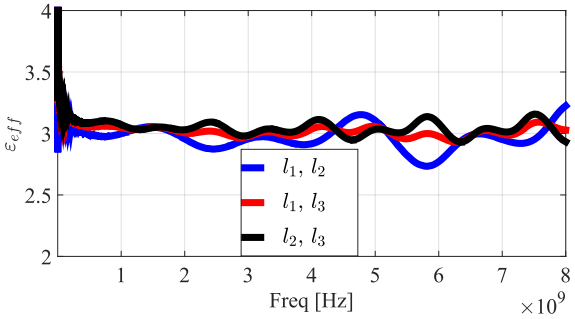


Fig. 2: The effective permittivity of the FR4 substrate extracted from measurements of test structures (i).

acterization and RF connector modelling, and (ii) structures for measuring the SMD capacitors and on-board calibration standards. Test structures (i), shown in Fig. 1a, consist of two SubMiniature version A (SMA) connectors [8] connected by CPWG transmission lines of several different lengths. The vias connecting the ground planes are placed along the trace, as recommended in [5].

Test structures (ii), shown in Figs. 1b, 1c, consist of pads for 2-port measurements of SMD-0603 components connected to SMA connectors by CPWG transmission lines. In Fig. 1b this test structure is used to characterize the DUT as a 2-port component, while in Fig. 1c the same test structure is used to measure the on-board open calibration standard. The on-board open consists of two parallel 1 pF capacitors soldered between each DUT pad and ground. All components of the on-board calibration standards have identical SMD-0603 package as the modelled SMD capacitors.

The CPWG transmission lines are designed to have the characteristic impedance of 50 Ω on the FR4 substrate. The transmission line geometric parameters, as well as the nominal substrate parameters used for their calculation, are summarized in Table I. The S-parameters are measured in the frequency range from 300 kHz to 8 GHz using a 2-port VNA [9]. Each measurement is taken in 1600 frequency points in the logarithmic scale using a 500 Hz resolution bandwidth and an averaging factor of 3. The modelling is done using Advanced Design System 2015.01 (ADS) [10].

B. Substrate characterization

The effective permittivity of the FR4 substrate is calculated using the S-parameter measurements of the test structures (i) described in Section II-A. The following

TABLE II: Substrate permittivity calculated using the phase difference of CPWG lines: l_1=25 mm, l_2=50 mm, l_3=100 mm.

used lines	$\varepsilon_{eff,avg}$	ε_r
l_1, l_2	2.973	4.848
l_1, l_3	3.017	4.935
l_2, l_3	3.039	4.980

TABLE III: Optimized values of the SMA connector and transmission line models.

parameter	value
C_1 [pF], L_1 [nH]	0.0044, 0.0326
C_2 [pF], L_2 [nH]	0.2798, 0.0010
C_3 [pF], L_3 [nH]	0.1573, 0.0466
l_{CPWG} [mm]	23.65
$\tan\delta$	0.0234

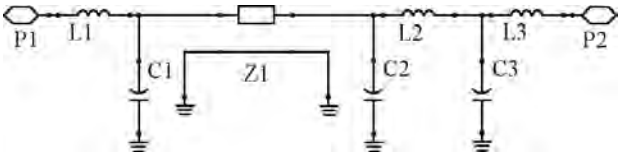


Fig. 3: The SMA connector model.

transmission line lengths are manufactured: l_1 = 25 mm, l_2 = 50 mm, l_3 = 100 mm. The phase difference between the S_{21} parameters of any two of these test structures is related to the difference of the physical line lengths by the equation:

$$\Delta\varphi = \beta\Delta l = \frac{2\pi}{\lambda_f}\Delta l = \frac{2\pi f\sqrt{\varepsilon_{eff}}}{c}\Delta l, \qquad (1)$$

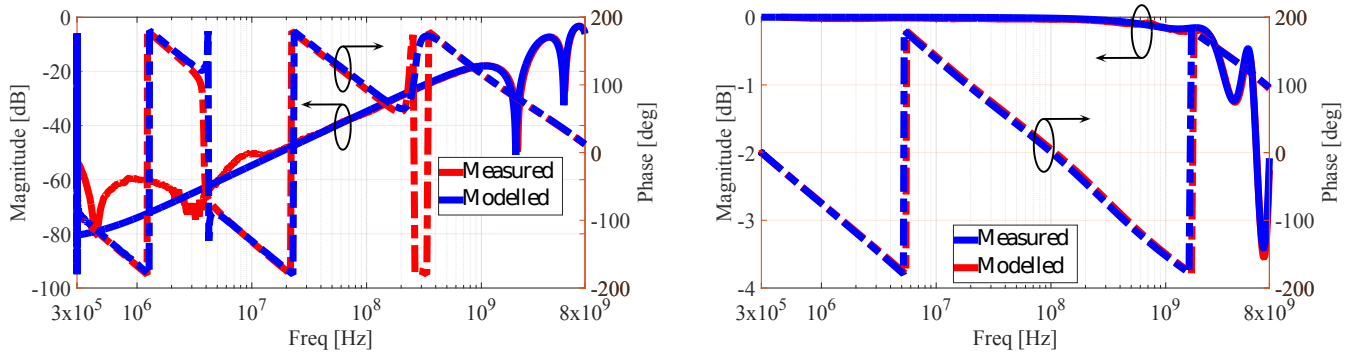
where Δl is the difference in physical line length of the two test structures, β is the wave number, $\Delta\varphi$ is the phase difference of the S_{21} parameters at each frequency f, and ε_{eff} is the effective permittivity. This relation enables the calculation of the effective permittivity as a function of frequency:

$$\varepsilon_{eff}(f) = \left(\frac{c\Delta\varphi}{2\pi f\Delta l}\right)^2, \qquad (2)$$

and it is shown in Fig. 2. The average value of the substrate permittivity over the frequency range from 1 GHz to 8 GHz is given in Table II for all three pairs of transmission line measurements. The value of the substrate permittivity ε_r is calculated from the average effective permittivity $\varepsilon_{eff,avg}$ using the CPWG transmission line model [12] and the frequency independent dielectric loss model of the substrate in ADS.

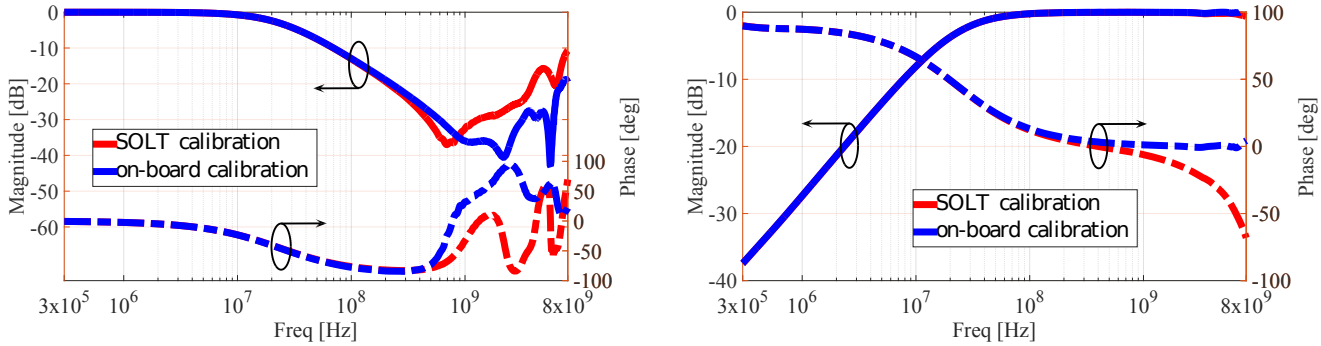
C. SMA connector modelling

The topology of the SMA connector model is a simplified version of the lumped-distributed connector model presented in [13]. The model is shown Fig. 3. The components L_1 and C_1 represent the transition between the coaxial cable and the SMA connector. The transmission



(a) $\mathbf{S_{11}}$ (b) $\mathbf{S_{21}}$

Fig. 4: Comparison between measurements and model of test structure (i) with the transmission line length l_1 = 25 mm.



(a) $\mathbf{S_{11}}$ (b) $\mathbf{S_{21}}$

Fig. 5: Comparison of the de-embedded S-parameters of the 68 pF capacitor for the SOLT and the on-board calibration.

line Z_1 models the coaxial section of the connector. The right-angle transition between the SMA connector and the CPWG line on the PCB is represented by the components C_2 and L_2, and the part of the transmission line that is directly below the connector is modelled by the components C_3 and L_3.

The values of the model parameters are obtained by minimizing the absolute difference between the measured and modelled S_{11} and S_{21} parameters of the test structures (i), described in Section II-A, using the random and gradient optimization solvers in ADS. The initial values of the optimized variables are chosen according to [13]. The optimization takes several minutes on an Intel® Core® i7 CPU @ 3.0 GHz and 12 GB of RAM.

The transmission lines l_1, l_2, l_3 are modelled by optimizing the CPWG model [12]. The lengths of these transmission lines are defined by a single variable l_{CPWG}, such that the modelled line lengths are equal to l_{CPW}, ($l_{CPWG} + 25$ mm), ($l_{CPWG} + 75$ mm). In this way the coupling effects between the SMA connector and the CPWG lines are taken into account consistently. The dielectric loss is modelled by optimizing the loss tangent of the substrate $\tan\delta$. The variables that are not optimized are the nominal impedance $Z_1 = 50\ \Omega$ of the SMA connector, the previously obtained substrate permittivity $\varepsilon_r = 4.848$, the copper conductivity σ_{Cu} = 58.5·10^6 S/m and copper thickness $t_{Cu} = 35$ µm. The random optimizer is used, followed by the gradient optimizer, with 100 iterations each. The final values of the optimized variables of the SMA connector model are given in Table III. The comparison between the measurements and the model of the test structures (i) with the line length l_1 = 25 mm is given in Fig. 4.

III. Measurement Results

The SMA connector and the CPWG line on each side of the DUT in the test structures (ii) is referred to as the fixture. In order to de-embed the fixture from the S-parameter measurements, the model of the fixture is built using the extracted characteristics of the substrate and the model of the SMA connector. The fixture return loss is better than 10 dB in the frequency range up to 4.9 GHz (the worst case return loss is 6.4 dB). The fixture insertion loss is less than 1.8 dB in the entire frequency range up to 8 GHz. The fixture model enables de-embedding of the fixture from the S-parameter measurements obtained by the VNA calibration kit, i.e. SOLT calibration up to the end of the coaxial cables.

The on-board calibration standards, i.e. the custom calibration kit, are soldered on the PCB pads where the component to be characterized will be soldered later. Then, each on-board standard is measured separately according

to the SOLT procedure and stored into the VNA memory. The short and through standards are implemented using 0 Ω resistors; the load standard are two 100 Ω resistors in parallel, and the open standard are two 1 pF capacitors in parallel. These SMD components have the same 0603 package as the characterized capacitors.

In the remainder of this paper, the measurements obtained using the custom calibration kit are referred to as "on-board calibration", while the DUT measurements obtained by de-embedding the fixtures are referred to as "SOLT calibration". The comparison between these two measurement methods is shown in Fig. 5, for the 68 pF capacitor in the SMD-0603 package.

IV. Modelling results

The models of the SMD capacitors are extracted using the π-model shown in Fig. 6. The admittances Y_1, Y_2 represent the parasitic coupling between the soldering pads and ground, while the admittance Y_3 represents the intrinsic SMD component. These admittances can be expressed in terms of the y-parameters using: $Y_3 = -(y_{12} + y_{21})/2$, $Y_1 = y_{11} - Y_3$, $Y_2 = y_{22} - Y_3$. The y-parameters are obtained from the de-embedded S-parameters using well-known relations [14].

The admittances Y_1 and Y_2 are modelled first. The results show that these admittances are similar. In order to simplify the model, they are modelled as identical components according to the mean value $Y_{1,avg} = (Y_1 + Y_2)/2$. The admittance $Y_{1,avg}$ is shown in Fig. 7a for the 68 pF capacitor measured using SOLT calibration. The capacitive character of this admittance can be observed. The value of $Y_{1,avg}$ admittance is similar for on-board calibration in magnitude and phase. Both Y_1 and Y_2 are modelled as the capacitances C_P. The value of the capacitance can be determined from the imaginary part of $Y_{1,avg}$ by using expression:

$$C_P = \frac{Im\{Y_{1,avg}\}}{\omega} = \frac{Im\{Y_{1,avg}\}}{2\pi f}. \quad (3)$$

The measured value of C_P obtained using Eq. (3) is frequency-dependent, as it is shown in Fig. 8. In order to obtain an optimal value of this parameter in the model, the value of C_P is optimized to minimize the mean square error with respect to the measured value in the frequency range where the frequency dependence of C_P is approximately flat: from 10 MHz to 8 GHz.

The admittance Y_3 is shown in Fig. 7b. The series resonance behaviour can be observed. Therefore, Y_3 is modelled using the series RLC model. Its impedance is equal to:

$$Z_3 = \frac{1}{Y_3} = R_S + j(\omega L_S - \frac{1}{\omega C_N}), \quad (4)$$

where C_N is the nominal capacitance, L_S is the series inductance, and R_S is the equivalent series resistance (ESR). The value of the nominal capacitance is calculated by neglecting the inductive component ωL_S in Eq. (4) for low frequencies:

$$Im\{Z_3\} \approx -\frac{1}{\omega C_N} = -\frac{1}{2\pi f C_N} \quad (5)$$

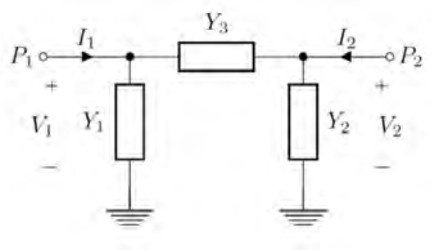


Fig. 6: π-model for SMD components extracted using y-parameters.

The nominal capacitance C_N can be expressed as:

$$C_N = \frac{-1}{2\pi f Im\{Z_3\}}, \quad (6)$$

The value of C_N in the model is optimized in the frequency range from 1 MHz to 10 MHz. Analogously, the value of the series inductance L_S is calculated by neglecting the capacitive component $1/\omega C_N$ in Eq. (4) for high frequencies:

$$Im\{Z_3\} \approx \omega L_S = 2\pi f L_S \ . \quad (7)$$

The value of L_S is equal to:

$$L_S = \frac{Im\{Z_3\}}{2\pi f} \ . \quad (8)$$

The value of L_S in the model is optimized in the frequency range from 1 GHz to 8 GHz. The value of the modelled series resistance R_S is optimized with respect to the real part of Z_3 in Eq. (4). The resistance value is optimised in the frequency range from 10 MHz to 100 MHz.

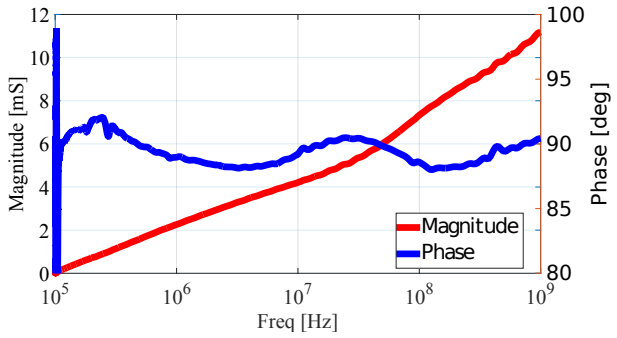
All optimizations are run in ADS using the random optimizer with 100 iterations. The frequency range for the optimization is chosen manually for each element of the model, and for each modelled capacitor.

The final schematic of the SMD capacitor model is shown in Fig. 9. The optimized values of the model elements are summarized in Table IV for the different capacitor values, as well as for both calibration methods. The comparison between the values of the model elements in Fig. 9 obtained by measurements and the optimized values is shown in Fig. 10, and the comparison of the S-parameters is shown in Fig. 11.

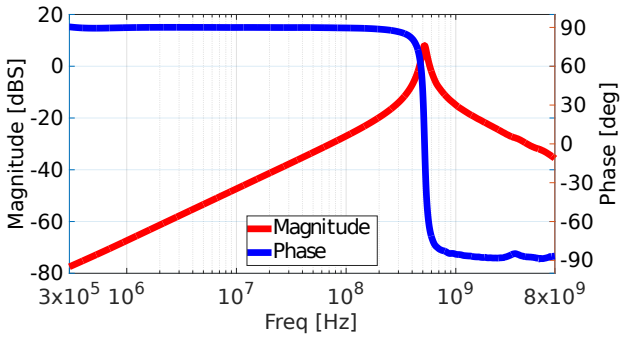
V. Discussion

The measured S-parameters of the test structures can be successfully used to build models of the SMA connector, the CPWG transmission lines and the SMD capacitors. The S-parameters obtained using the different calibration methods (SOLT and on-board), shown in Fig. 5, are consistent for frequencies up to 500 MHz. In the higher frequency range, the general trend of the S-parameters is consistent. The maximum model error in the magnitude of S_{21} is under 1 dB.

The model topology presented in Section IV is built based on the y-parameters obtained from measurements. It can be considered a physical model, since each element



(a) $\mathbf{Y_{1,avg}} = (\mathbf{Y_1} + \mathbf{Y_2})/\mathbf{2}$.



(b) $\mathbf{Y_3}$.

Fig. 7: Values of the admittances in the π-model obtained from the de-embedded S-parameters measured using SOLT calibration for 68 pF capacitor.

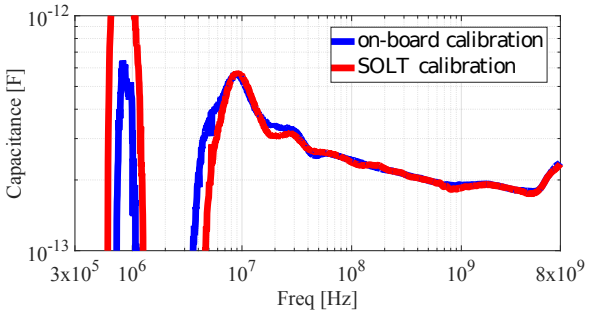


Fig. 8: Extracted value of the capacitance C_P in Eq. (3) for the 1 nF capacitor obtained using both calibration methods.

has a physical interpretation. The capacitors C_P model the parasitic capacitive coupling between the soldering pads and ground planes. The values for all capacitors and calibration methods are around 200 fF. This is in the expected order of magnitude for the SMD-0603 package. The capacitance to ground C_P is similar for both calibrations with the biggest difference between SOLT and on-board calibrations of 20% for the 10 nF capacitor. The element L_S represents the equivalent series inductance and its value is consistent and is between 1.1 nH and 1.6 nH. The element R_S represents the equivalent series resistance, and its value is between 0.1 Ω and 0.3 Ω.

VI. Conclusion

The SMD capacitor models are extracted from the measured S-parameters. The fixture consisting of SMA connectors and CPWG transmission lines is de-embedded by, firstly, standard VNA calibration enhanced by fixture modelling and, secondly, by on-board calibration standards. The two calibration methods are compared. The on-board calibration is consistent with the de-embeded results for frequencies up to 8 GHz. The parasitic element values of the modelled capacitors are similar for different capacitors in the same SMD-0603 package. The extracted models can be used to improve the design time for electronic circuits in high frequency applications.

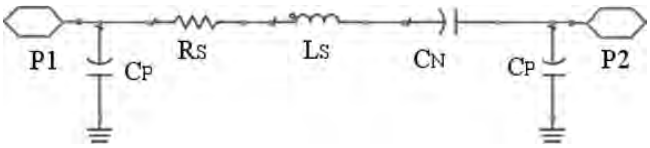


Fig. 9: Model of the SMD capacitors.

TABLE IV: Model components value for SOLT and on-board calibration.

nominal C	$C = 68$ pF		$C = 1$ nF		$C = 10$ nF	
calibration	SOLT	on-board	SOLT	on-board	SOLT	on-board
C_P/fF	283.2	273.5	247.1	202.9	195.1	221.4
L_S/nH	1.144	1.155	1.427	1.417	1.615	1.617
R_S/mΩ	135.9	117.9	157.5	149.2	126.1	127.4
C_N/pF	68.20	68.21	995.6	993.3	9187.4	9181.2
$f_{Y_3 res}$/MHz	517.2	517.2	133.4	134.2	40.65	41.12

Acknowledgment

This research is funded by ams AG, Premstaetten, Austria.

References

[1] P. Muthana, A. E. Engin, M. Swaminathan, R. Tummala, V. Sundaram, B. Wiedenman, D. Amey, K. H. Dietz, and S. Banerji, "Design, Modeling, and Characterization of Embedded Capacitor Networks for Core Decoupling in the Package," *IEEE Trans. Adv. Packag.*, vol. 30, no. 4, pp. 809–822, Nov 2007.

[2] A. Alaeldine, R. Perdriau, M. Ramdani, J. Levant, and M. Drissi, "A Direct Power Injection Model for Immunity Prediction in Integrated Circuits," *IEEE Trans. Electromagn. Compat.*, vol. 50, no. 1, pp. 52–62, Feb 2008.

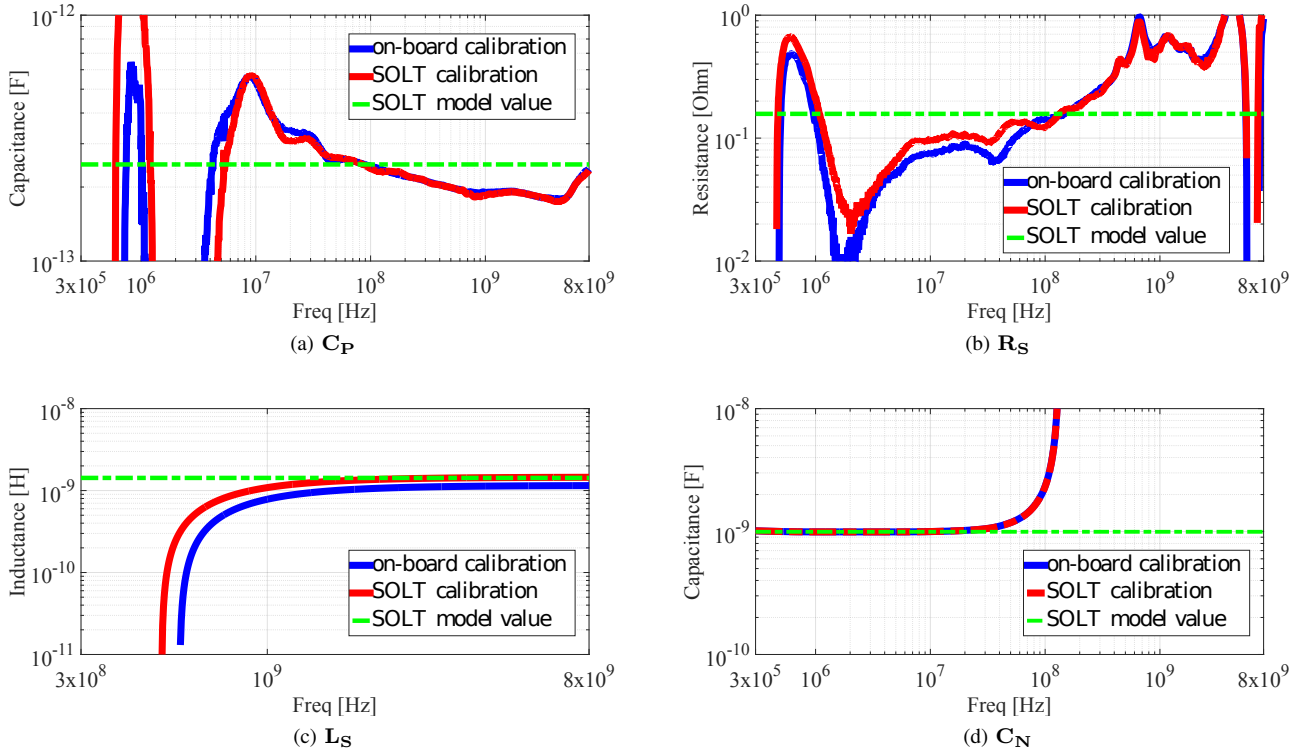
[3] Infineon, *PFC boost converter design guide, Appl. Note*, February 2016.

[4] K. Naishadham, "Experimental Equivalent-Circuit Modeling of SMD Inductors for Printed Circuit Applications," *IEEE Trans. Electromagn. Compat.*, vol. 43, no. 4, pp. 557–565, Nov 2001.

[5] B. Pejcinovic, V. Ceperic, and A. Baric, "Design and Use of FR-4 CBCPW Lines In Test Fixtures for SMD Components," in *ICECS*, Dec 2007, pp. 375–378.

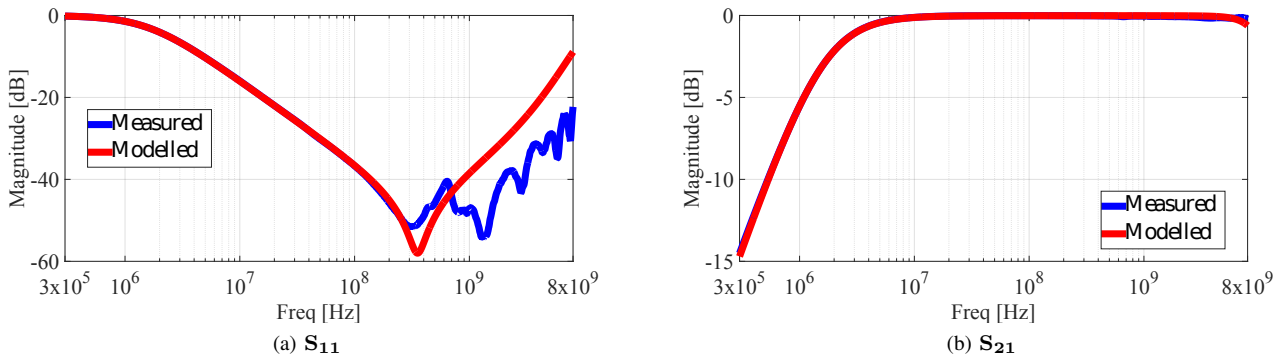
[6] P. R. B. Vitor, M. J. Rosario, and J. C. Freire, "Modelling SMD Capacitor for Microstrip Circuits," in *MELECON*, Apr 1989, pp. 717–720.

[7] K. Technologies, *Using the Time-Domain Reflectometer, Appl. Note*, August 2014.



(a) $\mathbf{C_P}$ (b) $\mathbf{R_S}$ (c) $\mathbf{L_S}$ (d) $\mathbf{C_N}$

Fig. 10: Comparison between the values of the parasitic elements of the 1 nF capacitor obtained using both calibration methods, and the optimised value used in the model.



(a) $\mathbf{S_{11}}$ (b) $\mathbf{S_{21}}$

Fig. 11: Comparison between measured and modelled S-parameters for on-board calibration method for 1 nF capacitor.

[8] Cinch Connectivity. RF coaxial, SMA, straight jack, 50 Ohm. [Online]. Available: cinchconnectivity.com/OA_MEDIA/specs/pi-142-0711-201.pdf

[9] Rohde&Schwartz. (2013) R&S ZVA / R&S ZVB / R&S ZVT Vector Network Analyzers Operating Manual. [Online]. Available: www.fer.unizg.hr/_download/repository/ZVA_ZVB_ZVT_Operating.pdf

[10] K. Technologies, *ADS 2015, Simulation-Analog RF*, 2015.

[11] Wcalc. (2009) Coplanar Waveguide Analysis/Synthesis Calculator. [Online]. Available: http://wcalc.sourceforge.net/cgi-bin/coplanar.cgi

[12] G. Ghione and C. Naldi, "Parameters of coplanar waveguides with lower common planes," *Electronics Letters*, vol. 19, no. 18, pp. 734–735, September 1983.

[13] T. Mandic, R. Gillon, B. Nauwelaers, and A. Baric, "Characterizing the TEM Cell Electric and Magnetic Field Coupling to PCB Transmission Lines," *IEEE Trans. on Electromagn. Compat.*, vol. 54, no. 5, pp. 976–985, Oct 2012.

[14] D. Pozar, *Microwave Engineering*. Wiley, 2004.

Impact of Capacitor Dielectric Type on the Performance of Wireless Power Transfer System

D.Vinko and P. Oršolić
University of Osijek, Faculty of Electrical Engineering, Department of Communications,
Osijek, Croatia
davor.vinko@etfos.hr

Abstract – In loosely coupled wireless power transfer systems, the efficiency is directly affected by the quality factor of the resonant LC tank in both the transmitter and the receiver. This paper studies the impact that the capacitor dielectric type has on the quality factor of the resonant LC tank. Experimental investigation is conducted for low ESR capacitor types and the performance of the wireless power transfer system is evaluated. Focus of the experimental evaluation is placed on the receiver's end, i.e. the evaluated parameters are current-voltage characteristics of the receiver and maximum power values. Capacitors are also compared with respect to their ESR values on different frequencies and price. The results show that the capacitor type has a significant impact on the performance of the wireless power transfer system. The correct choice of capacitor type can increase the efficiency of power transfer and the maximum achievable power up to 400 %.

I. INTRODUCTION

Wireless power transfer (WPT) is lately gaining more and more attention with wide spectrum of possible applications [1]-[3]. Applications differ in amount of transferred power, the impedance of the load circuit, the distance of wireless power transfer, the operating frequency, and these are just some of the parameters that can be altered to optimize power transfer. Analysis and the design of the WPT system in a given application is most often based on power and efficiency maximization [4], [5]. In systems that are utilizing wireless power transfer, a quality factor (Q factor) of an LC resonant tank has a significant impact on system performance [6]-[8].

WPT system consists of a transmitter and a receiver, Fig. 1. In WPT system the power is wirelessly transferred from the transmitter to the receiver through alternating magnetic field. Transmitter is represented by an AC voltage source U which drives resonant tank formed by C_1 and L_1. Receiver is represented with resonant tank formed by L_2 and C_2, and resistive load R_{LOAD}. M represents the mutual inductance of L_1 and L_2. In this paper we focus on a loosely coupled inductive WPT system [9], [10], which is characterized by coupling coefficient k below 0.1 [11].

$$k = \frac{M}{\sqrt{L_1 L_2}} < 0.1 \qquad (1)$$

With such setup, the resonant frequency of the WPT system is not significantly affected by mutual inductance of loosely coupled coils of the transmitter and the receiver. Therefore, with low coupling coefficient k, both LC tanks can be separately designed for the same desired resonant frequency, and the changes in mutual inductance M (due to different physical placement and alignment between coils) will not significantly affect the efficiency of wireless power transfer. Factors that do affect the performance of WPT system are parasitic parameters in resonant tank. Fig. 1 shows these parameters for the resonant tank of the receiver. Inductor L_2 has parasitic resistance R_L and parasitic capacitance C_L, while capacitor C_2 has following parasitic parameters: parallel resistance R_{Leak}, equivalent series inductance ESL and equivalent series resistance ESR. Both ESL and ESR are frequency dependent. Values of parasitic capacitance C_L and equivalent series inductance ESL affect the resonant frequency, while values of parasitic resistances, R_L, R_{Leak} and ESR, affect the quality factor of each component (L_2 and C_2), and consequently the quality factor of the resonant tank. With that in mind, the efficiency of the wireless power transfer is predominantly affected by the quality factor of components used.

Quality factor of the component (L or C) is the ratio of stored energy and dissipated energy, which also corresponds to the ratio of its reactance and resistance at a given frequency f.

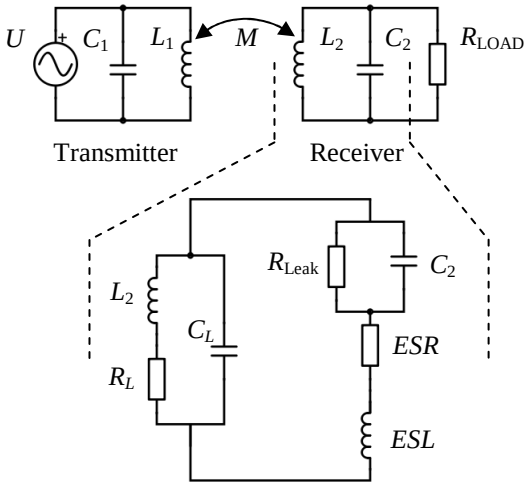


Figure 1. Wireless power transfer system

This work was sponsored by J. J. Strossmayer University of Osijek under project IZIP-2014-104 "Wireless power transfer for underground and underwater sensors".

$$Q_L = \frac{2\pi \cdot f \cdot L}{R_L} \quad (2)$$

$$Q_C = \frac{1}{2\pi \cdot f \cdot C \cdot ESR} \quad (3)$$

For LC tank, the Q factor represents the ratio of stored and dissipated energy in each period (cycle) of resonant frequency. Dissipated energy increases with increase of parasitic resistances (R_L, R_{Leak} and *ESR*), resulting with lower Q factor of the resonant tank, which decreases the efficiency of wireless power transfer.

Papers that address the impact of capacitor selection on system performance [12], [13] identify the *ESR* as the main source of losses. In WPT system (Fig. 1), the influence of R_{Leak} on system performance is negligible in comparison to the influence of *ESR*, which is why R_{Leak} is not included in (3).

II. *ESR* VS. CAPACITOR DIELECTRIC TYPE

Parasitic parameters of a capacitor (R_{Leak}, *ESL* and *ESR*) depend on a type of the dielectric used. In this paper we investigate the impact of the *ESR* on the performance of the WPT system. In order to be used in LC tank of a WPT system capacitor must be non-polarized. Most common non-polarized capacitor types are FILM and ceramic capacitors and both types are characterized as low-*ESR* capacitor types. Table I gives the list of all capacitors used in this paper. All capacitors are 10 nF capacitors with voltage rating of 100 V. Three capacitor types are evaluated: Ceramic, MLCC (Multi-Layer Ceramic Capacitor) and FILM capacitors. Besides the capacitor type, dielectric type is stated for each capacitor. Three FILM capacitors with same dielectric type are evaluated (dielectric types marked PET1, PET2 and PET3 in Table I). They differ in manufacturer, which leads to significant difference in unit price which is also given for each capacitor.

ESR value for each capacitor is measured using handheld LCR meter UT612, which can measure *ESR* value on predefined frequencies up to 100 kHz. These values are given in Table I and also shown on Fig. 2. It is important to note than *ESR* varies with frequency, and that this change is not linear and it varies with type of dielectric.

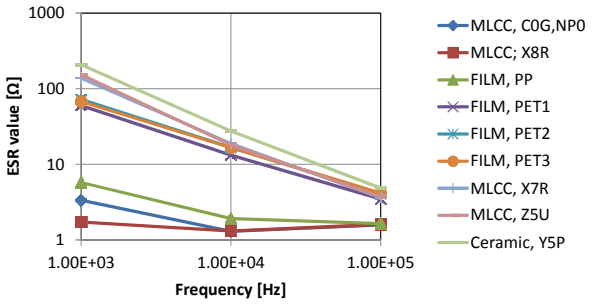


Figure 2. *ESR*-frequency characteristics of different capacitor types

Manufacturers commonly supply a single *ESR* value at frequency which differs from manufacturer to manufacturer. These common frequencies range from 1 kHz up to 100 MHz. With frequency dependent *ESR*, a provided single *ESR* value does not give the complete insight in capacitor performance.

III. MEASUREMENTS

During measurements, evaluated capacitors were placed (one at a time) in LC tank, parallel to a 100 µH air coil. Due to tolerance in capacitance value, the resonant frequency f_R of LC tank varied for each capacitor, and is given in Table II. The maximum deviation of resonant frequency is within 10%, which is the maximum tolerance of tested capacitors.

In WPT system (Fig. 1), both LC tanks (L_1, C_1 and L_2, C_2) must be adjusted to the same resonant frequency. The resonant frequency of receivers LC tank (L_2, C_2) changes for different capacitor types. To avoid the necessity of adjusting the resonant frequency of transmitters LC tank (L_1, C_1 in Fig. 1) for each tested capacitor, a measurement setup shown in Fig. 3 is used. LC tank is not used in transmitter, but only a transmitting coil L_1. This also

TABLE I. COMPARISON OF USED CAPACITORS

C = 10 nF; 100 V			*ESR* [Ω]		
Capacitor type	Dielectric type	Price [€]	@1 kHz	@10 kHz	@100 kHz
MLCC	C0G,NP0	1.48	3.36	1.3	1.58
MLCC	X8R	0.82	1.73	1.32	1.60
FILM	PP	0.52	5.75	1.92	1.64
FILM	PET1	2.55	60.1	13.24	3.47
FILM	PET2	0.76	71.8	16.7	4.09
FILM	PET3	0.11	67.7	16.62	4.13
MLCC	X7R	0.49	137.6	18.9	3.61
MLCC	Z5U	0.54	152.9	17.97	3.76
Ceramic	Y5P	0.29	207	27.6	4.86

TABLE II. MAXIMUM POWER POINT ANALYSIS

Dielectric type	R_{RES}=R_{LOAD} @MPP [Ω]	f_R [kHz]*	R_{LC} [Ω]	*ESR* @100 kHz [Ω] **	Q_{LC} @MPP
C0G,NP0	6500	159.9	1.55	1.58	64,7
X8R	6000	161.6	1.72	1.60	59,1
PP	6000	161.2	1.71	1.64	59,2
PET1	4000	163.9	2.65	3.47	38,8
PET2	3500	163.7	3.02	4.09	34,0
PET3	3500	165.9	3.10	4.13	33,6
X7R	2000	161.0	5.12	3.61	19,8
Z5U	2000	167.0	5.51	3.76	19,1
Y5P	1000	163.4	10.54	4.86	9,7

* Resonant frequency f_R is measured for parallel LC tank with L = 100 µH

** *ESR* values correspond to values given in Table I

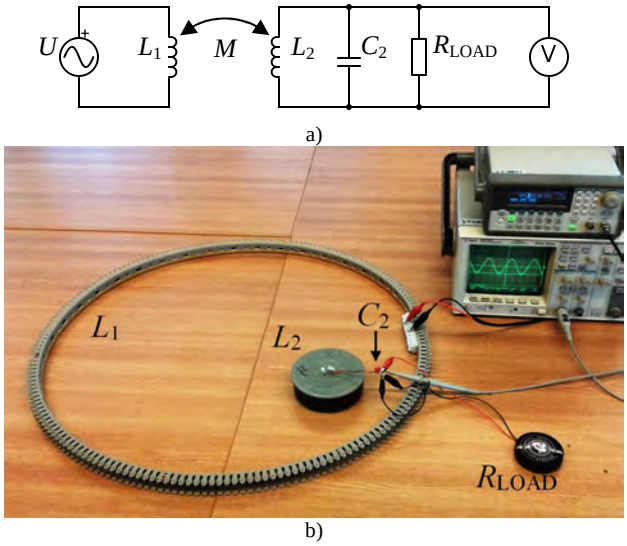


Figure 3. Measurement setup a) schematic and b) photograph

reduces the impact of mismatch in resonant frequencies of the transmitter and the receiver on measurement results.

The following component values are used for measurement: L_1 = 180 µH (single layer air coil with diameter of 60 cm, 10 turns of 1.6 mm wire), L_2 = 100 µH (single layer air coil with diameter of 11 cm, 25 turns of 1 mm wire), C_2 = 10 nF (different capacitor types as given in Table I), R_{LOAD} value is varied from 100 Ω to 1 MΩ in 13 discreet steps (100 Ω, 220 Ω, 560 Ω, 1 kΩ, 2.2 kΩ, 4.7 kΩ, 10 kΩ, 22 kΩ, 47 kΩ, 100 kΩ, 220 kΩ, 470 kΩ, 1 MΩ). As a voltage source *U*, Agilent 33250A arbitrary waveform generator is used, generating sine voltage waveform with peak-to-peak amplitude of 20 V. Frequency of the voltage source *U* is manually adjusted for each tested capacitor and it corresponds to resonant frequency f_R values given in Table I. Voltage across R_{LOAD} is measured using oscilloscope (represented as voltmeter in Fig. 3). Coils L_1 and L_2 are placed on the same plane, with coupling coefficient *k* under 0.1, which results with mutual inductance *M* under 14 µH.

A current-voltage output characteristic of the WPT receiver is measured for each tested capacitor, Fig. 4. Result show that with different dielectric type, a maximum output voltage of the WPT receiver varies from 25 V to 105 V. The variations of maximum output current are less prominent, ranging from 17 mA to 20 mA. There are three capacitor types that performed significantly better the rest, MLCC capacitors with C0G, NP0 and X8R

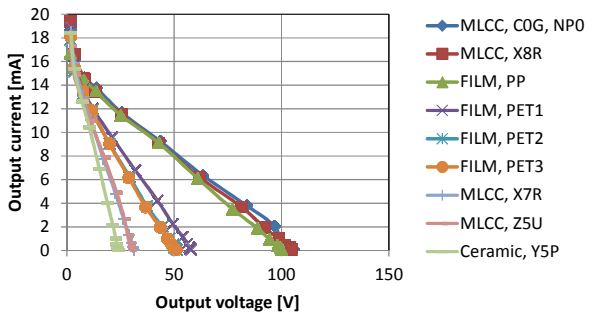


Figure 4. Current-voltage output characteristics of a WPT receiver for different capacitor types

dielectric type, and FILM capacitor with PP dielectric type. When compared to *ESR* measurement (Fig. 2), the same capacitors have the lowest *ESR* values on all tested frequencies.

Important parameter in WPT systems is a maximum instantaneous power that can be supplied by the receiver. Fig. 5 shows output power of a WPT receiver for different values of load resistance R_{LOAD} and different capacitor types. Highest output power is obtained for the same three capacitor types that have the lowest *ESR* values.

For each tested capacitor, the maximum power is available for a different value of load resistance. At maximum power point the load resistance is matched to the resistance of LC tank. At resonant frequency, the resistance of LC tank (R_{RES}) can be expressed as:

$$R_{RES} = R_{LOAD}@\text{MPP} = \omega L \times Q_{LC} = \frac{Q_{LC}}{\omega C}, \tag{4}$$

$$Q_{LC} = \frac{\omega L}{R_{LC}} = \frac{1}{\omega C \times R_{LC}}, \tag{5}$$

where Q_{LC} represents the quality factor of the resonant tank, R_{LC} represents parasitic resistance of LC tank (R_L + *ESR*), and ω is the angular frequency $2\pi f$. Since capacitor and inductor have equal reactance at resonant frequency, two equivalent expressions are given in (4) and (5).

To extract resistive loses of LC tank from measurement results, the following expression is derived from (4) and (5):

$$R_{LC} = \frac{(\omega L)^2}{R_{RES}} = \frac{1}{(\omega C)^2 \times R_{RES}} \tag{6}$$

The total parasitic resistance R_{LC} can be calculated by using measured values for resonant frequency f_R, and the resistance of LC tank R_{RES} is obtained experimentally by matching load resistance at maximum power point (4). Output power plots for each capacitor (Fig. 5) are drawn using 13 measurement points that do not necessarily correspond to the maximum power point. Therefore the values for R_{RES} given in Table II are extracted from output power plots (Fig. 5) and are best approximation of the

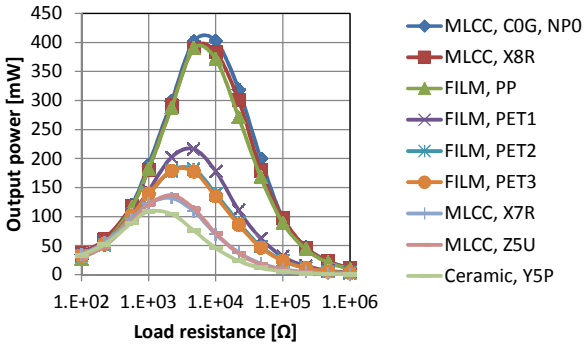


Figure 5. Output power of a WPT receiver for different capacitor types

maximum power point. Calculated values for R_{LC} are given in Table II.

Air coil used in LC tank has a DC resistance of 0.2 Ω. At resonant frequency of the LC tank, due to the skin effect, the coil resistance R_L increases. Calculated value of the coil resistance R_L, at resonant frequency, equals approx. 0.33 Ω. When compared with R_{LC} values given in Table II, the coil resistance makes from 3 % up to 20 % of the total parasitic resistance R_{LC}. The remaining more than 80% of R_{LC} is due to capacitor losses.

ESR values measured at 100 kHz are the closest to the resonant frequency (163 kHz ± 2.5%) and are also given in Table II. It can be noted that there is a good correlation between *ESR* values at 100 kHz and total parasitic resistance R_{LC}. Some discrepancies that occur can be attributed to the non-linear behavior of the *ESR*. Quality factor of resonant LC tank Q_{LC} is calculated (5) and the values are given in Table II.

IV. CONLUSION

Conducted experimental investigation has shown that the choice of capacitor type has a significant impact on performance of wireless power transfer system. Equivalent series resistance (*ESR*) of capacitor was identified as dominant factor which determines the performance of WPT system. Three capacitor types were tested: Ceramic, MLCC (Multi-Layer Ceramic Capacitor) and FILM capacitors, all considered to be capacitors with low *ESR* value. Measurements also showed that *ESR* value is more affected by the dielectric material than the capacitor type.

Best results are achieved by using C0G, NP0 and X8R dielectric for MLCC capacitors, and PP dielectric for the FILM capacitors. With use of those dielectric types system performance is improved by 400 %, with respect to Ceramic capacitor with Y5P dielectric, and the quality factor of resonant LC tank shows six-fold improvement. All tested capacitors were in price range from 0.11 to 2.55 €, and the top three range from 0.52 to 1.48 €. The best performance to price ratio is obtained for FILM capacitor with PP dielectric.

REFERENCES

[1] Y. Yang, X. Xie, G. Li, Y. Huang, Z. Wang, "A Combined Transmitting Coil Design for High Efficiency WPT of Endoscopic Capsule," IEEE International Symposium on Circuits and Systems (ISCAS), 2015, pp. 97 – 100.

[2] D. Futagami, Y. Sawahara, T. Ishizaki, I. Awai, "Study of high efficiency WPT underseas," IEEE Wireless Power Transfer Conference (WPTC), 2015, pp. 1 – 4.

[3] L. Olvitz, D. Vinko, T. Švedek, "Wireless Power Transfer for Mobile Phone Charging Device," Proceedings of the 35th International Convention MIPRO, 2012, pp. 141 – 145.

[4] G. Kim, B. Lee, "Analysis of Magnetically Coupled Wireless Power Transfer between Two Resonators Based on Power Conservation," IEEE Wireless Power Transfer Conference (WPTC), 2014, pp. 231 – 234.

[5] Y.-K. Jung, B. Lee, "Design of Adaptive Optimal Load Circuit for Maximum Wireless Power Transfer Efficiency," Asia-Pacific Microwave Conference Proceedings (APMC), 2013, pp. 1221 – 1223.

[6] O. Jonah, S. V. Georgakopoulos, M. M. Tentzeris, "Optimal Design Parameters for Wireless Power Transfer by Resonance Magnetic," IEEE Antennas and Wireless Propagation Letters, Vol. 11, pp. 1390-1393, 2012.

[7] I. Awai, T. Ishizako, "Transferred Power and Efficiency of a Coupled-resonator WPT system," IEEE MTT-S International Microwave Workshop Series on Innovative Wireless Power Transmission: Technologies, Systems, and Applications (IMWS), 2012, pp. 105 – 108.

[8] M. Dionigi, M. Mongiardo, "Coaxial Capacitor Loop Resonantor for Wireless Power Transfer Systems," The 7th German Microwave Conference (GeMiC), 2012, pp. 1 – 4.

[9] F. Lu, H. Zhang, H. Hofmann, C. Mi, "A High Efficiency 3.3 kW Loosely-Coupled Wireless Power Transfer System Without Magnetic Material," IEE Energy Conversion Congress and Exposition (ECCE), 2015, pp. 2282 – 2286.

[10] J. A. Russer, P. Russer, "Design Considerations for a Moving Field Inductive Power Transfer System," IEEE Wireless Power Transfer (WPT) Conference, 2013, pp. 147 – 150.

[11] K. A. Grajski, R. Tseng, C. Wheatley, "Loosely-Coupled Wireless Power Transfer: Physics, Circuits, Standards," Proceedings of IMWS-IWPT, 2012, pp. 9 – 14.

[12] S. Karys, "Selection of Resonant Circuit Elements for the ARCP Inverter," 10th International Conference on Electrical Power Quality and Utilisation EPQU, 2009, pp. 1 – 6.

[13] H. J. H. Since, S. Taninder S., B. J. Kai, "The Impact of Capacitors Selection and Placement to the ESL and ESR," International Symposium on Electronics Materials and Packaging EMAP, 2005, pp. 258 – 261.

Switching Speed and Stress Analysis for Fixed-fixed Beam Based Shunt Capacitive RF MEMS Switches

Anoushka Kumar A, Resmi R
LBS Institute of Technology for Women, Thiruvananthapuram, India
anoushkakumar4@gmail.com, resmilbs@gmail.com

***Abstract*—In this paper, effect of different materials on the reliability of RF MEMS shunt capacitive switch is analyzed. The effective von-Mises stress analysis is done on a fixed-fixed beam switch for materials like Titanium, Platinum, Gold, Aluminium and Copper. The maximum value of von-Mises stress obtained for each material was less than their corresponding ultimate tensile strength. Membrane using Titanium, Copper and Platinum can withstand more number of switching cycles. The variation in switching speed of a fixed-fixed beam structure using Aluminium, Platinum, Titanium, Copper and Gold is also analysed. Membranes using Titanium and Copper give better performance with respect to switching speed and reliability.**

***Keywords*—Fixed-fixed beam; von-Mises stress; Tensile strength**

I. Introduction

Radio Frequency Micro Electro Mechanical System (RF-MEMS) switches have been an attractive field for both scientific research and industry due to its promising applications in RADAR systems, Satellite communication systems, Wireless communication systems and Instrumentation systems. Compared to traditional GaAs FET and p–i–n diode switches RF MEMS switches have negligible power (few μ-watts) consumption, low insertion loss, high isolation, much lower intermodulation distortion, low cost and light weight[1]. RF MEMS switches replaced the traditional GaAs FET and p–i–n diode switches in RF and microwave systems. Typically MEMS switches are manufactured using surface micro-machining processes.

In RF-MEMS switches, the mechanical movement of switch membrane which can be either a fixed–fixed beam or a cantilever beam creates a short circuit or an open circuit in transmission line. This mechanical movement is achieved using electrostatic, piezoelectric, magnetostatic or thermal actuation. Even though electrostatic method requires a high actuation voltage it is the most prevalent one due to its near zero power consumption, small electrode size, thin layers and short switching time. In electrostatic actuation, electrostatic force is generated between fixed electrode and movable membrane for switching operation. The main limitations of RF MEMS switches include slow switching speed, high actuation voltage and hot switching. Fig. 1 shows the classification of RF MEMS switches.

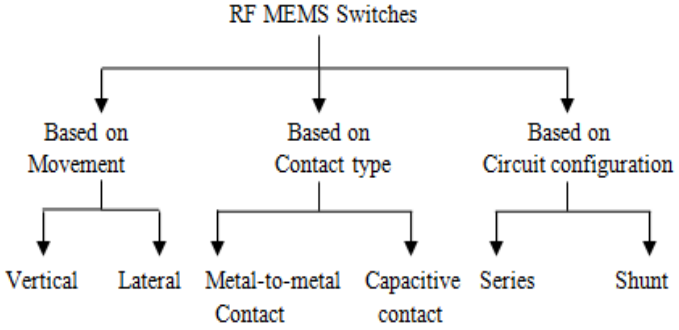


Figure 1. Classification of RF MEMS switches

Fig. 2 shows the shunt and series configurations of RF MEMS switches [2].

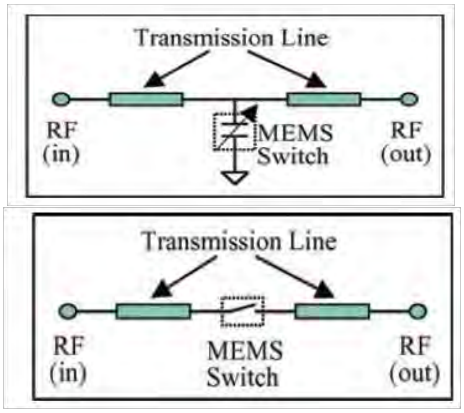


Figure 2. Shunt and series circuit configuration of MEMS switches

Various designs are existing on RF MEMS switches to tackle the limitations faced by switches. Stiff ribs around the membrane helps to reduce stiction and buckling effect in switches [3]. Meander in membrane reduces the required actuation voltage of a RF MEMS switch [4]. For highly reliable operation twin layered membrane is preferred [5]. Holes in switch membrane help to reduce actuation voltage and it also reduces squeeze film air damping [6].

II. SHUNT CAPACITIVE SWITCH

A. General Structure

RF MEMS shunt capacitive switch generally consists of a movable metal bridge, suspended at a height ‚g_0' above the center conductor. The dielectric layer is used above the center conductor, so that the switch membrane does not come into contact with the centre electrode during the actuation. Fig. 3 and Fig. 4 show the switch in OFF and ON states respectively [7].

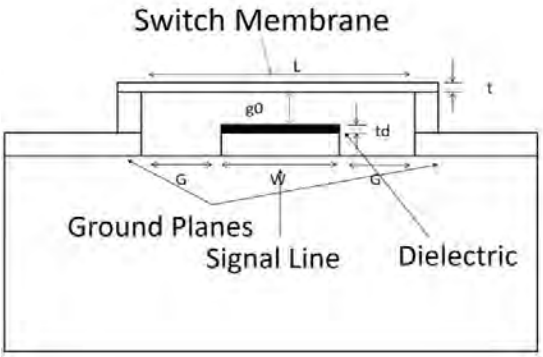


Figure 3. Cross sectional view of RF MEMS Shunt Switch (OFF)

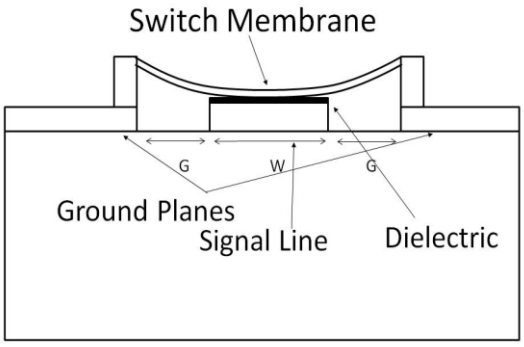


Figure 4. Cross sectional view of actuated MEMS Shunt switch (ON)

When a dc voltage is applied, an electrostatic force is produced between the beam and centre conductor. When this voltage exceeds pull-in voltage, the electrostatic forces overcome elastic recovery forces and pulls down the membrane over dielectric layer on signal line. It results in the formation of an open circuit between transmission lines and RF signal which prohibits the transmission of the signal. The pull-in voltage of fixed-fixed beam switch is given by

$$V_P = \sqrt{\frac{8k}{27\varepsilon_0 A} g_0^3} \quad (1)$$

Here k is the spring constant in N/m, g_0 is the initial gap in μm, ε_0 is the free space permittivity and A is the electrostatic area in μm^2. Fig. 5 shows the equivalent circuit model of shunt capacitive switch [1]. The sections of transmission line are of length *l*+*w*/2, where *l* is the distance from reference plane to the edge of membrane and *w* is the CPW centre conductor width.

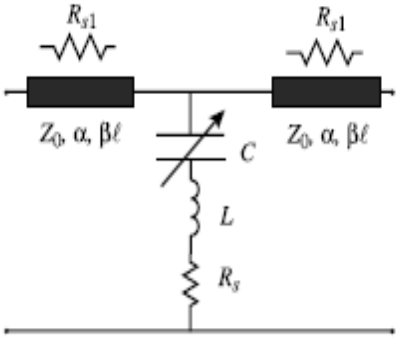


Figure 5. Equivalent C–L–R circuit model of shunt capacitive switch

B. Device Geometry

Fixed-fixed beam is usually preferred for the beam structure in MEMS switches because it provides better stability and much lower sensitivity to stress. Fig. 6 shows the 3D structure of a fixed-fixed beam switch modeled in COMSOL Multiphysics. The switch consists of a square plate which is suspended above a thin film of Silicon nitride (relative dielectric constant 7.5). Silicon nitride is the commonly used dielectric layer in shunt capacitive switches. There is a silicon counter-electrode below the substrate which is grounded. Four rectangular flexures are used at the corners to anchor the plate to the substrate. The switch dimensions and material parameters used for simulating the switch membrane are shown in Table I and Table II.

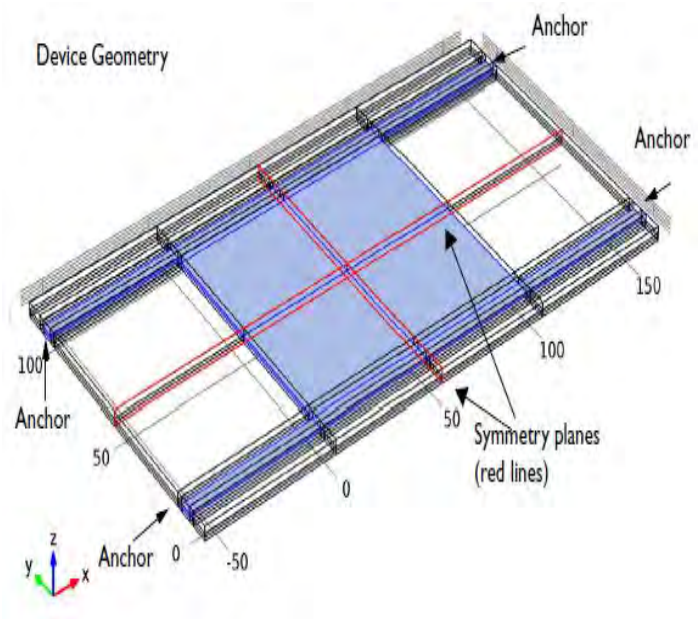


Figure 6. Modeled 3D structure of fixed-fixed beam switch in COMSOL Multiphysics

TABLE I. SWITCH SPECIFICATIONS

Component	Length(μm)	Width(μm)	Depth(μm)
Membrane frame	220	100	1
Contact area	100	100	1
Flexures	60	5	1
Dielectric layer	220	100	0.1
Gap	-	-	0.9

III. STRESS ANALYSIS USING FEM

The von-Mises stress analysis is mandatory to check whether the membrane will withstand a given load condition. If the maximum stress increases for a given gap height from the stress a material can withstand, it indicates the failure of design[8]. The effective von-Mises stress in membrane for the required force at specific spring constant is analyzed for some commonly available metals like Titanium, Platinum, Gold, Aluminium and Copper. Simulation of stress gradient is done using Finite Element Method in COMSOL Multiphysics.

Fig. 7 to Fig. 11 demonstrates the stress distribution in membrane using different materials. From Fig.7 to Fig.11, the colour variation across the surface of membrane of switch indicates the corresponding stress as provided in the colour chart at the right side of the plot. von-Mises stress profile obtained is different for different materials. The maximum stress can be seen along the flexures and minimum stress is obtained in the contact area.

TABLE II. SWITCH PARAMETERS

Material	Young's Modulus(GPa)	Poisson's ratio	Density(kg/m^3)
Copper	110	0.35	8700
Titanium	113	0.29	4540
Gold	79	0.44	19320
Aluminium	70	0.32	2700
Platinum	170	0.39	21450

Fig. 7 shows the maximum and minimum value of von-Mises stress obtained in membrane using Copper.

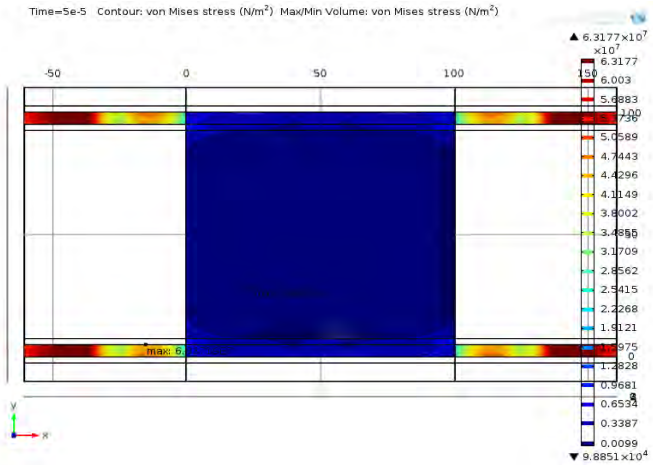


Figure 7. von-Mises stress demonstrating maximum stress of 68MPa in Copper membrane

Fig. 8 shows the maximum and minimum value of von-Mises stress obtained in membrane using Titanium

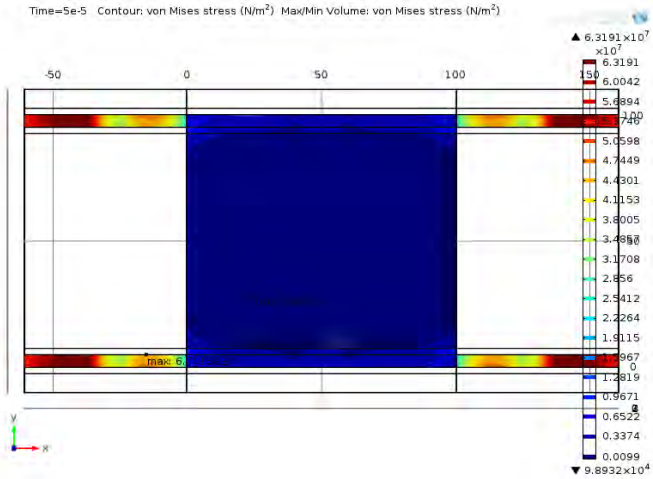


Figure 8. von-Mises stress demonstrating maximum stress of 68MPa in Titanium membrane

Fig. 9 shows the maximum and minimum value of von-Mises stress obtained in membrane using Gold

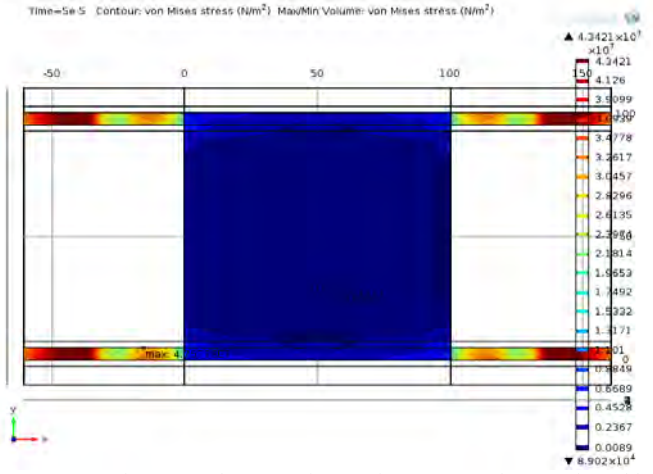


Figure 9. von-Mises stress demonstrating maximum stress of 48MPa in Gold membrane

Fig. 10 shows the maximum and minimum value of von-Mises stress obtained in membrane using Aluminium

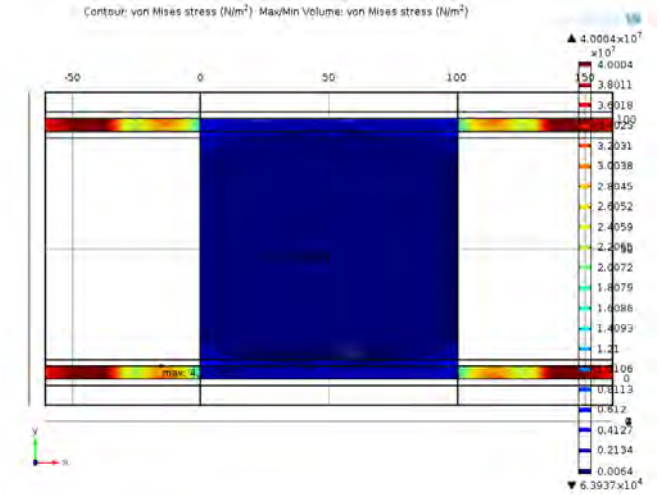


Figure 10. von-Mises stress demonstrating maximum stress of 43MPa in Aluminium membrane

Fig. 11 shows the maximum and minimum value of von-Mises stress obtained in membrane using Platinum

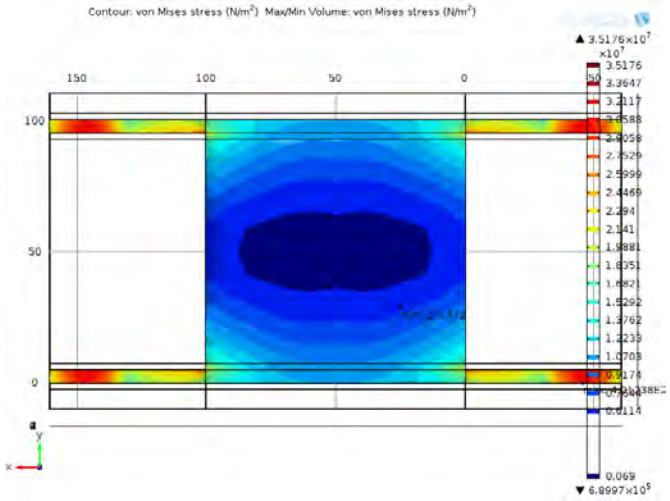


Figure 11. von-Mises stress demonstrating maximum stress of 40MPa in Platinum membrane

TABLE III. COMPARISON OF STRESS ANALYSIS

Material	Von-Mises stress Max (MPa)	Von-Mises stress Min (MPa)	Ultimate tensile strength (MPa)
Copper	68	0.59	210
Titanium	68	0.58	240-370
Gold	48	0.39	100
Aluminium	43	0.36	40-50
Platinum	40	0.27	200-300

The maximum and minimum value of von-Mises stress obtained for membrane using different materials along with their corresponding ultimate tensile strength are shown in Table III. Ultimate tensile strength is the value of maximum stress that a material can withstand while being stretched or pulled before breaking. From Table III it is clear that the maximum von-Mises stress obtained for Titanium, Copper, Gold, Aluminium and Platinum are less than their corresponding ultimate tensile strength. It indicates that design of membrane using these materials is safe for switching operation. Moreover, the maximum von-Mises stress obtained for Titanium, Copper and Platinum were much less than their corresponding ultimate tensile strength. Therefore membrane using these materials can be used for much more switching cycles than the membrane made using Gold or Aluminium.

IV. PULL-IN TIME ANALYSIS

One of the important parameters to be considered in the design of RF MEMS switches is the switching rate. Switching rate, also referred to as switching speed is the time required for the switch to respond at the output due to the change in control voltage. When a voltage is applied, electrostatic force build up and it tends to pull down the membrane and when this electrostatic force exceeds elastic recovery force of membrane, pull in occurs[9]. The pull-in time required by membrane using different materials is simulated on a fixed-fixed beam based shunt capacitive switch. As shown in Fig. 12, the membrane using Platinum takes more than 50µs for a displacement of 0.65µm.

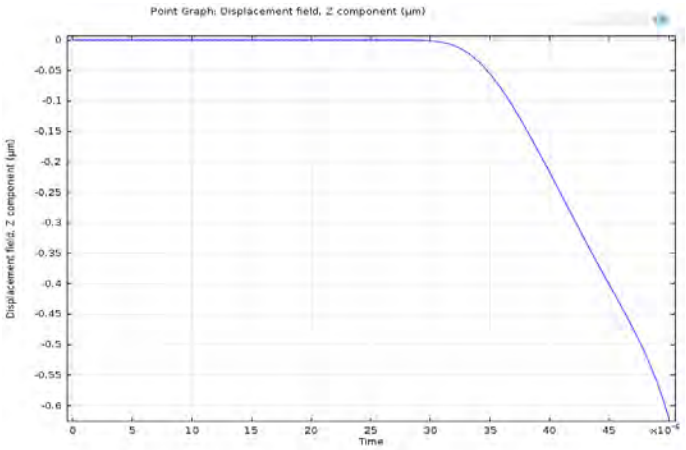


Figure 12. Displacement of the center of Platinum membrane in the modeled switch

When membrane using Copper is used, the time required for pull-in was obtained as 41µs for a displacement of 0.9µm. Fig. 13 shows the switching time required by membrane using Copper.

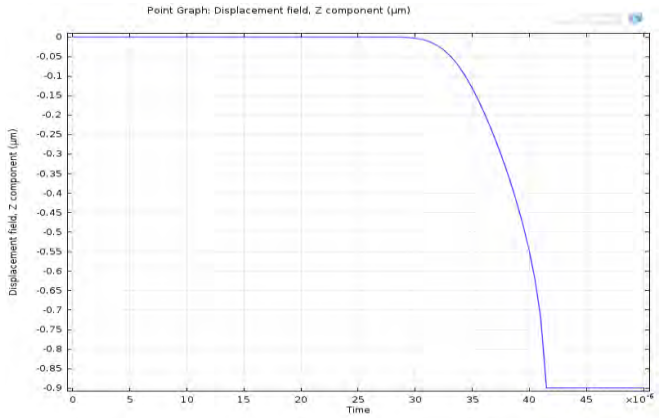


Figure 13. Displacement of the center of Copper membrane in the modeled switch

When membrane using Aluminium is used, the time required for pull-in was obtained as 37µs for a displacement of 0.9µm. Fig. 14 shows the switching time required by membrane using Aluminium.

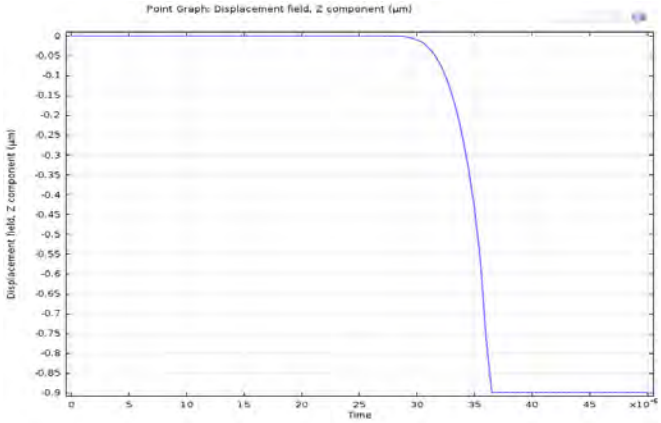


Figure 14. Displacement of the center of Aluminium membrane in the modeled switch

When membrane using Gold is used, the time required for pull-in was obtained as 45µs for a displacement of 0.9µm. Fig. 15 shows the switching time required by membrane using Gold.

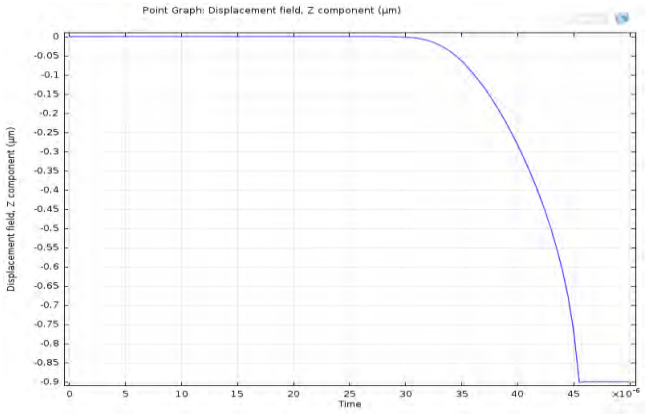


Figure 15. Displacement of the center of Gold membrane in the modeled switch

When membrane using Titanium is used, the time required for pull-in was obtained as 39µs for a displacement of 0.9µm. Fig. 16 shows the switching time required by membrane using Titanium.

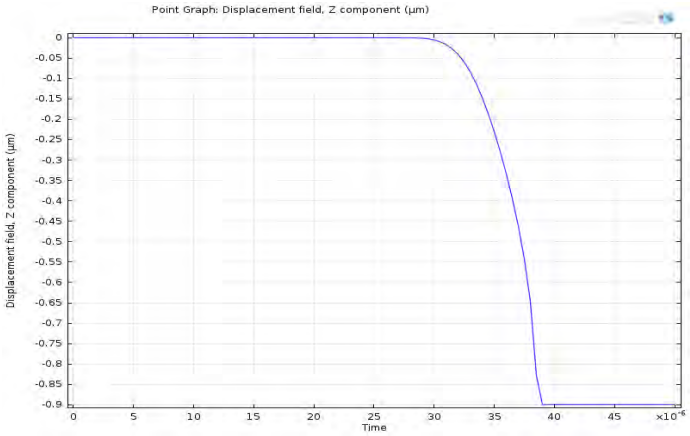


Figure 16. Displacement of the center of Titanium membrane in the modeled Switch

TABLE IV. SWITCHING TIME REQUIREMENT FOR DIFFERENT MATERIALS

Material used	Switching time(µs)
Titanium	39
Gold	45
Aluminium	37
Copper	41
Platinum	>50

From Table IV it is clear that among the five materials used for simulation of membrane, Platinum gives the worst switching speed. The switching time required by membrane using Titanium, Copper and Aluminium is less compared to membrane using Platinum and Gold. Thus for better switching speed conditions, Platinum and Gold are not a good choice.

V. CONCLUSION

Stress analysis is done for a fixed-fixed beam shunt capacitive RF MEMS switch for materials like Titanium, Platinum, Gold, Aluminum and Copper. Membranes using Titanium, Copper and Platinum can withstand much more switching cycles than the membrane made using gold or aluminium. Thus membranes using Titanium, Copper and Platinum are more reliable compared to membranes using Gold and Aluminium. Platinum and Gold are not a good choice for better switching speed conditions. Membranes using Titanium and Copper give better performance with respect to switching speed and reliability. When both reliability and switching speed is considered simultaneously, membranes made up of Titanium is found to be a better choice.

VI. REFERENCES

[1] Gabriel M Rebeiz, "RFMEMS Theory, Design and Technology", John Wiley and Sons Limited, New Jersey, 2002

[2] J Jason Yao, "Topical review RF MEMS from a device perspective", Journal of Micromechanics and Microengineering, 10 (2000), R9–R38

[3] Goldsmith CL, Lin T, Powers B, Wu W, Norvell B, " Micromechanical membrane switches for microwave applications", IEEE international MTT-S symposium digest, vol 1, pp 91–94,1995

[4] Muldavin, J.B, Rebeiz, G.M, "High-isolation CPW MEMS shunt switches, part 1: modelling", IEEE Trans. Microw. Theory Tech., pp1045–1052 , 2000

[5] Singh, T, Khaira, N, Sengar, J, "Stress analysis using multiphysics environment of a novel RF MEMS shunt switch designed on quartz substrate for low voltage applications",Trans Electr Electron. Mater,vol 2,2013

[6] Singh A, Dhanjr, "Design and modeling of a robust wideband Poly-Si and Au based capacitive RF MEMS switch for millimeter wave applications", Proc 2nd Int Conf Comput Sci, vol 1,Elsevier, GmbH, PB, pp 108–114,2014

[7] Bhadri, Baiju, "Electromagnetic Analysis of RF MEMS Switch", International Journal of Engineering Research & Technology, Vol. 3 Issue 9, September 2014

[8] Tejinder Singh, "Computation of Beam Stress and RF Performance of a Thin Film Based Q-Band Optimized RF MEMS Switch", Transactions on Electrical andEelectronic materials, vol. 16, no. 4, pp. 173-178, august 25, 2015

[9] S. Shekhar, 2 K. J. Vinoy, 3 G. K. Ananthasuresh, "Switching and Release Time Analysis of Electrostatically Actuated Capacitive RF MEMS Switches", Sensors & Transducers Journal, Vol. 130, Issue 7, pp 77-90, July 2011.

Performance Analysis of Micromirrors - Lift-off and von Mises stress

Sharon Finny, Resmi R

LBS Institute of Technology for Women, Thiruvananthapuram, India

sharon.finny2010@gmail.com, resmilbs@gmail.com

***Abstract*—Micromirrors are used in numerous applications such as optical switching, projection displays, biomedical imaging and adaptive optics. In this paper, the structural mechanical properties of an electrostatically controlled square shaped micromirror structure are studied. Lift-off analysis and stress analysis is also conducted and materials which can boost the performance of the micromirror are identified. Higher lift-off is observed for silicon-aluminium structure, but it is not recommended due to severe edge displacement and surface deformation. . Structural steel-aluminium material combination shows maximum lift-off with minimal surface deformation and the maximum value of von Mises stress obtained is less than the yield strength of the material thus ensuring safe operation of the structure. The modeling and analysis are done using COMSOL Multiphysics 4.3b.**

Keywords—micromirror; lift-off; von Mises stress;

I. INTRODUCTION

The progress of MEMS technology has led to the development of miniaturized optical devices which has a huge impact on a large number of optical applications. These include movable and tunable mirrors, lenses, filters and other optical structures [1]. Micromirrors are a much- researched area in Micro-Opto-Electro-Mechanical Systems (MOEMS). The movable micromirrors are micro optical components used for spatial manipulation of light. The incident light can be reflected to an expected direction by moving the mirror plate so as to modulate phase and/or the amplitude of the incident light [2]. Micromirrors have been used in a wide variety of applications including optical switching, projection displays, endoscopic imaging, barcode readers, laser beam steering etc [3].

Micromirrors are commonly classified on the basis of their method of actuation. The most common actuation methods include electrostatic actuation, electrothermal actuation, electromagnetic actuation and piezoelectric actuation. Electrostatic method of actuation is preferred over the other methods because of its low power consumption, faster response and simplicity of implementation. One of the drawbacks of electrostatic actuated mirrors is that they exhibit pull in, hysteresis and smaller vertical displacement range [4]. Prestressed actuators have been developed to address the issue of pull in, hysteresis and to obtain large vertical displacement [5].These actuators are operated electrostatically and their electromechanical behavior is influenced by the residual stress developed during the fabrication process [6]. The plating process usually controls this stress and it can be either compressive or tensile. The prestress level is normally set depending on the lift-off required.

II. MODELING OF MICROMIRRORS

The micromirror is modeled using COMSOL Multiphysics software. The geometry consists of a centre mirror plate surrounded by cantilever springs on each of the four sides. The 2D geometry of the micromirror is shown in Fig.1. Appropriate boundary conditions are selected, and then meshing is performed on the model to obtain final refined mesh with hexahedral elements. The meshed geometry is shown in Fig. 2.

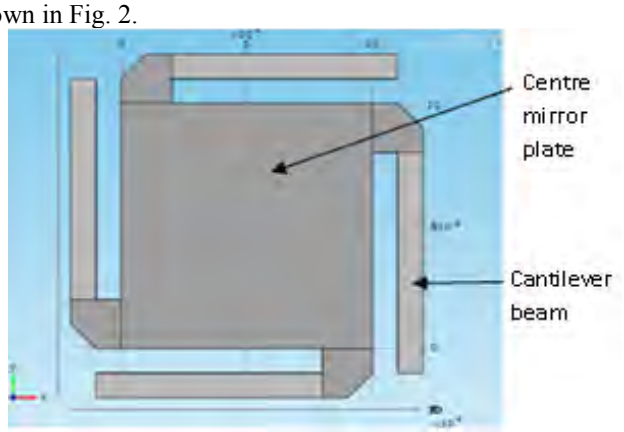


Figure 1. 2D Geometry of Micromirror.

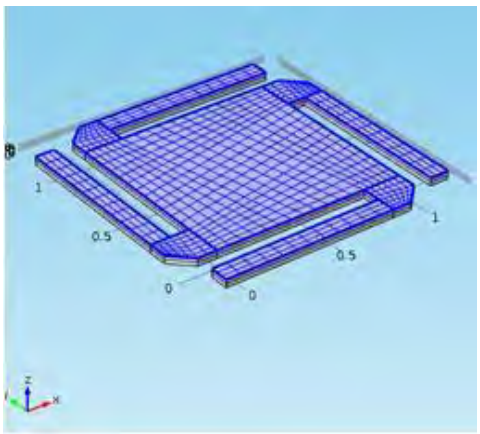


Figure 2. 3D Meshed Geometry of the micromirror.

III. Simulation Results

A. *Lift-off Analysis*

The lift-off analysis was performed in COMSOL Multiphysics to examine the effect of changing material combinations on the lift-off process. Fig. 3 - Fig. 8 show the lift-off when different materials such as aluminium, silicon, structural steel, iron, steel AISI 4340 and tungsten were used as the centre mirror plate. Aluminium was the material used for the cantilever beams. The initial normal stress was set as 5 GPa.

In Fig. 3 and Fig. 4, the maximum lift-off is present at the edges and edge displacement relative to centre plate is more than 0.2 mm. Also, there is surface deformation at the edges which makes the structure unsuitable for a mirror plate.

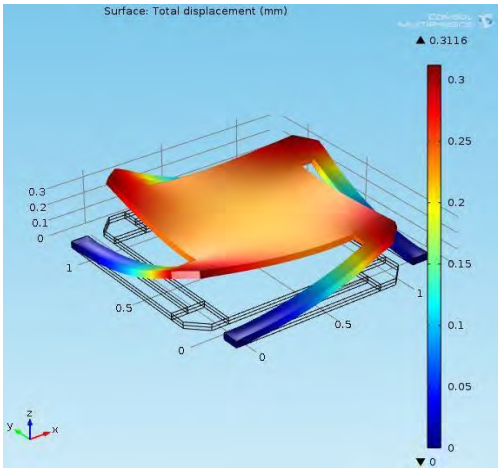


Figure 3. Total surface displacement for aluminium.

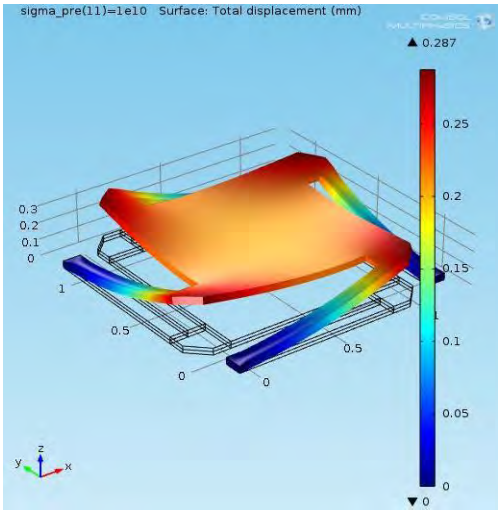


Figure 4. Total surface displacement for silicon.

For structural steel plate, the edge displacement is less compared to aluminium and silicon mirror plate as shown in Fig. 5. Steel deforms less because it is stiffer compared to aluminium and silicon. Structural steel has a Young's modulus of 200 GPa whereas that of aluminium and silicon is 69 GPa and 150 GPa respectively.

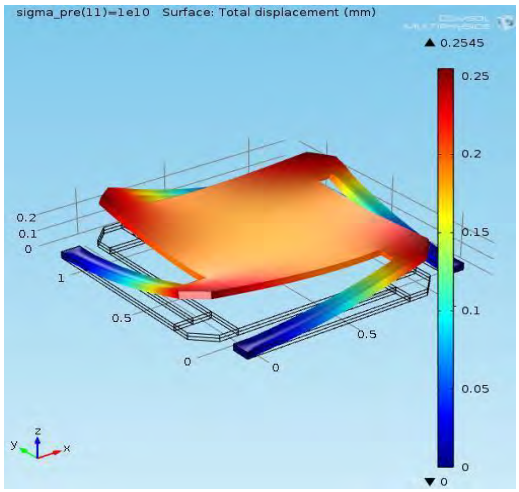


Figure 5. Total surface displacement for structural steel.

The maximum lift-off for iron and steel AISI 4340 plates are less than that of structural steel as shown in Fig. 6 and Fig. 7 respectively.

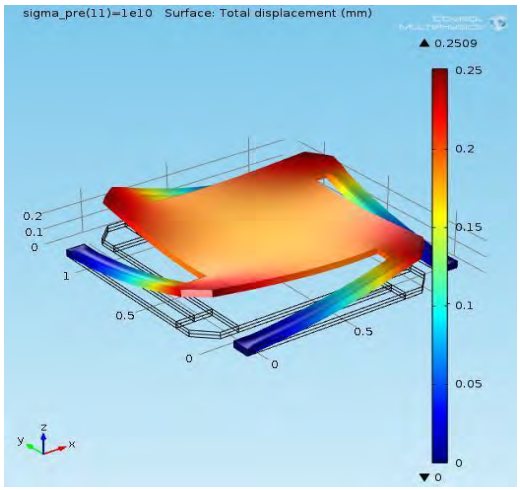


Figure 6. Total surface displacement for iron.

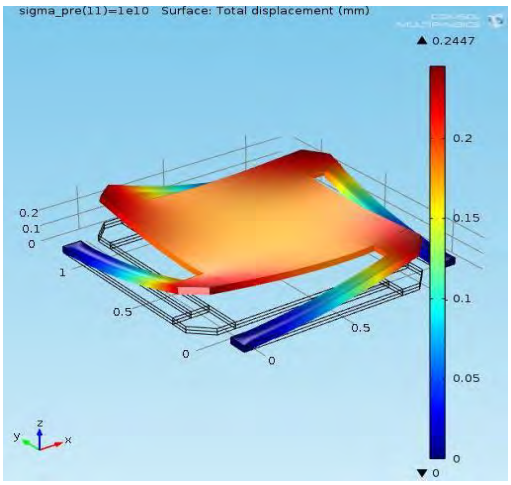


Figure 7. Total surface displacement for steel AISI 4340.

Tungsten has a Young's modulus of 411 GPa and it is stiffer compared to other materials. Therefore, tungsten-aluminum combination exhibits the least displacement as shown in Fig. 8

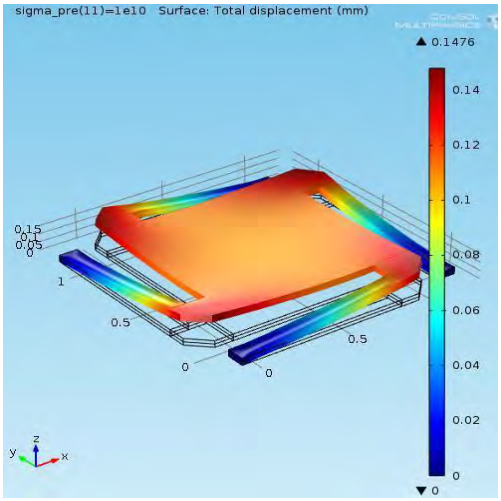


Figure 8. Total surface displacement for tungsten.

In all the material combinations analysed, the maximum displacement was observed at the ends of cantilever beam connected to the mirror plate and minimum at the opposite ends.

Table I shows a comparison of maximum lift-off for different material combinations of the centre mirror plate and the cantilever beams.

TABLE I. COMPARISON OF LIFT-OFF FOR DIFFERENT MATERIAL COMBINATIONS

Material of centre plate	Maximum Displacement or Lift-off (mm)
Aluminium	0.3116
Silicon	0.287
Structural Steel	0.2545
Iron	0.2509
Steel AISI 4340	0.2447
Tungsten	0.1476

Fig. 9. shows the total edge displacement for structural steel-aluminium combination which has a maximum value of 0.15 mm and therefore this material combination gives a comparatively stable structure with less surface deformation. In the case of other combinations the lift-off is too low.

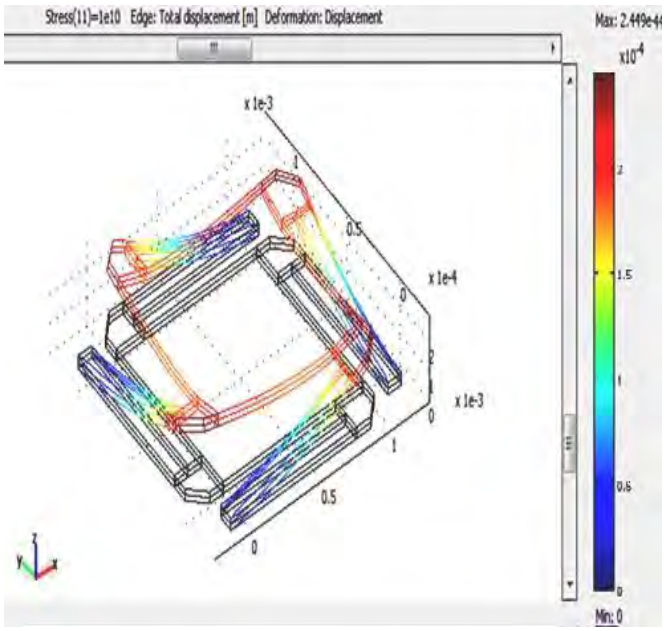


Figure 9. Edge: Total displacement: structural steel

B. *Displacement v/s Prestress Analysis*

Fig. 10 shows the variation of the total centre point deflection with different prestress levels for different centre plate materials. The mirror response is nearly linear in case of all the centre plate materials used. Silicon exhibits greater deflection compared to other materials for all the values of applied prestress, but it shows high displacement at the edges of the centre plate.

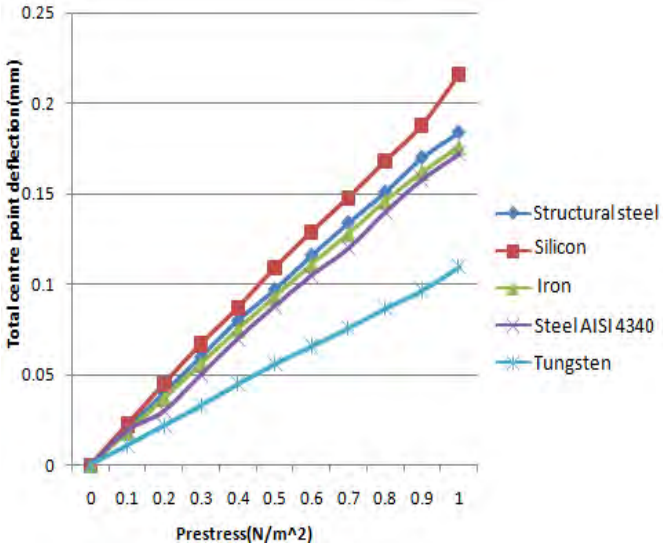


Figure 10. Total centre point displacement v/s prestress for different material combinations.

Fig. 11 shows the mirror's curvature along its centerline for a range of values of prestress. As per the graph, mirror bends more with increasing stress.

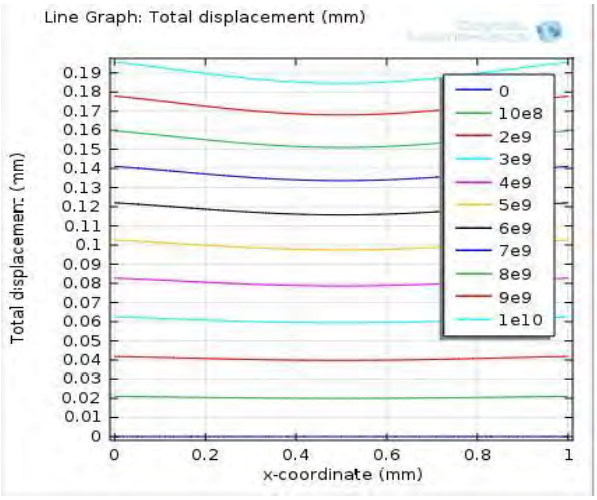


Figure 11. Total centre point deflection for different prestress levels: structural steel plate.

C. von Mises Stress Analysis

The effective stress at which yielding of any ductile material occur is known as the yield strength. In order to avoid static or fatigue failure of the structure, the von Mises stress induced in the material should be less than its yield strength. In (1), σ_v and σ_y denote von Mises stress and yield strength respectively.

$$\sigma_v \leq \sigma_y \tag{1}$$

The von Mises stress, σ_v can be expressed in terms of principal stresses $\sigma_1, \sigma_2, \sigma_3$ as given in (2).

$$\sigma_v = \left[\frac{(\sigma_1-\sigma_2)^2+(\sigma_2-\sigma_3)^2+(\sigma_3-\sigma_1)^2}{2}\right]^{1/2} \tag{2}$$

Fig. 12 shows the distribution of von Mises stress over the surface of the mirror structure with structural steel plate. The von Mises stress is high at the edges and decreases towards the centre of the mirror plate. To ensure the safety of structure against failure, von Mises stress must be less than the yield strength of the material (250 MPa). As per the Fig.13, the structure can be safely operated for prestress values less than 1 GPa.

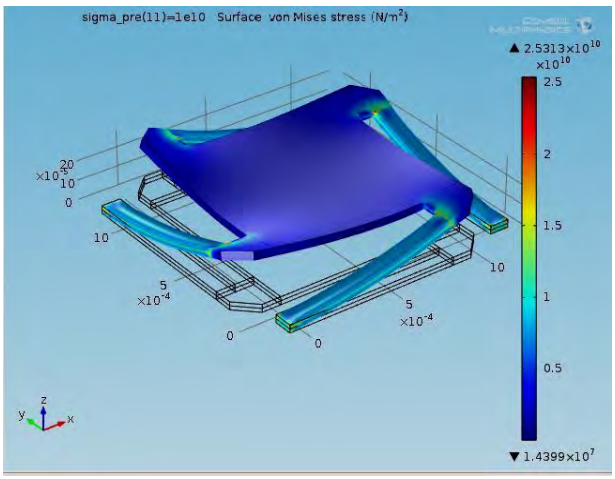


Figure 12. Surface von Mises stress: structural steel plate.

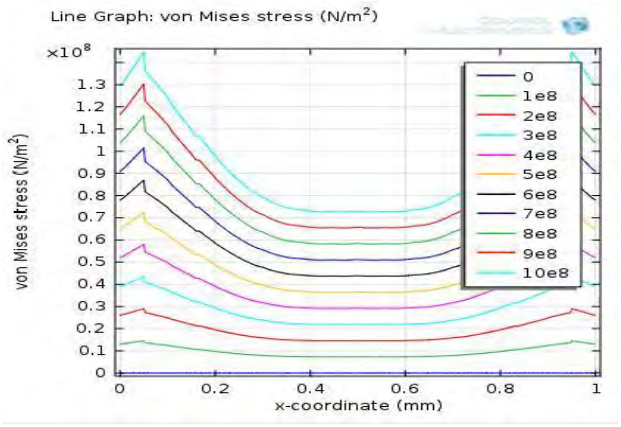


Figure 13. Surface von Mises stress for different prestress levels.

IV. CONCLUSION

In this paper, an electrostatically controlled micromirrors was modeled and analysed using COMSOL Multiphysics software. The lift-off analysis was carried out for different mirror plate materials. The effect of different prestress levels on the deflection of mirror plate was also studied. The result obtained shows that structural steel-aluminium micromirror has high lift–off with minimal surface deformation at the edges which improves the stability of the structure. The maximum value of von Mises stress is less than the yield strength of the material which ensures the safe operation of micromirror structure.

V. REFERENCES

[1] Olav Solgard, Asif A. Godil, Roger T. Howe, Luke P. Lee, Yves-Alain Peter, Hanes Zappe, "Optical MEMS:From micromirrors to complex systems," Journal of MicroElectroMechanical Systems, vol. 23, no.3, June 2014.

[2] Xingguo Xiong and Hanyu Xie, "MEMS Dual-mode Electrostatically Actuated Micromirror," Proceedingsof 2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1)

[3] Y. P. Zhu, W. J. Liu, K. M. Jia, W. J. Liao, and H. K. Xie, "A piezoelectric unimorph actuator based tip-tilt-piston micromirror with high fill factor and small tilt and lateral shift," *Sens. Actuators A, Phys.*, vol. 167, no. 2, pp. 495–501, 2011.

[4] Yuan Ma, Shariful Islam, and Ya-Jun Pan, "Electrostatic torsional micromirror With enhanced tilting angle using active control Methods", IEEE/Asme Transactions on Mechatronics, Vol. 16, No. 6, December 2011.

[5] M. Maheswaran and HarNarayan Upadhay, "A Study on Performance Driven Microfabrication methods for MEMS Comb-drive Actuators,"*Journal of Applied Sciences, 12: 920-928,* 2012.

[6] R. Sulima and S. Wiak, "Modelling of vertical electrostatic comb-drive for scanning micromirrors," *Int. J. Comput. Math. Electr. Electron. Eng.*, vol. 27, no. 4, pp. 780–787, 2008.

Material and Orientation Optimization for Quality Factor Enhancement of BAW Resonators

Reshma Raj R S and Resmi R
L B S Institute of Technology for Women, Thiruvananthapuram, India
reshmarajrs1991@gmail.com, resmilbs@gmail.com

Abstract - In this paper Quality factor of different Bulk Acoustic Wave resonators using various piezoelectric materials are analyzed. Higher the Quality factor, better is the resonator. High Quality factor provides higher signal to noise ratio, higher resolution and low power consumption. The results indicate that Rochelle Salt possess maximum value of Quality factor of about 2999.729. By changing the orientations maximum value of Quality Factor value obtained is about 42696.38 at 81◦. A MEMS based Bulk Acoustic Wave resonator is designed using COMSOL Multiphysics software.

Keywords - Bulk Acoustic Wave, Eigen frequency analysis, Quality Factor, BAW resonator

I. INTRODUCTION

BAW resonator is the core element of the BAW technology. BAW resonator is an electromechanical device in which, with the application of electrical signal a standing acoustic wave is generated in the bulk of piezoelectric material. In simple words, a device consisting of piezoelectric material is sandwiched between two metallic electrodes. To obtain the desired operating frequency, the natural frequency of the material and the thickness are used as design parameters. When the voltage is applied to the top electrode of the resonator, the bulk acoustic mode of the resonator is obtained from eigen frequency analysis.

The main mode of operation of BAW resonator [1] is the thickness or longitudinal mode, meaning that the bulk acoustic wave reflects the large plate surface and the resonance caused by the wave excited to the thickness direction.

A. Thin Film Bulk Acoustic Resonator

Thin Film Bulk Acoustic Resonator (FBAR or TFBAR) [2] is a device consisting of a piezoelectric material sandwiched between two electrodes and acoustically isolated from the surrounding medium with thicknesses ranging from several micrometres down to tenth of micrometres resonate in the frequency range of roughly 100 MHz to 10 GHz.In the wireless telecommunication world, film bulk acoustic wave resonators [3] are very promising for use as RF MEMS filters since they can be combined to make up duplexers (transmitter/receiver modules).

The FBAR is the chosen device in this study as an example of high frequency MEMS with thin film targeted for use in the ever growing wireless communications industry.

Bulk acoustic wave promises frequency in GHz range when integrated with RF circuits along with small size resonator and filters [4]. The resonator considered in the present paper is of thin film [5], [6] type in which the substrate is partially etched away on the back. The natural frequency of the material and Quality factor are used as design parameters to obtain the desired resonant frequency.

FBAR components offer small sizes, low cost, high quality factor, large power operation and compatibility with silicon low-cost process, enabling mass production and filter integration [7].

Fig.1 shows geometry of the modelled resonator. The lowest layer of the resonator is silicon substrate on top of which is the aluminum layer that operates as the ground electrode. A piezoelectric layer is laid over the ground electrode and above this piezoelectric layer is the metal electrode. The Perfectly Matched Layer (PML) effectively simulates the effect of propagation and absorption of elastic waves in the adjoining regions. The resonator dimensions used for simulating the resonator are shown in Table I.

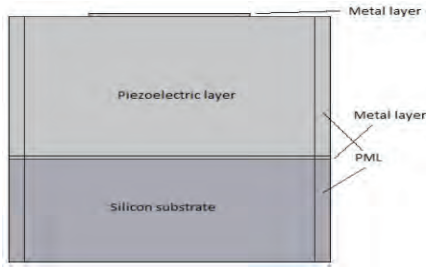


Figure 1. Geometry of Thin Film Bulk Acoustic Wave Resonator

TABLE I. RESONATOR DIMENSIONS

Parameters		Thickness	Width
Silicon layers		7μm	1.7mm
Metal layers (Aluminium)	top	0.2μm	500μm
	middle	0.2μm	1.7mm
Piezoelectric layer (Rochelle Salt)		9.5μm	1.7mm

II. EQUIVALENT CIRCUIT REPRESENTATION

For making circuit designs using MEMS resonators, an equivalent electrical circuit that describes their frequency dependent characteristics is needed. The equivalent Butterworth Van Dyke (BVD) lumped equivalent circuit of BAW resonator [8] is shown in Fig.2.

An obvious way to determine the frequency dependent admittance Y(ω) of a MEMS resonator is using a frequency response analysis. In this type of analysis, an actuation force of a frequency is applied to the surface of the resonator and the amplitude response is simulated. Because the Quality factor of MEMS resonators is usually high, this method requires a high density of frequency points to resolve a single resonance of the resonator. Moreover, fitting of Y(ω) is required to extract the equivalent circuit parameters.

In this paper an alternative method is described to directly extract the Quality factor from an eigen frequency analysis. The method will be applied to the resonator in Fig.2.

Eigen frequency and frequency domain analyses are different studies that are independent of each other. An eigen frequency step will gives list of all the natural frequencies of the system. Then we have to examine the deformed mode shape to see what displacement will be qualitatively. The normalization used in an eigen value problem is somewhat arbitrary. However, it is likely that the largest displacement will corresponds to the lowest mode.

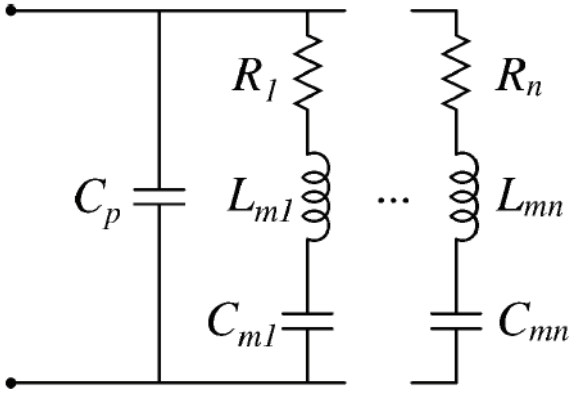


Figure 2. Equivalent Butterworth Van Dyke lumped element equivalent circuit of BAW resonator

Each branch contains a resistor, an inductor, and a capacitor, representing a resonance in the frequency response of the resonator. The admittance of the equivalent circuit is

$$Y(\omega) = \sum_{j=1}^{n} \frac{\frac{i\omega}{L_{mj}}}{\frac{1}{L_{mj}C_{mj}} - \omega^2 + i\omega\frac{R_j}{L_{mj}}} + i\omega C_p \qquad (1)$$

III. QUALITY FACTOR ANALYSIS

Quality factor is one of the most important characteristics of MEMS resonators, especially for vibrating structures where the resonant frequency variation is monitored.

Higher the Quality factor value better the resonator performance. Signal to noise ratio increases and power dissipation decreases. High Quality factor circuits can be used for various wireless applications. Quality factor value is a dimensionless parameter.

As we increase resistance value, the frequency range over which the dissipative characteristics dominate the behaviour of the circuit increases. Quality factor is related to the sharpness of the peak. Quality factor is an expression of the cyclic energy loss in an oscillating system.

In terms of energy, it is expressed as the total energy stored the system divided by the energy loss per cycle,

$$Q = 2\pi \times \frac{energy\ stored\ in\ the\ circuit}{energy\ dissipated\ per\ cycle} \qquad (2)$$

Quality factor analysis by using different materials is explained below. In Eigen frequency analysis there are 6 modes of frequencies. The lowest frequency for which these stationary waves are formed is called fundamental harmonic. The modes of oscillation have different shapes for different frequencies [9].

Here in this paper Quality factor is evaluated using eigen frequency,ω_i from the equation given below.

$$\text{Quality Factor},Q_i = \frac{Re\ \omega_i}{2I_m\omega_i} \qquad (3)$$

Table II shows the Quality factor analysis of Zinc Oxide. By using Zinc Oxide maximum value of Quality factor obtained is about 1330.08 at an Eigen frequency of 221.42 MHz.

TABLE II. QUALITY FACTOR VALUES OBTAINED AT DIFFERENT EIGEN FREQUENCY VALUES USING ZINC OXIDE

Eigen Frequency(MHZ)	Quality Factor
217.73	1199.7624
218.7	936.23781
219.55	1188.7183
220.63	936.51987
221.42	1330.0896
222.13	937.4436

Table III shows the Quality factor analysis of Aluminium Nitride. By using Aluminium Nitride maximum value of

Quality factor obtained is about 1808.26 at an Eigen frequency of 225.42 MHz.

TABLE III. QUALITY FACTOR VALUES OBTAINED AT DIFFERENT EIGEN FREQUENCY VALUES USING ALUMINIUM NITRIDE

Eigen Frequency(MHZ)	Quality Factor
214.8	1616.2625
216.03	1784.7664
221.3	1662.5064
222.09	1784.1761
225.42	1808.2671

Table IV shows the Quality factor analysis of Rochelle Salt. By using Rochelle Salt maximum value of Quality factor obtained is about 2999.72 at an Eigen frequency of 218.95 MHz.

TABLE IV. QUALITY FACTOR VALUES OBTAINED AT DIFFERENT EIGEN FREQUENCY VALUES USING ROCHELLE SALT

Eigen Frequency(MHZ)	Quality Factor
218.76	1165.6162
218.95	2999.7291
219.38	1185.5266
220.26	1083.3847
220.99	1228.3316

Table V shows the Quality factor analysis of Lithium Niobate. By using Lithium Niobate maximum value of Quality factor obtained is about 1360.2926 at an Eigen frequency of 222.88 MHz.

TABLE V: QUALITY FACTOR VALUES OBTAINED AT DIFFERENT EIGEN FREQUENCY VALUES USING LITHIUM NIOBATE

Eigen Frequency(MHZ)	Quality Factor
217.21	665.18125
219.17	1152.154
219.59	619.25414
221.03	1309.1761
222.57	681.24313
222.88	1360.2926

Quality factor Vs Eigen Frequency plot of different materials is shown in Fig.3 below.

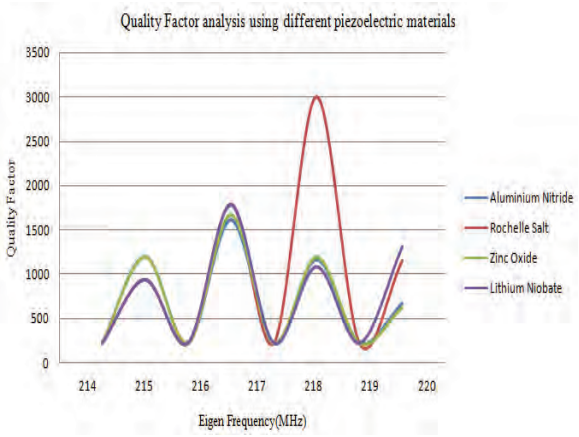


Figure 3.Quality factor analysis by using different piezoelectric materials

From the Quality factor analysis it is clear that by using Rochelle Salt a maximum Quality factor value of about 2999.729 is obtained.

Further analysis is done by using Rochelle Salt as a piezoelectric material via varying the orientations. Here, the piezoelectric layer is rotated along the X axis.

The eigen-frequencies and respective Quality factor change when the piezoelectric materials inside the resonator were rotated. This is due to the crystallographic effect [10] of piezoelectric materials. However, any angular misalignment from the axis of transduction disrupts the atomic linearity which might lead to acoustic losses at the atomic level, the ensemble of which might reflect on increase in the Quality factor of the resonator.

Table VI shows the Quality factor Vs Eigen frequency plot obtained using Rochelle Salt by rotating piezoelectric layer along 30°. Here maximum value of Quality factor obtained is about 17705.13955 at an Eigen frequency of 220.6 MHz.

TABLE VI. QUALITY FACTOR ANALYSIS BY ROTATING PIEZOELECTRIC LAYER ALONG 30◦

Eigen Frequency(MHZ)	Quality Factor
215.47	5601.25809
218.12	12669.76214
219.6	13173.77809
220.6	17705.13955
222.62	11062.03554

Table VII shows the Quality factor Vs Eigen frequency plot obtained using Rochelle Salt by rotating piezoelectric layer along 50°. Here maximum value of Quality factor obtained is about 40998.42342 at an Eigen frequency of 222.18 MHz.

TABLE VII. QUALITY FACTOR ANALYSIS BY ROTATING PIEZOELECTRIC LAYER ALONG 50◦

Eigen Frequency(MHZ)	Quality Factor
214.95	19496.29908
216.47	32516.34201
218.1	35075.73867
219.56	29980.49335
222.18	40998.42342

Table VIII shows the Quality factor Vs Eigen frequency plot obtained using Rochelle Salt by rotating piezoelectric layer along 60°. Here maximum value of Quality Factor obtained is about 32554.42619 at an Eigen frequency of 220.42 MHz.

TABLE VIII.QUALITY FACTOR ANALYSIS BY ROTATING PIEZOELECTRIC LAYER ALONG 60◦

Eigen Frequency(MHZ)	Quality Factor
214.36	32137.45978
217.7	28887.5745
220.42	32554.42619
221.27	18363.90608
224.99	17416.11227

Table IX shows the Q factor Vs Eigen frequency plot obtained using Rochelle Salt by rotating piezoelectric layer along 81°. Here maximum value of Quality Factor obtained is about 42696.38989 at an Eigen frequency of 218.84 MHz.

TABLE IX. QUALITY FACTOR ANALYSIS BY ROTATING PIEZOELECTRIC LAYER ALONG 81◦

Eigen Frequency(MHZ)	Quality Factor
216.15	23987.13199
216.76	13140.93297
218.84	42696.38989
220.4	38909.29835
223.42	26911.90005

Quality factor analysis by varying orientations is shown in Fig.4 below.

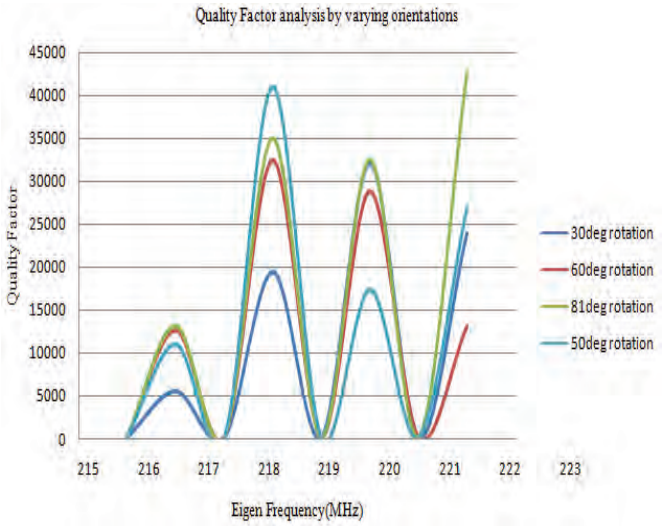


Figure 4.Quality Factor analysis at different orientations

IV. RESULTS

Eigen frequency analysis is used to find out Quality factor. All the design and modeling are done using COMSOL Multiphysics software. The Eigen frequency analysis plot shown in Fig.5 shows the lowest BAW mode of the structure that occurs at a particular frequency obtained via material optimization.

The eigen frequency values are shown in results as complex numbers wherein the real part provides the actual displacement frequency while the imaginary part is an indication of the extent of damping.

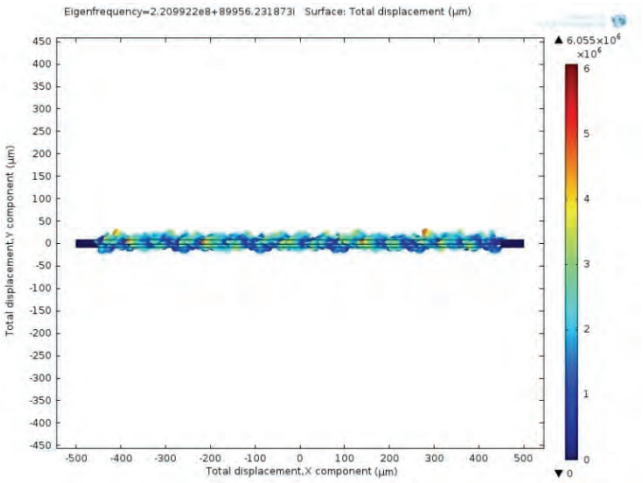


Figure 5.The lowest bulk acoustic mode of the resonator identified from the solutions of the Eigen frequency analysis via using Rochelle Salt as piezoelectric layer.

Fig.6 shows the lowest BAW mode of the structure that occurs at a particular frequency obtained via changing orientation. By changing orientations maximum value of Quality factor (42696.38989) is obtained by rotating piezoelectric layer along 81°. This is due to the crystallographic

effect of piezoelectric materials. The colour variation across the surface of the resonator and the corresponding displacement for that eigen frequencies are shown in the colour chart at the right side of the plot in Fig.5 and Fig.6.

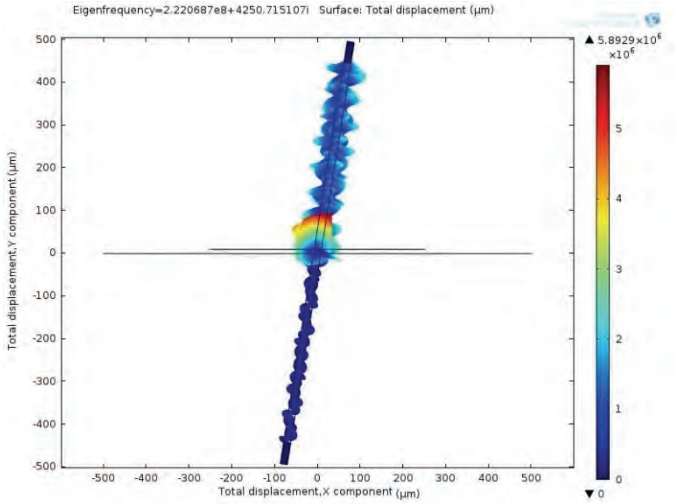


Figure 6. The lowest bulk acoustic mode of the resonator identified from the solutions of the Eigen frequency analysis by rotating piezoelectric layer along 81◦

V. CONCLUSION

Quality factor of different Bulk Acoustic Wave resonators using various piezoelectric materials are analyzed. It was found that by using Rochelle Salt a Quality factor of about 2999.729 is obtained at a frequency of 218.955MHz.This is the highest value obtained by analysing these piezoelectric materials. By changing the orientations it is found that at 81° maximum value of Quality factor (42696.38) is obtained.Higher the Quality factor value better the resonator and by improving Quality factor signal to noise ratio and sensitivity are also increased.

REFERENCES

[1] Sneha Dixit," Film Bulk Acoustic Wave (FBAR) Resonator", in Scientific andResearchPublications,Vol.5Iss.2,2015

[2]Gabriel.M.Rebeiz(2003),RF MEMS Theory, Design and Technology

[3] Tapani Makkonen, Antti Holappa, Juha Ella, and Martti M. Salomaa,"Finite element simulations of Thin-Film Composite BAW Resonators," IEEE Trans. On Ultrason, Ferroelectrics, and Frequency Cotrol, VOL. 48, no5,pp.1241-1258, 2001

[4] W.A. Burkland, A.R. Landlin, G.R. Kline, and R.S. Ketcham, "A thin-film-bulk-acoustic-wave resonator-controlled oscillator on silicon," IEEE Electron. Device Lett., vol. EDL-8, no. 11, pp. 531-533, 1987

[5] R.B. Stokes, J.D. Crawford, and D. Cushman, "Monolithic bulk acoustic filters to X-band in GaAs," in Proc. IEEE Ultrason. Symp., pp. 547-551, 1993

[6] T.W. Grudkowski, J.F. Black, T.M. Reeder, D.E. Cullen, and R.A. Wagner, "Fundamental-mode VHF/UHF miniature acoustic resonators and filters on silicon," Appl. Phys. Lett., vol. 37, no. 11, pp. 993-995, 1980

[7] R.F. Milsom and others, "Effect of mesa-shaping on spurious modes in ZnO/Si bulk-wave composite resonators," in Proc. IEEE Ultrason. Symp., pp. 498–503.

[8] K.M. Lakin, "Modelling of thin film resonators and filters," IEEE MTT-S Digest, vol. 1, pp. 149-152, 1992.

[9] Tapani Makkonen, Member, IEEE, Antti Holappa, Juha Ell¨a, and Martti M. Salomaa, "Finite Element Simulations of Thin-Film Composite BAW Resonators," in IEEE transactions on ultrasonics, ferroelectrics, and frequency control,vol.48,no.5,2007

[10] RanjanDey,D.Anumeha," Design and Simulation of Portable Fuel Adulteration Detection Kit",in Journal of Energy and Chemical Engineering, Vol.2Iss.2,PP.74-80,2014

Impact of Propagation Medium on Link Quality for Underwater and Underground Sensors

G. Horvat, D. Vinko and J. Vlaović
University of Osijek, Faculty of Electrical Engineering, Department of Communications
Osijek, Croatia
goran.horvat@etfos.hr

Abstract - With the rapid development of wireless networks new application domains are continuously proposed. One of these applications includes an underwater and underground sensor which relies on the properties of diminished RF performance to send the gathered information to a base station. Although many researchers deal with the theoretical background of the channel and propagation models, very few experimental studies regarding the link quality were conducted. Due to the fact that new and more powerful sensor nodes are continuously being developed, the question that arises is how these new and more powerful classes of wireless sensors can handle the harsh environment in underwater and underground networks. Therefore, in this paper the authors analyze the impact of soil, water, moisture and other parameters on the link quality between two sensor nodes in various scenarios. The obtained results present a basis for designing and planning an underground/underwater sensor network.

I. INTRODUCTION

The performance of wireless sensor nodes (WSN) depends on the environment and the medium in which they operate. The underwater and underground environment significantly differ from the terrestrial environment where WSNs are commonly used. There is a number of papers proposing new protocols to cope with the propagation problems in both underwater [1] - [4] and underground [5] environment. If we compare underwater and underground environment, the research focuses on a different set of problems. In underwater applications, the focus is set on problems such as deployment algorithms for node placement [6], [7], powering and energy efficiency issues [8], [9]. In underground communications, the focus is mainly set on various applications in coal and metal mines [10] - [13], and underground train tunnels [14]. The security aspect [15], [16] and localization/tracking [17] is also being discussed but mainly for WSN applications in underground environments.

When it comes to communication, researchers are exploring different approaches to extend the communication range in the underwater environment by using wireless sensor nodes with optical [18], [19] and acoustic communication [20], [21]. This is due to high attenuation of the electromagnetic (EM) wave [22] in water. Additional circuitry (optical and acoustic transducers) used to extend the communication range has a negative impact on power consumption and battery life of the sensor node. In both, the underground environment (such as coal mines and train tunnels) and terrestrial environment, a propagation medium is still air. Thus, standard EM wave propagation is used for communication.

This paper focuses on sensor node performance when the node is transmitting EM waves through water or ground. There are some papers that deal with such channel modelling and there are few experimental studies in underwater [23] and underground [24] environment, but the link quality has not yet been thoroughly evaluated.

Furthermore, as new sensor nodes are continuously being developed (higher RF sensitivity, higher RF power), the question that arises is how these new and more powerful classes of wireless sensors can handle the harsh environment in underwater and underground networks. Therefore, in this paper the following parameters are being analyzed in various scenarios: Received Signal Strength Indicator (RSSI), Link Quality Indicator (LQI), Round Trip Time (RTT), and Packet delivery probability (p_{rr}). The main objective of underground communication is to analyze the impact of communication distance and soil moisture on the monitored parameters. The goal of underwater communication is to analyze how the distance and additional impurities (salinity) affect the observed parameters.

This paper is structured as follows: the measurement testbed for measuring underwater and underground communication is proposed in Section II. In Section III the measurement results for underground communication are presented, whereas in Section IV the measurement results of underwater communication are presented. Section V gives the conclusion of the proposed work and the guidelines for future work.

II. MEASUREMENT TESTBED

In order to examine the link quality parameters in the underwater and underground networks a testbed that consists of a point-to-point WSN and a testing chamber was proposed. Used WSN topology is based on a star topology, with one network Coordinator and *n* End nodes which form a WSN. Hence, in the proposed topology the first node is a network Coordinator whereas the second node is an End node. The diagram of the proposed measurement testbed is shown in Fig. 1.

This work was sponsored by the J. J. Strossmayer University of Osijek under the project IZIP-2014-104 "Wireless power transfer for underground and underwater sensors"

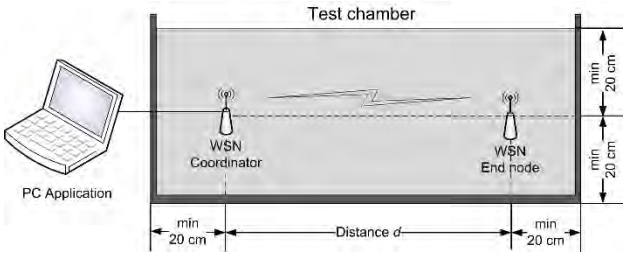


Figure 1. The measurement testbed for underwater and underground sensors

The network Coordinator is connected to a personal computer (PC) via a USB cable, which is also used for powering the Coordinator node. The End node is a battery powered node. An application designed for testing and measurement purposes is situated on the PC that communicates with the Coordinator. Both nodes are housed in waterproof containers to prevent contamination by moisture. The distance between the node's antennas and the test chamber wall is greater than the diameter of the 1st Fresnel zone (> 0.18 m); no relevant part of RF radiation will propagate outside the test chamber. One of the analyzed testbed parameters is the distance between the nodes which refers to the distance between node antennas. The nodes are assumed to be placed on the same height and both antennas are directed to each other; PIFA (PCB Inverted F Antenna) with a gain of 2.28 dBi. A block diagram of the WSN node is shown in Fig. 2.

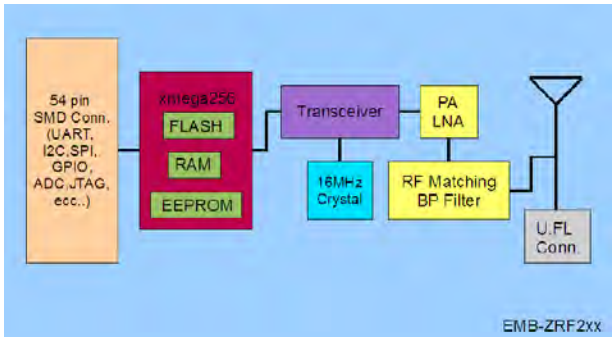


Figure 2. Block diagram of the used WSN node

The node is composed of a microcontroller and a RF section which consists of a transceiver and a Power Amplifier / Low Noise Amplifier (PA/LNA). With the proposed architecture the used nodes have higher TX power and higher RX sensitivity. The characteristics of the used WSN node are shown in Table I.

TABLE I
WSN NODE PARAMETERS

Parameter	Value
Frequency and modulation	2.4 GHz, DSSS, 250 kBit/s
Transciever	AT86RF231
Microcontroller	Atmel ATxmega256A3U
Communication stack	Atmel LightWeight Mesh
Transmission power	20 dBm
Receiver sensitivity	-105 dBm
Packet generation interval	uniform (1s)
Number of packets sent	100
Packet size	100B
macMaxCSMABackoffs	4
macMaxFrameRetries	3

The proposed methodology for measuring the link quality is based on convergecasting properties of the WSN; multiple nodes transmit measured values to a single coordinator of the network, but taking into account round trip communication. In the proposed methodology the sensor node sends a packet towards the coordinator and waits for the reception of an acknowledgement packet. In this manner each sensor node has a feedback information whether the measured data reached the destination, what is the delay (Round Trip Time, RTT) of the packet and what is the packet delivery probability (p_{rr}). During the testing, the node will send 100 packets, each separated by a time delay of uniform distribution U (0.1 s, 1 s). After the packets are sent, the end node will analyze the data and transmit the link quality information to the coordinator.

This methodology is proposed due to the fact that in harsh environments often End nodes lose communication with the Coordinator due to changes in environmental factors. With the proposed methodology a sensor node is aware of the link quality information and can therefore decide to transmit the data via RF or to store the data for future transmission (upon link quality improvement).

In the course of the experiment two scenarios were analyzed: underwater and underground communication by means of filling the test chamber with soil or water.

III. MEASUREMENT RESULTS – UNDERGROUND COMMUNICATION

First scenario estimates link quality in underground communication in WSNs. The proposed methodology of the experiment consists of several steps. The first step consists of measuring the link quality for various distances in a constant moisture of soil for uplink and downlink communication. In this step the testing chamber was filled with commercial soil of constant soil moisture of 56.2 %. The approximate density of the soil was 300 kg/m^3. Measurement was performed each 10 cm until the distance of 120 cm. The measurement results for *RSSI* value are shown in Fig. 3

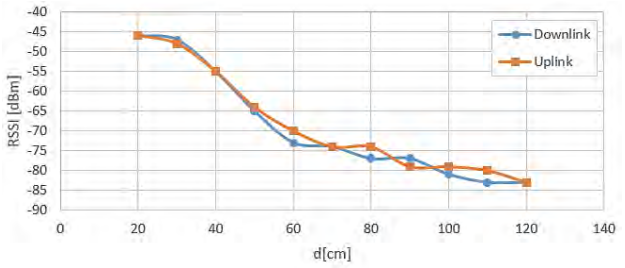


Figure 3. RSSI for various distances vs. Downlink and Uplink channel (underground – soil moisture 56.2 %)

As seen in Fig. 3 as the distance of nodes increases, the *RSSI* value decreases accordingly. The trend can be seen as a log distance model, widely known as log normal shadowing model. If the graph is shown as a log distance diagram, it can be seen that it is possible to determine the

path loss exponent for underground communication [26]. First it is important to define the path loss of a communication link as:

$$PL[dB] = P_t[dB] - P_r[dBm] + G_t[dB] + G_r[dB] \quad (1)$$

where PL refers to the path loss, P_t refers to the transmitted Power of the node (20 dBm), P_r refers to the received power (RSSI), and G_t and G_r refer to the antenna gain of the receiver and transceiver, respectively (2.21 dB). From (1) it is possible to calculate a linear regression line and determine the regression coefficients (Fig. 4).

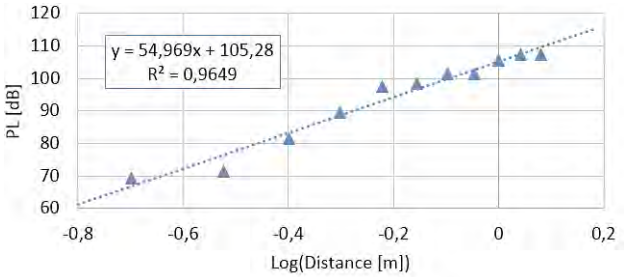


Figure 4. Path loss for logarithmic distance with linear regression (underground)

It can be seen from Fig. 4 that there is high correlation with linear regression line and measurement data (coefficient of determination value of 0.96) vs. the logarithmic distance between nodes. As there are some variations from the linear model these differences can be accounted to multipath fading, reflections diffractions and similar radio irregularity phenomenon [25]. These variations can be modeled using a very widely used log-normal shadowing model:

$$PL(d) = PL(d_0) + 10 \cdot n \cdot \log\left(\frac{d}{d_0}\right) + X_\sigma \quad (2)$$

where PL is the path loss at referenced distance $PL(d0)$, n is the path loss exponent that is dependent on the propagation medium and X_σ is the normally distributed random variable with zero mean and standard deviation σ. The derived linear regression equation from Fig. 4 is

$$y = 54.969\,x + 105.28 \quad (3)$$

To determine the path loss exponent of the underground communication we correlate equations (3) and (2), which results in determining path loss exponent model as follows:

$$PL(d) = 105.28dB + 10 \cdot 5.496 \cdot \log(d) \quad (4)$$

Then, from the analysis of other parameters of the link quality analyzed in this step (LQI, RTT, p_{rr}), it can be concluded that the LQI ranges from 98.8 % to 100 % of the link quality and the RTT ranges from 99 ms to 118 ms (averaging 106.6 ms) without any direct correlation with the distance or the $RSSI$ value. Also, the packet delivery probability averages 98.6 % with no direct correlation with $RSSI$ or distance. It is important to notice that the used protocol stack uses packet retransmission on the MAC layer which improves the reliability of the communication and achieves high packet delivery probability.

TABLE II
UNDERGROUND EMPIRICAL PATH LOSS MODEL

Parameter	Value
Path loss equation	$10{\cdot}5.496{\cdot}\log(d) + 105.28$
n	5.496
$PL(d_0 = 1\text{ m})$	105.28 dB
X_σ	2.6 dB

The next step in measurements includes the increasing of the soil moisture and measuring link quality parameters. Soil moisture was controlled by adding tap water, with quality parameters described onwards in Section IV. First measured link quality parameter is $RSSI$ (Fig. 5)

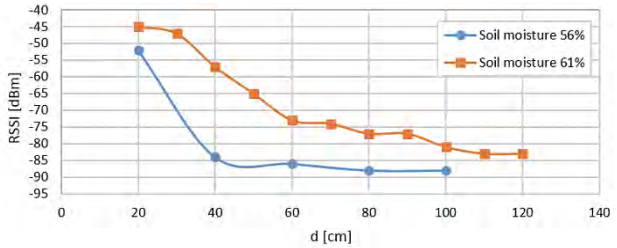


Figure 5. RSSI for various distances and soil moisture content (underground)

In Fig. 5 the previous measurement containing 56 % of moisture was shown alongside with the increased soil moisture measurement of 61 %. It can be concluded that an increase in soil moisture of only 5 % drastically increases signal attenuation and increases the path loss. This can be seen in Fig. 6 where the $RSSI$ value is shown for a distance of 50 cm in regards to soil moisture.

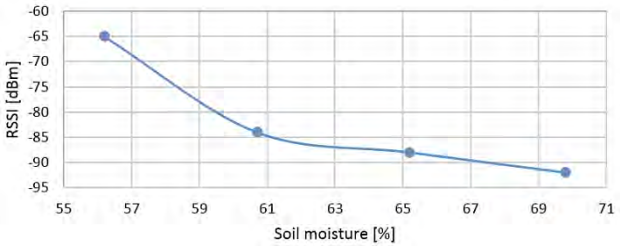


Figure 6. RSSI vs. soil moisture content at distance of 50cm (underground)

From Fig.6 it is evident that the increase in soil moisture above 70% will result in communication link failure at distances greater than 50 cm. This presents a very important factor that needs to be taken into consideration upon planning of an underground WSN. The influence of soil moisture and distance for round-trip-time (RTT) was analyzed next (Fig. 7).

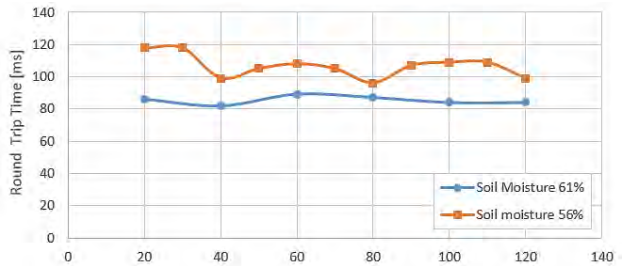


Figure 7. Measurements of RTT vs. distance vs. soil moisture (underground)

From the performed measurements shown in Fig. 7 it can be concluded that the distance has no significant effect on the *RTT* parameter. However, soil moisture has a significant impact on the *RTT* with a negative trend, meaning that the increase in moisture (increase in RF attenuation) reduces average *RTT* for the sent packet. This can be accounted for the reduction of reflections and diffractions due to the increase in path loss of the medium. This can also be seen for higher moisture contents where *RTT* value settles at the minimum value (Fig. 8).

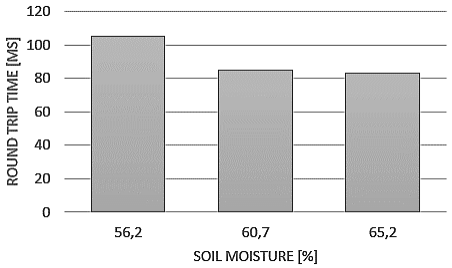


Figure 8. Measurements of RTT vs. distance vs. soil moisture (underground)

By increasing the soil moisture beyond the 65%, the communication is drastically affected and the packet delivery probability drops to only 9 % (91 % of the packets do not reach the destination – the Coordinator). In these conditions the *RTT* increases due to the increase in retransmissions (Fig. 9).

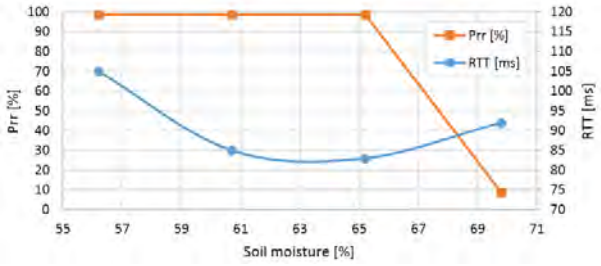


Figure 9. *RTT* and p_{rr} for various soil moisture content (underground)

From the measurement results it can be concluded that for underground communication the most influenced parameter is soil moisture. Soil moisture can significantly degrade link quality and cause communication link failure. On the other hand, by increasing soil moisture a drop in RTT was observed that could be used as a model for advanced communication protocols.

In comparison with path propagation in air, at maximum distance (120cm) the RSSI value for LOS air propagation equals -47 dBm, RTT equals 71 ms and packet delivery probability equals 100%, representing perfect communication condition. This shows the amount of attenuation of RF propagation induced by the soil or similar propagation medium.

IV. MEASUREMENT RESULTS – UNDERWATER COMMUNICATION

A second scenario involves the analysis of underwater communication in WSNs considering link quality parameters. The proposed methodology of the experiment consists of several steps. The first step consists of measuring the link quality for various distances in water for uplink and downlink communication. In this step, the test chamber is filled with water (approximately 110 L) and the nodes are positioned so that the distance between the antenna and the edge of the chamber is greater than the size of the 1st Fresnel zone. Nodes with antennas are positioned in a hermetically sealed enclosure made from 0.1mm nylon material, with PVC support. Measurement was conducted using tap water at temperature of 15 °C with conductivity of 870 µS/cm, hardness of 295 mg $CaCO_3$, °D 16.5 and TDS 460 mg/L [27]. This value for conductivity is very close to the limit for brackish water (1000 µS/cm).

Measurement was conducted at various distances where the antennas were positioned in a straight line. Measurement results for Uplink and Downlink value of RSSI link quality parameter are shown in Fig. 10.

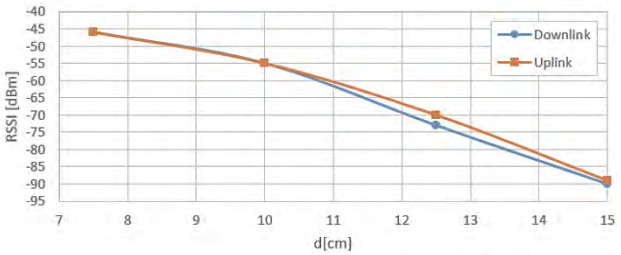


Figure 10. RSSI for various distances vs. Downlink and Uplink channel (underwater)

As seen from measurement results Uplink and Downlink channel results are consistent, with little deviation, resulting in an almost symmetrical communication link. On the other hand, the path loss in underwater communication exhibits large amounts of attenuation which results in a maximum communication distance of only 15 cm. This can be accounted for large value of water conductivity, CaCO3 and TDS. On the other hand, other link quality parameters are not significantly affected by the change in distance, resulting in mean value of packet delivery probability of 98.3%, mean value of RTT of 115.8 ms and average LQI of 254.5, not having any correlation with RSSI or distance between the nodes.

On the other hand, as water quality parameters have a profound impact on the communication and link quality, the second step consists of adding impurities (salinity) in order to increase the conductivity of water. Salinity was increased up to 3 ‰ and RSSI was measured at the distance of 10 cm between the End node and the Coordinator. Results are shown in Fig. 11.

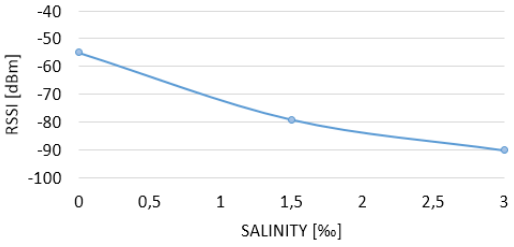


Figure 11. RSSI for different values of salinity at 10cm distance (underwater)

From the measurement results it can be concluded that the increase in salinity drastically affects the RSSI value; increases the path loss in the channel. However, other link quality parameters are not significantly affected by the change in distance; not having any correlation with salinity concentration.

V. CONCLUSION

This paper presents a study on the impact of propagation medium on link quality parameters for underwater and underground Wireless Sensor Networks. As the propagation medium can be versatile, this paper analyzes the influence of soil moisture on the propagation in underground networks and impurity levels in underwater communication. Measurements were performed with the proposed testbed composed of two WSN nodes; a Coordinator node and an End node. Nodes communicate within 2.4GHz band with TX power of 20 dBm, receiver sensitivity -105 dBm and PIFA antenna with gain of 2.21 dBi. The testbed is composed of a testing chamber filled with medium (soil/water) and a PC used to analyze the communication. Link quality parameters that were analyzed within this paper are *RSSI*, *LQI*, *RTT* and p_{rr}.

Underground communication: By performing detailed measurements for various distances an empirical propagation model is proposed where the path loss exponent *n* and the path loss were calculated. From the measurement results it can be concluded that soil moisture has a profound effect on the link quality where the increase in soil moisture can reduce communication range by a factor of two, for only 10 % increase in soil moisture. An interesting phenomenon was observed whereby increasing soil moisture the communication delay RTT was reduced.

Underwater communication: Similarly to underground communication an analysis was performed for various distances, regarding link quality parameters. Tap water was used with described water quality parameters. It can be concluded that the path loss in water with relatively high conductivity (850 µS/cm) is significant, resulting in a maximum communication distance of only 15 cm. As the content of impurities significantly affects communication, the next step involved adding impurities in order to increase the conductivity – salt. The increase in water salinity drastically affects the RSSI value; increases the path loss in the channel. However, other link quality parameters are not significantly affected by the change in distance; not having any correlation with salinity concentration. When salinity is 3 ‰, the maximum communication distance drops to 10 cm.

Future work on the topic of underground communication targets calculating the path loss exponent for various soil moisture contents and various soil densities in order to present an empirical path-loss model for underground communication. Also, as the soil moisture affects RSSI, soil composition such as the presence of salts, or ions in the soil will change soil electromagnetic properties, affecting radio propagation. These properties will be investigated in future work. Furthermore, similar approach will be taken in underwater communication where the path-loss model will be proposed based on different water quality factors (conductivity, salinity, TDS etc.) and isolating the most influenced factor for underwater communication. Also, the measurement data will be compared with theoretical expectations from the available literature – existing propagation models.

ACKNOWLEDGMENT

The authors would like to thank Krunoslav Aladić, PhD from Croatian veterinary institute, Branch - Veterinary department Vinkovci for the analysis of soil moisture content.

REFERENCES

[1] Umar, A.; Akbar, M.; Iqbal, Z.; Khan, Z.A.; Qasim, U.; Javaid, N., "Cooperative partner nodes selection criteria for cooperative routing in underwater WSNs," in *Information Technology: Towards New Smart World (NSITNSW), 2015 5th National Symposium on* , vol., no., pp.1-7, 17-19 Feb. 2015

[2] Fahim, H.; Javaid, N.; Qasim, U.; Khan, Z.A.; Javed, S.; Hayat, A.; Iqbal, Z.; Rehman, G., "Interference and Bandwidth Aware Depth Based Routing Protocols in Underwater WSNs," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2015 9th International Conference on* , vol., no., pp.78-85, 8-10 July 2015

[3] Liaqat, T.; Javaid, N.; Ali, S.M.; Imran, M.; Alnuem, M., "Depth-Based Energy-Balanced Hybrid Routing Protocol for Underwater WSNs," in *Network-Based Information Systems (NBiS), 2015 18th International Conference on* , vol., no., pp.20-25, 2-4 Sept. 2015

[4] Shah, M.; Javaid, N.; Imran, M.; Guizani, M.; Khan, Z.A.; Qasim, U., "Interference Aware Inverse EEDBR protocol for Underwater WSNs," in *Wireless Communications and Mobile Computing Conference (IWCMC), 2015 International* , vol., no., pp.739-744, 24-28 Aug. 2015

[5] Zhiping Zheng; Shengbo Hu, "Research challenges involving cross-layered communication protocol design for underground WSNS," in *Anti-counterfeiting, Security and Identification, 2008. ASID 2008. 2nd International Conference on* , vol., no., pp.120-123, 20-23 Aug. 2008

[6] Felamban, M.; Shihada, B.; Jamshaid, K., "Optimal Node Placement in Underwater Wireless Sensor Networks," in *Advanced Information Networking and Applications (AINA), 2013 IEEE 27th International Conference on* , vol., no., pp.492-499, 25-28 March 2013

[7] Khalfallah, Z.; Fajjariz, I.; Aitsaadiz, N.; Langar, R.; Pujolle, G., "2D-UBDA: A novel 2-Dimensional underwater WSN barrier deployment algorithm," in *IFIP Networking Conference (IFIP Networking), 2015* , vol., no., pp.1-8, 20-22 May 2015

[8] Amruta, M.K.; Satish, M.T., "Solar powered water quality monitoring system using wireless sensor network," in *Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on* , vol., no., pp.281-285, 22-23 March 2013

[9] Parmar, J.K.; Mehta, M., "A cross layered approach to improve energy efficiency of underwater wireless sensor network," in *Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on* , vol., no., pp.1-10, 18-20 Dec. 2014

[10] Dohare, Y.S.; Maity, T.; Paul, P.S.; Das, P.S., "Design of surveillance and safety system for underground coal mines based on low power WSN," in *Signal Propagation and Computer Technology (ICSPCT), 2014 International Conference on* , vol., no., pp.116-119, 12-13 July 2014

[11] Jin-ling Song; Heng-wei Gao; Yu-jun Song, "Research on Transceiver System of WSN Based on V-MIMO Underground Coal Mines," in *Communications and Mobile Computing (CMC), 2010 International Conference on* , vol.2, no., pp.374-378, 12-14 April 2010

[12] Longsheng Liu; Yue Li; Zhijun Zhang; Zhenghe Feng; Wenming Li; Da Zhang, "Experiment on underground propagation characteristic using CC110-based WSN," in *Antennas and Propagation Society International Symposium (APSURSI), 2013 IEEE* , vol., no., pp.1922-1923, 7-13 July 2013

[13] Xu Huping; Wu Jian, "Metal mine underground safety monitoring system based on WSN," in *Networking, Sensing and Control (ICNSC), 2012 9th IEEE International Conference on* , vol., no., pp.244-249, 11-14 April 2012

[14] Cammarano, A.; Spenza, D.; Petrioli, C., "Energy-harvesting WSNs for structural health monitoring of underground train tunnels," in *Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on* , vol., no., pp.75-76, 14-19 April 2013

[15] Guofang Dong; Bin Yang; Yang Ping; Wenbo Shi, "A secret handshake scheme for mobile-hierarchy architecture based underground emergency response system," in *Advanced Communication Technology (ICACT), 2015 17th International Conference on* , vol., no., pp.54-58, 1-3 July 2015

[16] Li Rong, "A study of the security monitoring system in coal mine underground based on WSN," in *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on* , vol., no., pp.91-93, 27-29 May 2011

[17] Li Zhang; Xunbo Li; Liang Chen; Sijia Yu; Ningcong Xiao, "Localization system of underground mine trackless facilities based on Wireless Sensor Networks," in *Mechatronics and Automation, 2008. ICMA 2008. IEEE International Conference on* , vol., no., pp.347-351, 5-8 Aug. 2008

[18] Anguita, D.; Brizzolara, D.; Ghio, A.; Parodi, G., "Smart Plankton: a Nature Inspired Underwater Wireless Sensor Network," in *Natural Computation, 2008. ICNC '08. Fourth International Conference on* , vol.7, no., pp.701-705, 18-20 Oct. 2008

[19] Ghelardoni, L.; Ghio, A.; Anguita, D., "Smart underwater wireless sensor networks," in *Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of* , vol., no., pp.1-5, 14-17 Nov. 2012

[20] Misra, S.; Ghosh, A., "The effects of variable sound speed on localization in Underwater Sensor Networks," in *Australasian Telecommunication Networks and Applications Conference (ATNAC), 2011* , vol., no., pp.1-4, 9-11 Nov. 2011

[21] Yong-sheng Yan; Hai-yan Wang; Xiao-hong Shen; Fu-zhou Yang; Zhao Chen, "Efficient convex optimization method for underwater passive source localization based on RSS with WSN," in *Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on* , vol., no., pp.171-174, 12-15 Aug. 2012

[22] Uribe, C.; Grote, W., "Radio Communication Model for Underwater WSN," in *New Technologies, Mobility and Security (NTMS), 2009 3rd International Conference on* , vol., no., pp.1-5, 20-23 Dec. 2009

[23] Abdou, A.A.; Shaw, A.; Mason, A.; Al-Shamma'a, A.; Cullen, J.; Wylie, S., "Electromagnetic (EM) wave propagation for the development of an underwater Wireless Sensor Network (WSN)," in *Sensors, 2011 IEEE* , vol., no., pp.1571-1574, 28-31 Oct. 2011

[24] Stuntebeck, E.P.; Pompili, D.; Melodia, T., "Wireless underground sensor networks using commodity terrestrial motes," in *Wireless Mesh Networks, 2006. WiMesh 2006. 2nd IEEE Workshop on* , vol., no., pp.112-114, 25-28 Sept. 2006

[25] G. Horvat, D. Šoštarić and D. Žagar, "Using radio irregularity for vehicle detection in adaptive roadway lighting," MIPRO, 2012 Proceedings of the 35th International Convention, Opatija, 2012, pp. 748-753.

[26] A. Alsayyari, I. Kostanic and C. E. Otero, "An empirical path loss model for Wireless Sensor Network deployment in an artificial turf environment," Networking, Sensing and Control (ICNSC), 2014 IEEE 11th International Conference on, Miami, FL, 2014, pp. 637-642

[27] Dadić, Ž., PREGLED KVALITETE PITKE VODE U HRVATSKOJ, Priručnik o temeljnoj kakvoći vode u Hrvatskoj, WaterLine, Quality water systems.

Electrical Field Intensity Model on the Surface of Human Body for Localization of Wireless Endoscopy Pill

Bojan Lukovac, Ana Koren, Antonija Marinčić, Dina Šimunić
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000
Zagreb, Croatia
E-mail(s): bojan.lukovac@ericsson.com, {ana.koren, antonija.marincic, dina.simunic}@fer.hr

***Abstract*—Constant struggle in medical science to develop new, more advanced methods of medicine application and reducing the invasiveness of a range of tests and procedures is present. For achieving this ambitious goal, it is necessary to find and develop new methods and technologies and to applicate them. The issue of localizing a range of devices inside the human body is one of them. Solution to this problem would enable the upsurge of nanorobots (nanobots) for micro-operations, application of strong medicine on limited areas to lessen its side effects, conducting the nanotoxicology reports or reducing the discomfort of some invasive procedures like endoscopy. In this paper, the focus is on the latter and an example of an endoscopy with a pill is given. Due to maximizing the effectiveness of endoscopy by the use of the pill there is a need for optimizing the technology used for its localization. SEMCAD tool is used in modelling of the scenario and executing the simulations.**

Keywords— WBAN; Electrical Field; Dipole; Endoscopy; Localization; SEMCAD

I. INTRODUCTION

Medical field is in a constant search for new, more advanced methods of medicine application and reduction of the invasiveness of a range of tests and procedures, one of which is endoscopy. Endoscopy is a procedure used for viewing the internal organs and vessels of a patient's body without making incisions. It is a common and powerful diagnostic tool for digestive diseases. Classical endoscopy is described as an extremely unpleasant and painful procedure which doesn't allow the exploration of parts of digestive tract (i.e. small intestine). The endoscopy, in which a pill is swallowed by the patient, travels through digestive system and documents it by taking photographs already exists, however it hasn't replaced the classical method. This is due to it being less exact as the localization of the pill (and any device in general) inside of human body is not precise enough due to several challenges. Current procedure of diagnosis via endoscopy pill consists of the doctor manually viewing all the images provided by the pill, which can sometimes be up to 100.000 images, depending on the manufacturer. Currently there is no automatic filtering of non-relevant images. When the image where there's an indication of potential problems is found, difficulty of deducing the exact location in the digestive tract presents itself, using solely the images surrounding it. Thus, the process is not precise enough to use in any operative procedures and it is usually necessary to repeat endoscopy with classical methods in order to confirm the diagnosis. Localization of the endoscopy pill implies determining its location, whether in real-time or retroactively, in respect to a referent point. It is expected for the solution to be portable, and not requiring the hospitalization of the patient for the time of procedure (time necessary for the pill to travel the digestive tract is approximately 12 hours). Furthermore, human body consists of bones and tissue which are arranged unevenly. Tissue has different density in different locations and is heterogeneous, which is why the issue of precise localization inside the human body remains open. In the next chapter (Chapter 2), the explanation of the selected antenna model is given, followed by the simulation and results of the model using SEMCAD (Chapter 3). Finally, conclusions are drawn in the Chapter 4.

II. PROPOSED ANTENNA MODEL

Endoscopic capsules are medical diagnostic devices 32 mm long and 11 mm in diameter consisting of a camera module, battery, antenna and transmitter along with additional sensors that depend on the manufacturer. They were developed in order to reduce the discomfort inherent to traditional endoscopic procedures and allow for detection of medical issues of the otherwise unobservable small intestine. Design of the antenna is an iterative process which demands multiple calculations with the goal of optimizing all the relevant parameters for working on the frequency specified. The antenna contained in the pill used in simulations is modelled in SEMCAD tool as a simple dipole antenna, consisting of two cylindrical parts with diameter of 1 mm, length of 10 mm and spacing of 1 mm. Antenna in the simulations is assumed to be a perfect electric conductor (PEC) in order to simplify the experiment. As shown in Figure 1, the antenna is sheathed with dielectric cylinder to protect it from contact with the surrounding tissue. Material used for the cylinder is Rogers

TMM-10 substrate which has electrical conductivity of 0.00087 S/m, relative permittivity 9.2 and density 1000 kg/m3).

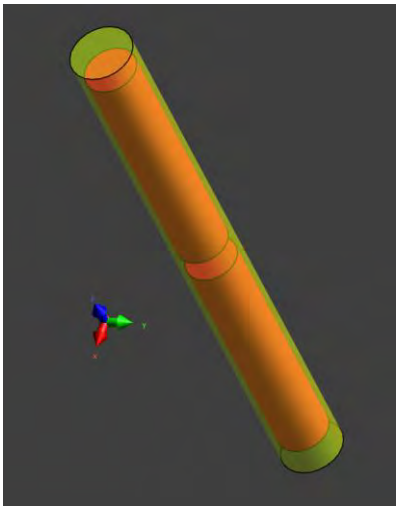


Figure 1. Isometric view of the antenna modelled in SEMCAD

Model of the pill is placed onto a several locations in the body, though the majority of the measurements are executed on one of the three locations.

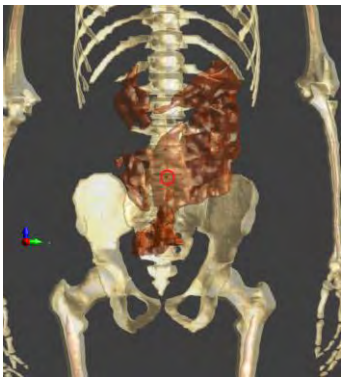
III. Simulation and Results

For the simulations, the software SEMCAD (Sim4Life) has been used, namely its most detailed model of the human body, Duke – a 36 year old man. First location of the antenna is in the large intestine in position surrounded by the pelvic bones (Figure 2a). Second location is the lumen of small intestine in the proximity of navel, on the front end part of the abdomen, which is a location that enables us to observe the difference in attenuation between front part of the body (with just a thin layer of muscle, fat and skin) and the back part (where attenuation is caused by intestine, organs, spine, muscles, fat and skin). This is shown in Figure 2b. Final location measured is the middle of the stomach lumen, which is characterized by the presence of the rib cage and multiple organs which significantly influence the strength of the signal (Figure 2c). These locations are selected as they represent well the various conditions inside the human's body. Furthermore, in these three locations, measurements have been performed with harmonic signal and Gauss impulse; the pill has been rotated parallel to all three axes and; the measurements in the space diagonal have been performed. The intensity of the electrical field and Poynting vector has been measured. Poynting vector was selected as an additional method due to practical reasons.

a)



b)



c)



Figure 2. Locations of the antenna marked with a red circle: pelvic bones (2a), lumen of small intestine, (2b) and middle of stomach lumen (2c)

Expected diagram of dipole antenna radiation in free space is shown in Figure 3. Areas in the red specter present lower power while the green areas present higher relative power. All the simulations have been performed using the antenna transmitting at 2.4 GHz frequency, and the source characteristics being 5 V voltage and internal resistance 50 Ω.

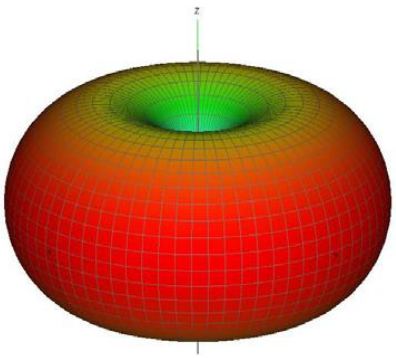
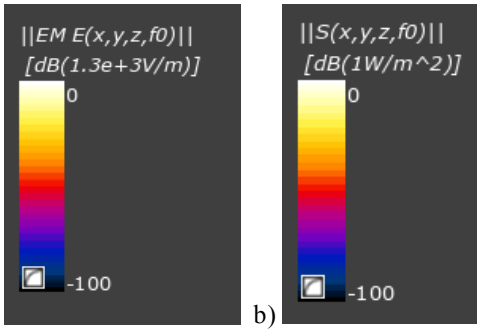


Figure 2. Diagram of dipole radiation in free space

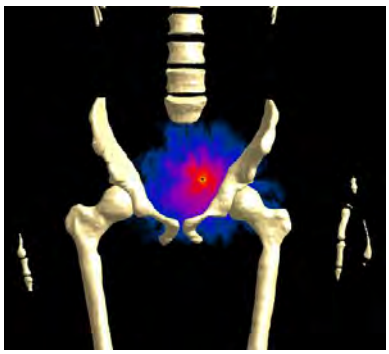
Following figures (Figure 4a and 4b) show the legend for the following simulations (Figures 5 through 7).



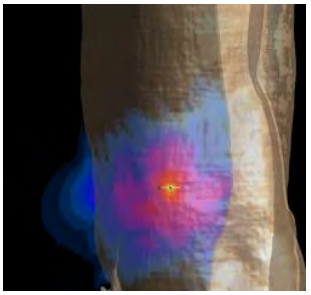
a) b)

Figure 4. Reference values of electrical field intensity (left) and Poynting vector value (right; legend applicable for all simulations)

Furthermore, the radiation of the dipole can be observed on the following figures, as well as the position (localization) of the pill. In Figures 5a and 5b it is clear that the intensity of the electrical field suffers from strong attenuation caused by the pelvis bones.



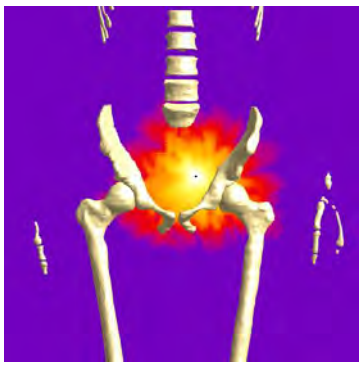
a)



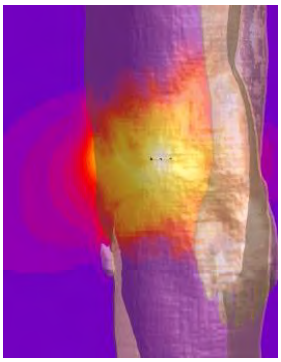
b)

Figure 5. Electrical field intensity (front and side view)

As an additional method, Poynting vector was measured. Results for the same scenario are given in Figures 5c and 5d below.



a)



b)

Figure 5. Poynting vector value in the first scenario (front and side view)

This leads to the following conclusions. First, electrical field intensity is less indicative than the value of Poynting vector due to lesser loss of intensity inside the body. Secondly, the area in the pelvis region strongly affects the radiation of dipole antenna; pelvis bones cause reflection and attenuation which causes the diagram of radiation starts to look almost isotropic. The results of the second simulation scenario are presented in Figure 6.

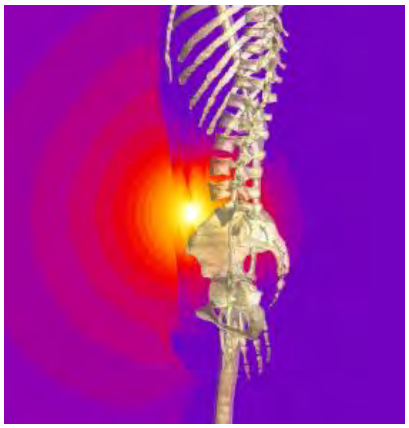
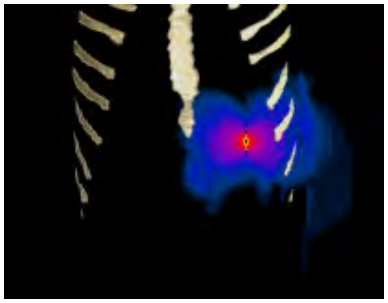


Figure 6. Poynting vector value in the second scenario (side view)

Again, several conclusions can be drawn. The proximity of the radiation source to skin causes propagation of electromagnetic wave and distorts the radiation diagram. Organs, bones and tissue cause the expected loss of wave strength in all other directions. The surroundings have much larger influence on the radiation diagram than the orientation of the pill. Thus, precise localization of the pill would require an algorithm which would define measurements on a large number of locations in a body so it would be possible to eliminate the influence of the wave propagation in the skin. Results of the third and last scenario are given below in Figures 7a and 7b.

a)



b)

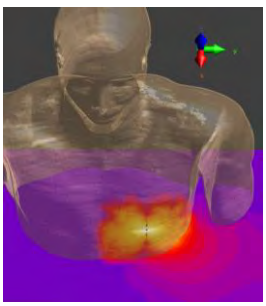


Figure 7. Electrical field intensity and Poynting vector value in the third scenario

The stomach has proved to be the best location for all tests of dipole antenna since the result received has the expected doughnut diagram of dipole antenna radiation, most likely due to the homogenic nature of the stomach lumen where the antenna is located. The importance of this result is great since the recognition of the expected value potentially enables the assessment of the pill's orientation.

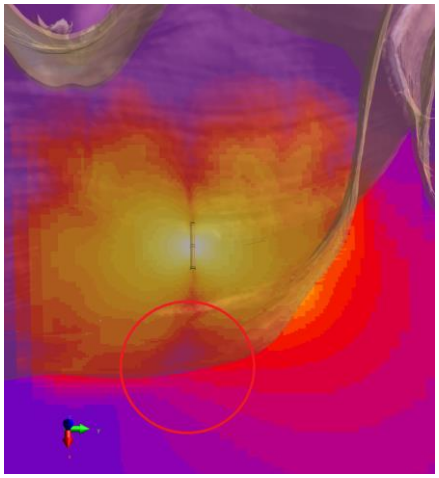


Figure 8. Radiation diagram distortion on the skin surface

Closer analysis of the radiation gives an insight into specific interactivity of the wave and skin and is shown on Figure 8. On the area, marked with a red circle, it is clearly visible how weak the signal is in the area between the pill and skin (in line with the expectations) but then greatly increases through the skin surface, due to refractive index between air and skin. Thus, it would be advisable to perform the measurements at approximately 1 cm distance from the skin surface.

IV. CONCLUSION

The issue of localizing a range of devices inside the human body is extremely relevant in today's medical science as the solution to this problem would enable the upsurge of nanorobots (nanobots) for micro-operations, application of strong medicine on limited areas to lessen its side effects, conducting the nanotoxicology reports or reducing the discomfort of some invasive procedures, endoscopy included. Several conclusions were drawn from the past simulations in SEMCAD. The greatest impacts on the propagation of electromagnetic wave inside human body have transitions between bones and other tissue and transitions between skin and air. Low electrical conduction and density of the bones has the largest effect on the wave propagation having in mind the ratio of bone volume in the total body volume. Choice of antenna is of crucial importance. The Rogers TMM-10 substrate cylinder sheathed dipole antenna gives realistic radiation diagrams. In some of the simulated scenarios (locations) the parameters of the environment (surrounding area) impacts the radiation diagram severely, to the point where it is impossible to draw conclusions of the pill position or orientation. Thus, it is necessary to develop algorithms which would take into account the environment and eliminate its effects (e.g. wave propagation through skin) in

order to ensure precision. Furthermore, measurements should be conducted without skin contact. Point of refractive index between skin and air deforms radiation diagram and makes the assessment of the position and orientation of the antenna more difficult. Hence, it is advisable to move the measuring instrument at least 1 cm away from the skin surface to avoid diagram deformation. Lastly, graphs of electrical field intensity and Poynting vector both describe similar radiation diagrams. However, due to, by far, easier and more practical measuring of the received power (which is directly proportional to the Poynting vector) using a sensor in comparison with measuring the electrical field intensity it is advisable to use measurements of Poynting vector (i.e. received power) for calculations.

REFERENCES

[1] Penazzio M. et al., Small-Bowel Capsule Endoscopy and Device-Assisted Enteroscopy for Diagnosis and Treatment of Small-Bowel Disorders: European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline, Endoscopy 2015 nr. 47, 2015., pp. 352.-376.

[2] Ladas, S.D: et al. European Society of Gastrointestinal Endoscopy (ESGE): Recommendations on Clinical Use of Video Capsule Endoscopy to Investigate Small-Bowel, Esophageal and Colonic Diseases, Endoscopy 2010 nr. 42, 2009., pp. 220.-227.

[3] Woods, S.P., Constadinou, T.G., Wireless Capsule Endoscope for Targeted Drug Delivery: Mechanics and Design Considerations, IEEE Transactions on Biomedical Engineering, Vol. 60, nr. 4, April 2013., pp. 945.

[4] Yuce, M., Dissanayake, T., Easy-to-Swallow Antenna and Propagation, IEEE Microwave Magazine, nr. 14, June 2013., pp. 74-82.

[5] Wen-cheng Wang et al., Experimental Studies on Human Body Communication Characteristics Based Upon Capacitive Coupling, Body Sensor Networks (BSN), 2011 International Conference on, Dallas, USA, May 2011., pp. 180 – 185.

[6] Sazonov, E., Neuman, M. R., Wearable Sensors: Fundamentals, Implementation and Applications, Academic Press – Elsevier, San Diego, USA, 2014., pp. 462.

[7] Francesco Merli et al., Design, Realization and Measurements of a Miniature Antenna for Implantable Wireless Communication Systems, IEEE Transactions on Antennas and Propagation, Vol. 59, nr. 10, October 2011., pp. 3544-3555.

[8] Vallejo, M., et al., Accurate Human Tissue Characterization for Energy-Efficient Wireless On-Body Communications, Sensors (Basel) Vol. 13., nr. 6., 2013., pp. 7546.-7569.

Wide band current transducers in power measurment methods - an overview

Roman Malarić, Željko Martinović *, Martin Dadić, Petar Mostarac, Žarko Martinović **
Faculty of Electrical Engineering and Computing/
Department of Electrical Engineering, Fundamentals and Measurements
Unska 3, 10000 Zagreb, Croatia
*COMBIS / Baštijanova 52A, Zagreb, Croatia
**Danieli Systec, Katuri 17, Labin, Croatia
e-mail:, roman.malaric@fer.hr, zeljko.martinovic@combis.hr, martin.dadic@fer.hr, petar.mostarac@fer.hr
z.martinovic@systec.danieli.com

Precise power measurement is usually done at national metrology institutes. In recent years researchers have devoted substantial time toward a measurement of power under non-sinusoidal conditions, especially in power grids because of renewable energy increase. The challenge is not only to measure power at 50 Hz, but also to measure harmonics and inter harmonics. To measure power at needed accuracy, accurate and frequency independent current and voltage transducers are needed up to 100 kHz. In this paper an overview and state of the art is given for current transducers used for this purpose. This mainly includes AC shunts of coaxial design, calculable resistors, precise transformers and methods for their characterization.

I. INTRODUCTION

In order to transform the currents to voltages suitable for analog to digital conversion (usually 1 V RMS) to obtain power and energy, precise current shunts or shunt/transformer combination are used for this purpose. This, in particular, is important for the measurement of power and energy under distorted waveform conditions. Current shunts therefor represent one of the most important parts of power measuring system (fig.1.).

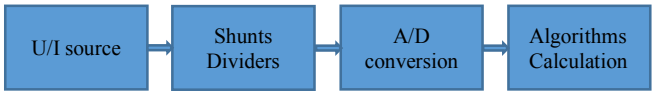


Fig. 1. Precise power measurement system

II. CURRENT SHUNTS

Coaxial shunts have a better frequency behavior than shunts which are not coaxially connected [1], and smaller AC-DC current transfer differences up to 30 kHz or even 100 kHz. Shunt design have been mostly based on work done at Dimitry Ivanovich (VNIIM) but there have been few exceptions and different designs. The most often used resistors are low inductance metal foil type. NRC [2] shunts for currents up to 200 mA are a star configuration of resistors soldered on a single-sided printed circuit board. For the current ranges from 0.5 to 10 A, the shunts are Mendeleev type built of three plates connected by a number of ribs. The plates and the ribs are cut from the double-sided copper-clad boards. The shunt resistor consists of a number of Vishay type S102C metal film resistors, concentrically mounted between the output and the middle plates, and connected in parallel by the copper layers of these plates. The input plate is removed from the two output plates by the length of the rib switch minimizes the inductive coupling between the input and the output circuits. This design is similar to the SP type [3][4] which are designed from 50 mA to 100 A, CSIRO shunts [5] and SIQ shunts [6] which are coaxial type ranging from 100 mA to 20 A, with frequency range from 10 Hz to 30 kHz. The input and output sides are separated by 17-cm-long crossbars that are made of double-sided printed circuit boards. The number of crossbars and the diameter of the circular plates depend on the current shunt's nominal current. CMI shunts [7] are also cage type designed to be well suited for use with planar multi-junction thermal converters (PMJTCs). The voltage drop across every shunt (30 mA to 10 A) in parallel with a 90 PMJTC is 1 V at nominal current. A calculable model was developed for these shunts using lumped circuit elements. Temperature coefficients were reduced by using a suitable combination of resistors types (S102C, S102K and Z201). New generation of shunts were produced based on results obtained from the analysis of the lumped element model of the original shunt [8]. At BEV [9] a manganine foil is wound as the resistive element around a cylindrical core made of glass fiber epoxy. One copper foil at the outside brings the low potential back to the current input, another copper foil underneath the manganine brings the high potential from the input to the output side. All foils are soldered onto copper discs which are mechanically connected to the core. This principle applies for shunts in the range from 2 A to 100 A, except that for shunts from 2 A to 10 A the core is made of a brass tube instead of copper foil. Shunts from 100 mA to 1 A is made with metal film resistors soldered in parallel in a 4-terminal arrangement with a small loop in both the current and potential path. The output voltage at nominal current is 0,4 V. INRIM [10] current shunts cover the range from 5 mA to 20 A with nominal voltage of 1 V to cover the input voltage of planar multi-junction thermal converter (PMJTC). The shunts with value from 5 mA to 100 mA have a single SMD resistor connected between the central

conductor and a return wire at the external coaxial screen. Shunts up to 2 A are made with a disk of a double-sided printed circuit board with conductive layer removed from one side except for the edge, and on the other side, a ring is machined out of the conductive layer. The conductive parts and the external edges that are connecting the two sides are coated with gold, with 10 SMD resistors soldered across to the insulating ring. For ranges up to 20 A the disks are made of a thermally conductive ceramic material 10 mm thick with a diameter of 80 mm. The resistive parts are made with manganin foils that are fixed by thermally conductive adhesive resin. The foils are then patterned to the proper shape. At JV[11], a new type of shunt for ac–dc current transfer was developed for frequencies from 10 to 100 kHz and current ranges of 30 mA–10 A with uncertainties smaller than ±9 μA/A using surface-mount resistors of the cylindrical metal-electrode face-bonding (MELF) type which are cheap and easily available with temperature coefficients below 10 μΩ/Ω per K. The shunts have very low and calculable ac–dc current transfer differences in the frequency range of 10 Hz–100 kHz. INTI [12] presented three different shunt designs for AC-DC current transfer which has been evaluated. The nominal current values are 5 A and 10 A with nominal output of 1 V. One of the shunt is similar to cage type design and two other differ from known designs, however the paper lacks the comparison results. Commercially available shunts include Fluke A40B series, Guildline 7340 and 7350 series, Transmille, and some of the already mentioned in this paper from National Metrology Institutes. A new AC current shunt was also designed at NIM that uses minalpha wires instead of commercial available resistors [13]. Shunt showed good results up to 100 kHz.

A. Use of transformers to lower current level

Sometimes current transformers are used to lower the currents to acceptable levels and are then burdened with low current shunts. INMETRO [14] developed a special two-stage transformer, which is coupled to a standard shunt of 10 Ω. At NIM [15] a set of shunts is used for current ranges of 0.1, 0.2, 0.5, 1, 2, 5, 10, and 20 A, while a current transformer is adopted for 50 A. The full-scale output signals of the voltage dividers and the shunts are 0.8 V regardless of the voltage/current range, which are the input signals of the corresponding two DVMs. INTI [16] developed current transformers with 100 mA output value which are used together with 10 Ω AC shunt to obtain 1 V output. At PTB Primary AC power standard [17] current transformers together with AC shunt is used as well. INRIM [18] current to voltage converter is made up of a double-stage current transformer with a precision 1 Ω AC resistor connected to its output.

B. High current shunts

Recent developments have seen the realization of 100 A shunt with considerable uncertainty improvement so that a current transformer is not needed any more up to 100 A [19] that have AC-DC difference below 20μA/A up to 100 kHz or with [4] AC-DC difference of -36 μA/A up to 100 kHz. These are designed in similar manner like the lower value shunts using Mendeleev cage type. NRC high current shunts [20] have been produced in cooperation with BEV. Each shunt consists of three coaxial cylinders mounted between four copper plates on a glass fiber epoxy cylinder. The middle cylinder, the resistor of a shunt is manufactured from a thin manganine foil. The AC-DC difference is lower than 100 μA/A at 100 kHz. BEV Institute [21] also manufactures high current TEE connector needed for calibration of these shunts. There are also some commercial shunts like Guidline 7340 and Fluke A40B that are available up to 100 A.

C. Calculable AC-DC resistors

The AC-DC difference of shunts can be obtained using the modelling the shunts with lumped elements and calculating the frequency response and by calibrating the shunts with AC bridge which needs a calculable resistance standard or comparing it against thermal voltage transfer standards. Several types of the calculable ac/dc resistors have been presented in literature as coaxial, bifilar, quadrifilar and octofilar designs. The time constant for coaxial and bifilar resistors was analyzed [22]. Usually a simple geometrical structure is used. As for more recent research, the calculable coaxial resistor was designed [23] based on the coaxial line with a cylindrical shield which can be described by relatively simple equations for the real and imaginary parts of the impedance. The resistors consist of a straight even ohm resistance wire surrounded by a 51 mm diameter coaxial brass case with nominal value of 1000 Ω. However, the resistor was evaluated only at one frequency of 1592 Hz with good results. In an inter-comparison [24] of calibration systems for AC shunts up to audio frequencies (10 kHz) between the NRC, JEMIC, and NIST is presented. The comparison was implemented with a calculable transfer ac/dc shunt, designed by JEMIC at frequencies up to 10 kHz in each laboratory. Both laboratories use for that purpose its own calculable AC-DC resistors and AC bridges. The NRC uses two calculable ac/dc quadrifilar resistors of 100 and 1000 Ω described by Gibbons [25]. JEMIC [26] and NIST [27] compares a 0.1 Ω bifilar reference current shunt with calculable amplitude and phase response. A new design for calculable coaxial resistors [28] with the values of 1 kΩ, 10 kΩ and 12,906 kΩ have been modeled by using Mathcad computer program and evaluated in. Also, a new coaxial resistors [29] for use with quantum Hall effect-based impedance measurements was designed at PTB and CMI. The improvement compared with existing resistors is in easier handling of inner resistive wire from the shield.

D. AC Shunt modelling

Shunts differ from calculable AC-DC resistors as they are not so simple to model having many elements such as ribs, plates and resistors. Some of the shunts were just calibrated using thermal voltage converters, some have been evaluated using simple models, and some of the shunts have been modeled using detailed equivalent schemes. A calculable model [30] for Mendeleev type shunt was developed using lumped circuit elements which can be used to calculate trans impedance, ac–dc difference, and the phase angle error of a shuns. It is based only on calculations of all component

values from the geometry and material properties (except resistors). Compared with measured results the difference is less than 6 μΩ/Ω in the AC-DC difference and 110 μrad in the phase angle error at frequencies up to 100 kHz. SIQ shunts [31] model was made using lumped element modeling and then analytical transfer function and input impedance of the shunt were derived from the model and used to calculate frequency response and input impedance as a function of frequency. These calculations were then compared to the calibrated values of AC/DC difference and against measured input impedance. Simplified model of BEV foil shunts [32] have been presented. Also, the simplified circuit of INRIM shunts [10] have been presented but without mathematical modeling. NRC shunts [2] have been modeled assuming that shunt consists of several identical components. The input and output front plates were modeled as R−C−R T-networks, the ribs as (R + L−C−R + L) T-networks, and the shunt resistor as (R + L)|| C assuming that the shunt is a two terminal-pair device. The model, even though too imprecise for a theoretical characterization of a "calculable" shunt was used in selecting the length of the ribs. The longer ribs increase the distance between the input and the output, thus reducing the coupling between the two circuits, which is a major source of the shunt ac–dc difference. Shorter ribs, which were eventually used in the design with lower capacitance, decrease the internal capacitance of the shunt. The number of ribs and resistors is again a compromise as with increasing the number of ribs increases the nominal power rating of the shunt but also increases the internal capacitance. JV shunts [11] are modeled using lumped circuit elements. The model includes approximations to the parasitic reactive and resistive components from the chosen shunt geometry. Component values are measured and/or calculated from the geometry and material properties. Finally, important paper is the analysis of shunt TVC [33] combination which was done. This paper describes the relationship between the overall ac-dc differences of a shunt thermal-converter combination.

E. Measurement of AC-DC current difference of shunts

An AC-DC difference is defined as:

$$\delta_{AC-DC} = \frac{V_{AC}-V_{DC}}{V_{DC}} \qquad (1)$$

The AC-DC current transfer difference is the most important characteristics of AC shunt and can be measured in three ways: inter-comparison with reference shunt, calibrated with thermal voltage standards and by measuring it with precise AC bridge. For example, SIQ shunts have been compared to current shunts [6] with known AC-DC. The AC-DC difference was less than 20 μΩ/Ω for frequencies up to 30 kHz and currents up to 20 A. However to properly characterize current shunts it is necessary to compare them with thermal voltage converters which have a flat AC-DC characteristics up to 100 kHz. At CMI [34] step-up procedure is used to measure AC-DC current transfer difference from 1 mA up to 10 A in the frequency range 10 Hz – 100 kHz by comparing it with PMJTC 10 mA AC standard. The AC-DC difference was less than 50 μΩ/Ω for all frequencies and currents. The same method was also used to determine frequency characteristics of NRC shunts [20]. The AC-DC current difference for in-house built shunts from 100 mA to 10 A is less than 100 μΩ/Ω up to 100 kHz, but for commercial shunts from 30 A up to 100 A the AC-DC difference is quite larger particularly at frequencies above 10 kHz. JV shunts [11] were evaluated for AC-DC difference using PMJTC in a digital bridge setup for comparing the ac–dc current differences is similar to the system described by Rydler [35]. A complete step-up from 10 mA to 10 A was made where the unknown shunt/PMJTC for the next higher current is calibrated at the current level of the known converter and then used at its rated current. At 10 mA, the shunt/PMJTC was calibrated directly against the primary reference for the ac–dc current difference. The other steps were 30 mA, 100 mA, 300 mA, 1 A, 3 A, 5 A, and 10 A. The AC-DC difference was determined to be less than 20 μΩ/Ω for currents from 50 mA to 10 A and frequencies up to 100 kHz. As the frequency characteristic of a shunt–thermal converter combination depends on the frequency characteristic of the shunt [2] the frequency characteristic of the thermal converter, and a mutual inductance between the shunt and the thermal converter which can be neglected for this particular shunt design, and quit some simplifications. The AC-DC transfer difference of a shunt–thermal converter combination δ_i can be presented [32] as follows:

$$\delta_i \approx \frac{R_{TVC}}{R_S+R_{TVC}} \times \delta_V + \frac{R_S}{R_S+R_{TVC}} \times \delta_I + \delta_R \qquad (2)$$

from where the AC-DC current difference for shunts can be calculated. In addition AC-DC difference of current shunts can be evaluated using different [36] AC bridges and resistors with calculable AC response.

F. Phase angle error

The phase angle error calculation from the model can be validated by measuring it, even though it is a difficult task as the shunts are designed to have very low phase angle error. Most of the commercial shunt do not even specify phase angle error. Measurement of shunt phase angle errors have been performed [37] in various ways in SP research center. Four terminal inductance are determined using an LCR-meter and calculable inductance standards and calculated phase angel errors are then verified at frequencies up to a few kHz by comparison with digital sampling wattmeter [38] for shunts from 0,05 A to 10 A. Later [39] it was extended 1 MHz using the phase comparator comprising of fast and accurate digitizers. The agreement between the different methods, for measurement of phase angle difference of two shunts, was within a few prad at 1500 Hz. Phase angle errors of shunts [40] with rated currents at 50 and 100 A at frequencies from 25 to 100 kHz have been determined using a three-branch binary inductive current divider to measure the phase angle errors of high current shunts against a phase angle reference standard with only one step. Later [41] it was extended to 200 kHz. A new method [42] was described to determine the

phase angle errors of ac shunts by measuring the inductance and distributed capacitance. For this purpose several units have been developed: 1 Ω shunt of coaxial design as the time constant standard, a coaxial inductor with identical structure as the time constant standard and a four-terminal mutual inductor for the measurement of the inductance of the time constant standard.

In [43] a method is proposed to determine phase angle of shunts using group of micro potentiometer resistors designed at CSIRO whose phase angle error can be described, in the first approximation with a formula. With the buildup process the phase angle can be determined for currents from 100 mA up to 20 A and frequencies from 40 Hz to 200 kHz. Level dependence of phase angle error method have been proposed [44] because in the build-up procedures, the unknown shunts for the next higher rated current are calibrated at the current level of the known shunts, and then it is used at its rated current. A wideband phase comparator has been developed at INRIM for high current shunts [45]. The two-input digital phase detector is realized with a precision wideband digitizer connected through a pair of symmetric active guarded transformers to the outputs of the shunts under comparison. The system is suitable for comparing shunts in a wide range of currents, from several hundred of mill ampere up to 100A, and frequencies ranging between 500Hz and 100 kHz. The system has been used for international comparison of current shunts [46] from 10 A to 100 A.

G. DC characterization of shunts

To properly characterize AC shunts and calculate measurement uncertainty contribution of shunts in power measurement, in addition to phase angle errors and AC-DC difference it is also necessary to evaluate shunt performance for power coefficient (level dependence), temperature coefficient and drift of DC resistance of shunt.

Level dependence of step up procedure has been [47] evaluated. The step-up procedure is based on the assumption that the ac–dc difference of a shunt and thermal converter combination is the same at two current levels used in the step-up procedure and can produce systematic errors that are added in each step and can be considerable at the end of the chain. The authors argue that because of their design, it was not expected to find an appreciable low frequency level dependence of the shunts but in PMJTC. In this paper TEE connector used to connect two shunts has been analyzed as well. The level dependence of two different shunts have been also evaluated [48] in order to reduce the contribution of the current level dependence in the uncertainty budget.

The temperature coefficients (TCR) of the shunts and drift [6] were evaluated. The temperature coefficient (TCR) was measured in a temperature chamber, where the temperature was set to 23 °C, 18 °C, and 28 °C. This temperature change was large enough to determine the TCRs at the shunt's working point. The TCRs proved to be linear in the 18 °C–28 °C temperature range within the measurement uncertainty. The temperature rise due to the load current was calculated based on the resistor power coefficient specification. Drift was during several month periods by comparison with the known reference resistor with the direct current comparator resistance bridge. Measurements showed that this drift was fairly linear with time and that a linear drift slope could be calculated for each current shunt. Drift was measured from 2,8 μΩ/Ω to 13,8 μΩ/Ω per year for all shunts. Complete DC characterization of shunts in ppm level have been conducted [49]. The set of four working resistance standards (0,1 Ω - 0,001 Ω) and Fluke 8508A reference multi-meter have been used for comparison purpose. The system established in CMI which also includes oil bath and thermostatic chamber was used to characterize foil and cage shunts up to 100 A. The second paper [50] gives full DC characterization of AC shunts in the range of 30 mA to 10 A. The characterization includes the long-term drift, the temperature coefficient, and the power coefficient, which all appear to have effects on the level of 1-10 μΩ/Ω. The shunt investigated was designed and built by JV [11]. In table 1, current and frequency ranges for all shunts manufactured at different national metrology institutes and companies are summarized:

Table 1: Available current shunts

Manufacturer	Current range [A]
SIQ	100 μA – 100 A
NRC	1 mA–100 A
INRIM	5 mA – 20 A
SP	10 mA - 100 A
CMI	30 mA - 10 A
JV	30 mA to 10 A
CSIRO	100 mA - 20 A
BEV	10 mA to 100 A
VNIIM	1 A – 10 A
INTI	5 A – 10 A
Fluke	1 mA – 100 A
Transmille	1 mA – 100 A
Guildline	10 mA – 100 A

III. CONCLUSION

In this paper an overview of wide band current transducers is presented. All available papers presenting the available AC shunts and transformers are included, and commercial AC shunts are also mentioned. A special section has been added that presents the methods and procedures for the calibration and characterization of shunts and transformers in respect to their AC-DC difference, phase angle error and DC characterization.

IV. ACKNOWLEDGEMENT

In the name of our research group (Department of Electrical Engineering Fundamentals and Measurements from Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia) I would like to thanks Croatian Science Foundation for their support and motivation. This work has been fully supported by Croatian Science Foundation under the project Metrological infrastructure for smart grid IP-2014-09-8826.

LITERATURE

[1] M. M. Halawa, Implementation and Verification of Building-up AC-DC Current Transfer Up to 10 A, International journal of electrical engineering & technology (IJEET), 2012, pp. 262 – 273

[2] P. S. Filipski, and M. Boecker, AC–DC Current Shunts and System for Extended Current and Frequency Ranges, IEEE Transactions on Instrumentation and Measurement, 2006, pp. 1222 – 1227

[3] S. Svensson and K. E. Rydler, A Measuring System for the Calibration of Power Analyzers, IEEE Transactions on Instrumentation and Measurement, 1995, pp. 316 – 317

[4] K. E. Rydler and V. Tarasso, Extending ac-dc current transfer measurement to 100 A, 100 kHz, Precision Electromagnetic Measurements Digest, 2008, pp. 28 - 29

[5] I. Budovsky, A micropotentiometer-based system for low voltage calibration of alternating voltage measurement standards, Precision Electromagnetic Measurements Digest, 1996, pp. 497 - 498

[6] B. Voljc, M. Lindic, and R. Lapu, Direct Measurement of AC Current by Measuring the Voltage Drop on the Coaxial Current Shunt, IEEE Transactions on Instrumentation and Measurement, 2009, pp. 863 – 867

[7] V. N. Zachovalova, AC-DC current transfer difference in CMI, Precision Electromagnetic Measurements Digest, 2008, pp. 362 – 363

[8] V. N. Zachovalova, M. Síra and P. Bednari, New generation of cage-type current shunts at CMI, 20th IMEKO TC4 International Symposium, 2014, pp. 59 – 64

[9] M. Garcocz, P. Scheibenreiter, W. Waldmann, and G. Heine, Expanding the measurement capability for AC-DC current transfer at BEV, Precision Electromagnetic Measurements Digest, 2004, pp. 461-462

[10] U. Pogliano, G. C. Bosco, and D. Serazio, Coaxial Shunts as AC–DC Transfer Standards of Current, Precision Electromagnetic Measurements Digest, 2008, pp. 30-31

[11] K. Lind, T. Sørsdal and H. Slinde, Design, Modeling, and Verification of High-Performance AC–DC Current Shunts From Inexpensive Components, IEEE Transactions on Instrumentation and Measurement, 2008, pp. 176 - 181

[12] L. D. Lillo, H. Laiz, E. Yasuda and R. García, Comparison of three different shunts design for AC-DC current transfer, 17th IMEKO TC4 International Symposium, 2004,

[13] J. Zhang, X. Pan, H. Huang, L. Wang and D. Zhang, A coaxial AC shunt with calculable AC-DC difference, Precision Electromagnetic Measurements (CPEM), 2010, pp. 597 – 598

[14] A. M. R. Franco, E. Tóth, R. M. Debatin, and R. Prada, Development of a power analyze, 11th IMEKO TC-4 Symposium, Trends in electrical measurements and instrumentation, 2001, pp 1-5

[15] L. Zuliang, W. Lei, L. Min, L. Lijuan, and Z. Hao, Harmonic Power Standard at NIM and Its Compensation Algorithm , IEEE Transactions on Instrumentation and Measurement, 2009, pp. 180-187

[16] R. Carranza, S. Campos, A. Castruita, T. Nelson, A. Ribeiro, E. So, L. D. Lillo, A. Spaggiari, D.Slomovitz, D. Izquierdo, C. Faverio, H. Postigo, H. Díaz, H. Sanchez, J. Gonzalez and Á. Z. Triana, Precision Electromagnetic Measurements (CPEM 2014), pp. 302-303

[17] E. Mohns, G. Ramm, W.G.K. Ihlenfeld, L. Palafox and H. Moser, The PTB Primary Standard for Electrical AC Power , MAPAN - Journal of Metrology Society of India, 2009, pp. 15 - 19

[18] U. Pogliano, Use of Integrative Analog-to-Digital Converters for High-Precision Measurement of Electrical Power, IEEE Transactions on Instrumentation and Measurement, 2002, pp. 1315 - 1318

[19] B. Voljc, M. Lindic, B. Pinter, M. Kokalj, Z. Svetik, and R. Lapuh, "Evaluation of a 100 A current shunt for the direct measurement of AC current, IEEE Transactions on Instrumentation and Measurement, 2013, pp. 1675–1680

[20] P. S. Filipski and M. Boecker, AC-DC current transfer standards and calibrations at NRC, Simposio de Metrología, Mexico, 2006

[21] M. Garcocz, P. Scheibenreiter, W. Waldmann, G. Heine, Expanding the measurement capability for AC-DC current transfer at BEV, Precision Electromagnetic Measurements Digest, 2004, pp. 461-462

[22] H.Fujiki, A.Domae and Y-Nakamura, Analysis of the time constant for the bifilar calculable ac/dc resistors, Precision Electromagnetic Measurements, 2002, pp. 342-343

[23] R. E. Elmquist, Calculable Coaxial Resistors for Precision Measurements, Instrumentation and Measurement Technology Conference, 1999, pp. 1468 - 1471

[24] E. So, D. Angelo, T. Tsuchiyama, T. Tadokoro, B. C. Waltrip, and T. L. Nelson, Intercomparison of Calibration Systems for AC Shunts Up to Audio Frequencies, IEEE Transactions on Instrumentation and Measurement, 2005, pp.507 – 511

[25] D. L. H. Gibbings, A design for resistors of calculable a.c./d.c. resistance ratio, Electrical Engineers, Proceedings of the Institution, 2010, pp. 335 - 347

[26] T. Tsuchiyama and T. Tadokoro, Development of a high precision AC standard shunt for AC power measurement, Precision Electromagnetic Measurements, 2002, pp. 254 - 255

[27] O. B. Laug, T. M. Souders, and B. C. Waltrip, A Four-Terminal Current Shunt With Calculable AC Response, National Institute of Standards and Technology, 2004, pp. 1-56

[28] Y. Giilmez, G. Gulmez, E. Turhan, T. Ozkan, M. Cinar, L. Sozen, A new design for calculable resistor, Precision Electromagnetic Measurements, 2002, pp. 348 – 349

[29] J. Kucera, E. Vollmer, J. Schurr and J. Bohacek, Calculable resistors of coaxial design, Measurement science and technology, 2009,

[30] V. N. Zachovalova, On the Current Shunts Modeling, IEEE Transactions on Instrumentation and Measurement, 2014, pp. 1620-162

[31] B. Pinter, M. Lindic, B. Voljc, Z. Svetik, and R. Lapuh, Modeling of AC/DC current shunts, Precision Electromagnetic Measurements (CPEM), 2010, pp. 599–600.

[32] M. Garcocz, P. Scheibenreiter, W. Waldmann, and G. Heine, Expanding the measurement capability for AC-DC current transfer at BEV, Precision Electromagnetic Measurements Digest, 2004, pp. 461–462

[33] J. R. Kinard, T. E. Lipe, and C. B. Childers, AC-DC difference relationships for current shunt and thermal converter combinations, Precision Electromagnetic Measurements, 1991, pp. 352–355

[34] V. N. Zachovlova, M. Sira and J. Streit, Current and frequency range exstension of AC-DC current transfer difference measurement system at CMI, Precision Electromagnetic Measurements (CPEM), 2010, pp. 605 – 606

[35] K. E. Rydler, High precision automated measuring system for AC–DC current transfer standards, IEEE Transactions on Instrumentation and Measurement, 1993, pp. 608–611

[36] E. So, D. Angelo, T. Tsuchiyama, T. Tadokoro, B. C. Waltrip, and T. L. Nelson, Intercomparison of Calibration Systems for AC Shunts Up to Audio Frequencies, IEEE Transactions on Instrumentation and Measurement, 2005, pp. 507 – 511

[37] K.E. Rydler and V. Tarasso, A method to determine the phase angle errors of an impedance meter, Precision Electromagnetic Measurements Digest, 2004, pp. 123-124

[38] S. Svensson, K.E. Rydler, and V. Tarasso, Improved model and phase-angle verification of current shunts for ac and power measurements Precision Electromagnetic Measurements Digest, 2004 p. 82-83

[39] K. E. Rydler, T. Bergsten and V. Tarasso, Determination of Phase Angle Errors of Current Shunts for Wideband Power Measurement, Precision Electromagnetic Measurements (CPEM), 2012, pp. 284-285

[40] X. Pan, J. Zhang, H. Shao, W. Liu, Y. Gu, X. Ma, B. Wang, Z. Lu, and D. Zhang, Measurement of the Phase Angle Errors of High Current Shunts at Frequencies up to 100 kHz, IEEE Transactions on Instrumentation and Measurement, 2013, pp. 1652 - 1657

[41] J. Zhang, X. Pan, W. Liu, Y. Gu, B. Wang, and D. Zhang, Determination of Equivalent Inductance of Current Shunts at Frequency Up to 200 kHz, IEEE Transactions on Instrumentation and Measurement, 2013, pp. 1664 – 1668

[42] X. Pan, J. Zhang, X. Ma, Y. Gu, W. Liu, B. Wang, Z. Lu, and D. Zhang, A Coaxial Time Constant Standard for the Determination of Phase Angle Errors of Current Shunts, IEEE Transactions on Instrumentation and Measurement, 2013, pp. 199-204

[43] I. Budovsky, Measurement of Phase Angle Errors of Precision Current Shunts in the Frequency Range From 40 Hz to 200 kHz, IEEE Transactions on Instrumentation and Measurement, 2007, pp. 284 – 288

[44] X. Pan, Q. Wang, J. Zhang, S. Zeng, W. Chen, Measurement of the level dependence in phase angle errors of the high current shunts, Precision Electromagnetic Measurements (CPEM), 2012, pp. 160-161

[45] U. Pogliano, B. Trinchera and D. Serazio, Wideband digital phase comparator for high current shunts, Precision Electromagnetic Measurements (CPEM), 2010, pp. 135-136

[46] G. C. Bosco, M. Garcocz, K. Lind, U. Pogliano, G. Rietveld, V. Tarasso, B. Voljc, and V. N. Zachovalova, Phase Comparison of High-Current Shunts up to 100 kHz, Precision Electromagnetic Measurements (CPEM), 2010, pp. 229-230

[47] J. D. de Aguilar, R. Caballero, and Y. A. Sanmamed, Realization and Validation of the 10 mA–100 A Current Standard at CEM, IEEE Transactions on Instrumentation and Measurement, 2014, pp. 1753 – 1759

[48] T. Funck and M. Klonz, Improved AC–DC Current Transfer Step-Up With New Current Shunts and Potential Driven Guarding, IEEE Transactions on Instrumentation and Measurement, 2007, pp. 361-364

[49]V. N. Zachovalova, M. Sira, J. Streit, and L. Indra, Measurement system for high current shunts DC characterization at CMI, Precision Electromagnetic Measurements (CPEM), 2010, pp. 607-608

[50] G. Rietveld, J. H. N. van der Beek, and E. Houtzager, DC Characterization of AC Current Shunts for Wideband Power Applications, IEEE Transactions on Instrumentation and Measurement, 2011, pp. 2191 – 2194

Laboratory model for design and verification of synchronous generator excitation control algorithms

S. Tusun*, I. Erceg* and I. Sirotić*
* Faculty of Electrical Engineering and Computing/Department of Electric Machines, Drives and Automation, Zagreb, Croatia
stjepan.tusun@fer.hr, igor.erceg@fer.hr, igor.sirotic@fer.hr

Abstract - This paper presents a laboratory model of synchronous generator excitation system based on National Instruments cRIO real-time industrial controller. It was specially designed for development and verification of classical linear and modern nonlinear excitation control algorithms. Real-time Clarke and Park transformations were implemented on the FPGA for measurements of the generator load angle, voltages and currents in the generator *dq*-frame. Automatic voltage regulator (AVR) and power system stabilizer (PSS2A) were implemented and experimentally verified on an 83kVA synchronous generator. Tests of step changes for generator voltage and mechanical power references, and test of transmission line disconnection were conducted. Experimental results were compared to simulation results of the designed model in Matlab/Simulink.

I. INTRODUCTION

Electrical power system is one of the largest technical systems. Increasing power demands [1] and introduction of new technologies, mostly based on power electronic [2], are the main reasons for changes and growth of the power system. Most of the produced electric energy comes from the synchronous generators installed in hydro, thermal and nuclear power plants. Long distances between power plants and operation of synchronous generator near its capability limits can lead to poor damping of electromechanical oscillations among group of generators or between specific areas of the power system [3]. These oscillations reduce power transfer capability limits and can lead to collapse of the power system [4].

One way to enhance the damping is by controlling synchronous generators excitation. Modern excitation systems are equipped with automatic voltage regulator (AVR) and power system stabilizer (PSS). Design and tuning of PSS is accomplished via linear techniques around an operating point [5-6]. However, synchronous machine and power system are highly nonlinear and generator operating point can change significantly due to load and topology changes and large disturbances. Therefore, nonlinear excitation strategies were proposed, such as: adaptive and intelligent control, feedback linearization, Lyapunov theory and energy shaping based controllers.

Nonlinear excitation control strategies mostly have complex control algorithms and require measurement or estimation of generator load angle and network parameters. Commercial available excitation systems are closed for modifications of the control algorithms and have predefined control options which are based on standard excitation types [7]. Implementation and testing of new excitation control strategies on these systems are very limited which was the main motivation for development of a new and open digital control system for excitation control.

Excitation control system DIRES21 was the first digital system developed on the Faculty of electrical engineering and computing in Zagreb (Department of electrical machines, drives and automation). It was based on four DSP ADMC300 processors [8]. The programming and configuration of control algorithms was accomplished by graphical oriented software development tool. There were more than 300 predefined blocks manually implemented via assembler language. After DIRES21, another system was developed based on Texas Instrument DSP TMS320F281 [9]. These two digital systems were specially designed for excitation control and special skills and programming knowledge was required for implementation of control algorithms. They had limited communication capabilities and there was no simple way for transferring large blocks of measurement data.

Over time the need for new and standardized development environment, which allowed fast prototyping and testing trough graphical development environment, grew. In [10] first prototype of excitation control system based on National Instruments cRIO platform was experimentally tested. This paper is continuation of [10] and it presents a new laboratory model for design, verification and fast prototyping of excitation control algorithms in graphic based software. Classical automatic voltage regulator (AVR) and power system stabilizer (type PSS2A) were implemented on the digital system using standard LabView blocks and experimental tested were conducted to show the system performances.

II. LABORATORY MODEL

Laboratory model used for verification of the excitation control algorithms is depicted in Fig. 1. The laboratory model enables the verification of the control algorithms in various control regimes (reference voltage and mechanical power step changes, turning on and off transmission lines, short circuit experiments, etc.).

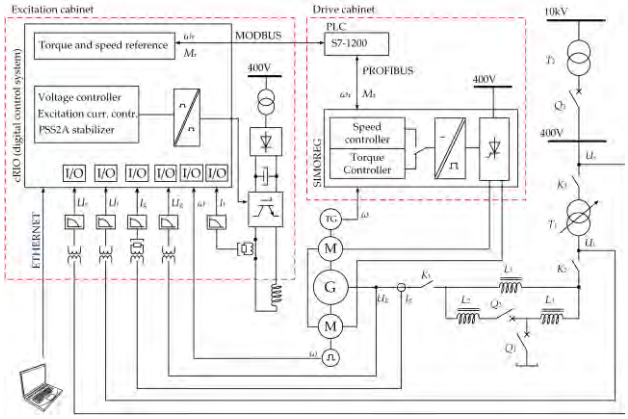


Figure 1. Laboratory model of 83kVA synchrnous generator

Laboratory model is located at the Faculty of Electrical Engineering and Computing in Zagreb and consists of:

- Salient synchronous generator mechanically coupled with two DC motors
- Transformer and inductances that simulate two parallel transmission lines,
- Thyristor rectifier for control of the DC machines armature current,
- IGBT rectifier for the synchronous machine excitation,
- Digital control systems for excitation control algorithms.

Synchronous generator (83kVA) is governed by two DC motors (44kW each) which are connected in series. Parameters of synchronous generator and DC motors are given in Appendix A (Tab. III and IV).

Inductors L1, L2 and L3 are used to represent parallel transmission lines. Disconnection of transmission line and three phase short circuits tests can be performed by operating circuit barkers Q2 and Q1. Transformer T1 is used for simulation of network voltage changes in range of ±10%. Connection of the T1 to the 10kV grid is accomplished by transformer T2. Parameters of inductors and transformers are given in Appendix A (Tab. V and VI).

DC motors are powered from the thyristor rectifier. Protection functions and parameterization of the rectifier is done by standard procedures [11]. Built-in speed and torque controllers are configured as two stage controllers. In first stage when the generator is not connected to the grid speed controller maintains synchronous speed (near nominal). In second stage when the generator circuit barker is closed (the generator is connected to the grid) torque control is activated. Speed and torque references are set by Siemens S7-1200 PLC. Also, the PLC is used as communication interface between the tyristor rectifier and cRIO digital system that controls the generator excitation.

Generator excitation is powered from independent voltage source trough IGBT rectifier (Fig. 1.). Output voltage of the IGBT rectifier is controlled by PWM signal from digital control system.

III. DIGITAL CONTROL SYSTEM

Digital control system, used for the control algorithm programming and data acquisition, is based on the National Instruments CompactRIO platform (cRIO-9014). cRIO is a combination of a real-time controller, reconfigurable IO modules (RIO) and FPGA module. I/O modules are used for data acquisition and analog to digital conversion (ADC). Generator currents and voltages are first measured by voltage and current transducers and then passed through analog anti-aliasing filters (Fig 1.).

cRIO program consists of microcontroller program loops and FPGA program loops. Microcontroller program loops are executed in real time and monitored by the real-time operating system (RT OS) while FPGA program loops are executed on dedicated hardware (Fig. 2).

A. Microcontroller program loops

Two program loops are periodically executed on microcontroller: communication loop and control loop (Fig. 2.) For the control and communication loop implementation, standard Labview library programming blocks were used. By using the standard blocks optimal code execution is guaranteed.

Main task of the communication loop is the transfer of the measured data, control signals and user interface commands. The loop is executed periodically with the frequency of 10 Hz. Modbus/TCP communication protocol is used for communication with Siemens PLC S7-1200.

Control loop is used for execution of the excitation control algorithms. Execution frequency of this loop is 500 Hz. Input measurements are converted by FPGA ADCs, scaled in the FPGA, and used for calculation of the output control value.

B. FPGA program loops

Three time critical program loops are implemented on FPGA module: counter loop, PWM loop and ADC measurement loop.

Counter loop is used for the digital encoder impulse counting. Positive and negative edge of the encoder's A and B impulses are counted. From counter value the generator rotor angle and speed is calculated.

Control signal for the IGBT rectifier is generated in the PWM loop. Triangular carrier signal with frequency of 1250 Hz is used for the PWM signal generator.

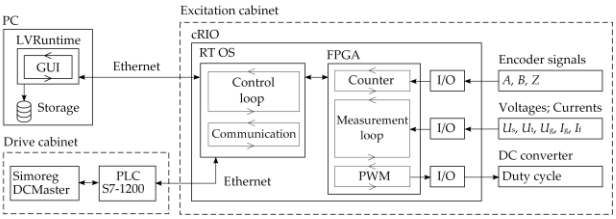


Figure 2. Structure of digital control system

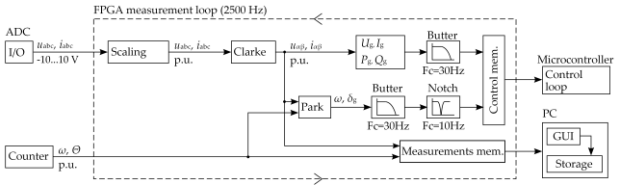


Figure 3. Strucutre of ADC measuremnt loop

ADC loop is used for signal conditioning and calculation of the Clarke and Park transform (Fig. 3). This loop is executed at the frequency of 2500Hz and it is synchronized with the PWM loop.

Signals from the ADCs are scaled to relative units (pu) and the Clarke transform is used for calculation of the $\alpha\beta$ components:

$$i_{\alpha\beta} = \begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} = \frac{2}{3} \begin{bmatrix} 1 & \frac{-1}{2} & \frac{-1}{2} \\ 0 & \frac{\sqrt{3}}{2} & \frac{-\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} \quad (1)$$

Furthermore, the Park transform is used to calculate the generator voltage dq components:

$$u_{dq} = \begin{bmatrix} u_d \\ u_q \end{bmatrix} = \begin{bmatrix} \sin\theta & -\cos\theta \\ \cos\theta & \sin\theta \end{bmatrix} \begin{bmatrix} u_\alpha \\ u_\beta \end{bmatrix} \quad (2)$$

Generator load angle is then calculated from the voltage dq components [12].

Generator active and reactive power, effective values of the generator currents and voltages are also calculated from $\alpha\beta$ components. These signals are then filtered by second order Butterworth filter. Additionally, the generator angular speed and load angle are filtered with 10Hz Notch filer to damp base mechanical frequency. These two digital filters are part of the standard LabView FPGA library. In the end, values of filtered signals are transferred to microcontroller memory and used as inputs in control loop.

Generator speed, rotor angle and $\alpha\beta$ components of the measured signals are transferred to the microcontroller shared memory. These values are then collected by GUI application which is executed on a PC. GUI application is used for transfer and storage of measurement data. Furthermore, GUI is used for references setup and user control of the digital control system.

IV. CLASSICAL EXCITATION CONTROL SYSTEM

Classical generator excitation control, synchronous generator [3-4], transmission lines and power system were modeled in Matlab/Simulink (Fig. 4.).

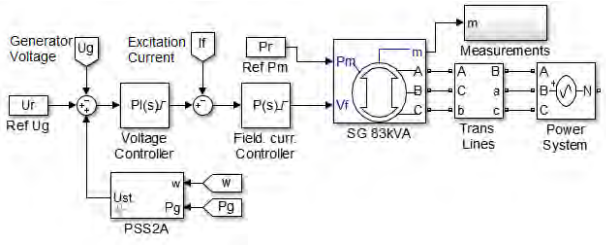


Figure 4. Matlab/Simulink model of classical excitation control sysem

Excitation control system consists of the AVR and PSS. The AVR control loop consists of an inner field current control loop and outer generator voltage control loop. Field current controller is proportional (P) and generator voltage controller is proportional integral (PI). Parameters of generator voltage and field current controller used for simulation and experiments are given in Tab. I.

PSS was modeled by transfer function blocks from Matab/Simulink as PSS2A type [7]. Main task of the PSS is the damping of synchronous generator electromechanical oscillations [3-4]. Parameters of the PSS are given in Tab. II

V. EXPERIMENTAL VERIFICATION

For experimental verification of the presented laboratory model AVR and PSS2A were implemented on the digital control system.

Discrete PI controller was implemented using standard PID block from LabView [13].

To implement the PSS2A on the digital control system it was necessary to convert its transfer function from continuous to discrete time domain. This step was accomplished by use of bilinear transformation in Matlab/Simulink.

To validate AVR and PSS2A following tests were performed on the laboratory model:

- voltage reference change,
- mechanical power change, and
- disconnection of transmission line.

For each experiment the generator terminal voltage, active and reactive power, load angle, angular speed and field current are shown (Fig. 5 to 8).

A. Voltage reference change

Voltage reference step change from 1.0 to 0.9 pu experiment was conducted for the case of the synchronous generator connected to the distribution network. The mechanical power reference of the generator was constant (0.5 pu). Also, simulation in Matlab/Simulink was

TABLE I. PARAMETERS OF THE GENERATOR VOLTAGE AND FIELD CURRENT CONTROLER

Controller	P gain	I gain	Upp. lim.	Low.lim.
Voltage [pu]	10	20	4	0
Field curr. [pu]	3	-	3	0

TABLE II. PARAMETERS OF PSS2A

$T_{w1,w3}$	T1	T2	T6	T7	T8	T9
1 s	0.3 s	0.05 s	0 s	1 s	0.2 s	0.09 s
Ks1	Ks2	Ks3	**N**	M	Upp. Lim.	Low. Lim.
5	0.5	1	1	4	0.1 pu	-0.1 pu

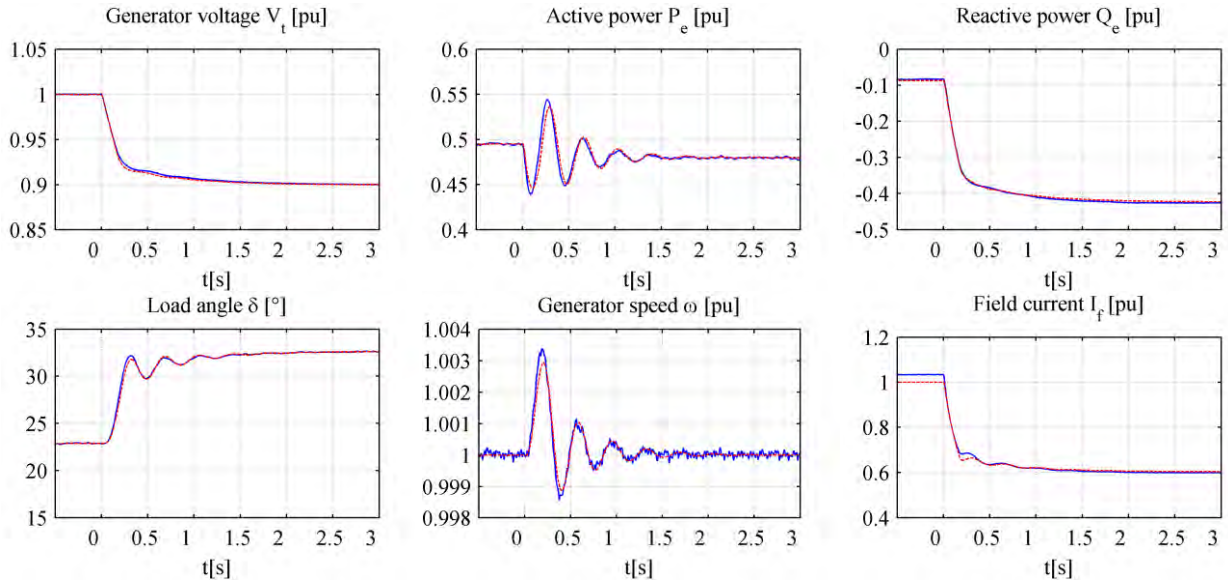


Figure 5. Comparison of experimental (blue) and simulation (red dash) results for a voltage reference change for the system with AVR

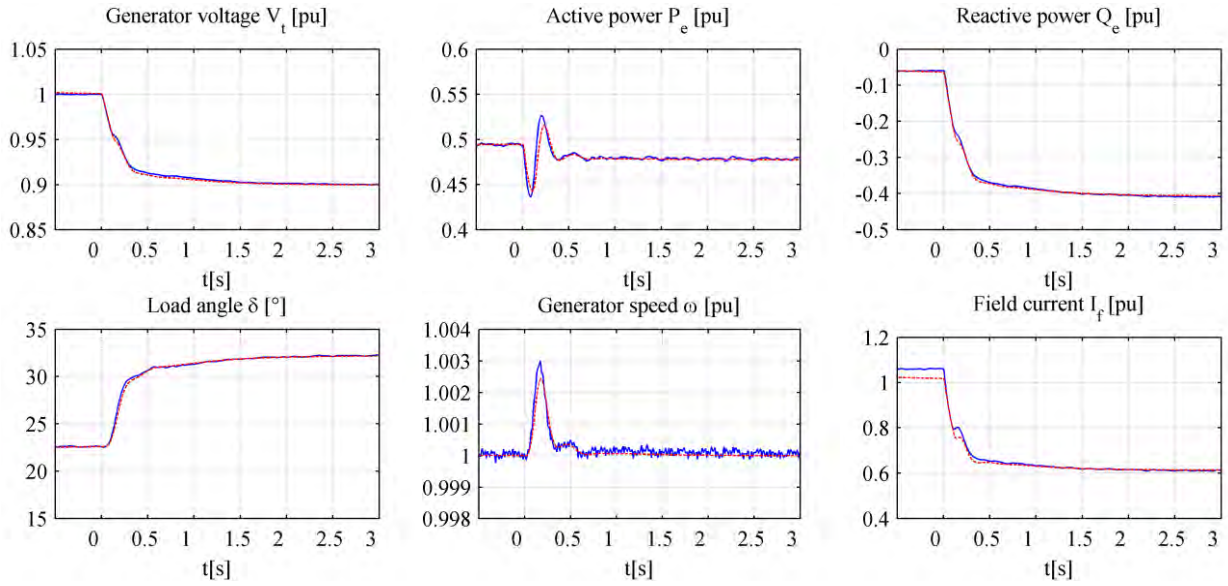


Figure 6. Comparison of experimental (blue) and simulation (red dash) results for a voltage reference change for the system with PSS2A

performed with the same conditions for validation of the proposed laboratory model.

As it can be seen from Figs. 5 and 6 there is no significant difference between simulation and experimental results, in transient and steady state. The greatest difference is in the signal of field current due to unmodeled hysteresis in simulation model.

Oscillations in signal of the active power, angular speed and load angle were presented for the system with AVR (Fig. 5.). These oscillations were successfully damped by adding PSS2A signal to AVR input (Fig. 6.). Active power feedback loop was not used in the laboratory model so there were differences in pre and post steady state values of active power (Figs. 5. and 6.).

B. Mechanical reference change

The mechanical reference step change from 0.3 to 0.6 pu experiment was conducted for the case of the synchronous generator connected to the distribution network. The generator voltage reference was constant (1.0 pu).

In Fig. 7 comparison of experimental results for the system with without and with PSS2A are shown. As it was expected, system with PSS2A better damps electromechanical oscillations than system with AVR (without PSS2A). Also, it is important to note that system with PSS2A has grater overshot in signal of generator voltage which is tradeoff for better oscillation damping.

C. Disconnection of transmission line

Experiment of disconnection of one of the transmission lines was conducted while the generator voltage reference (0.9 pu) and mechanical power reference (0.4 pu) were constant. Disconnection of the line (inductors L1 and L2) was done by circuit breaker Q2 (Fig.1.).

As it can be seen from Fig. 8 system with PSS2A better damps electromechanical oscillations than the system with AVR (without PSS2A).

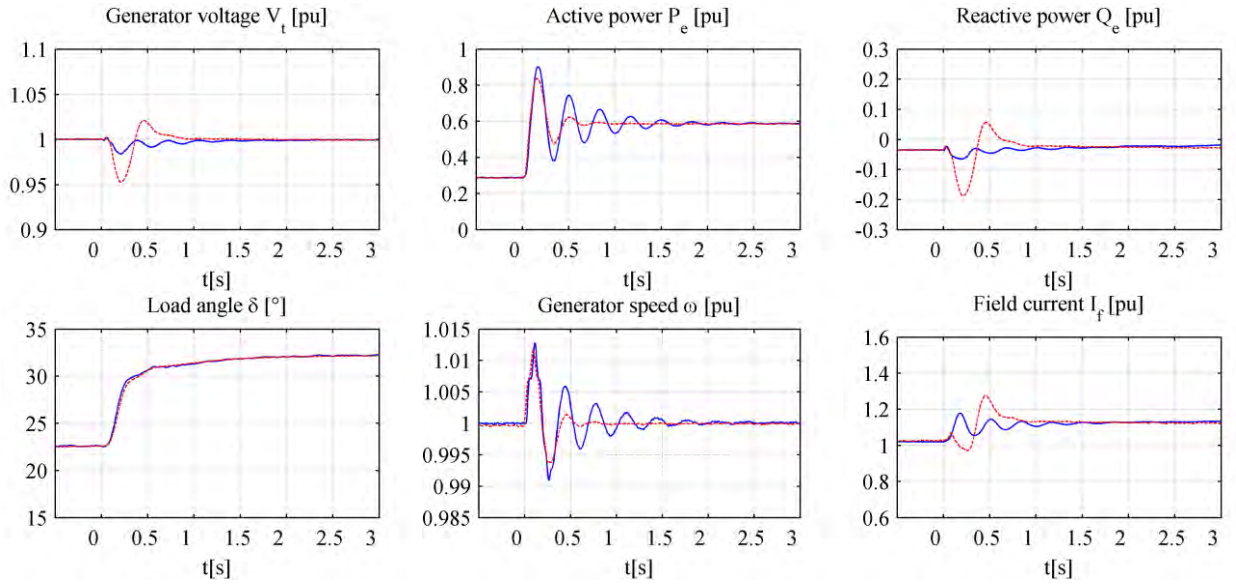


Figure 7. Experimental results for a mechanical power reference change for the system with AVR (blue) and system with PSS2A (red dash)

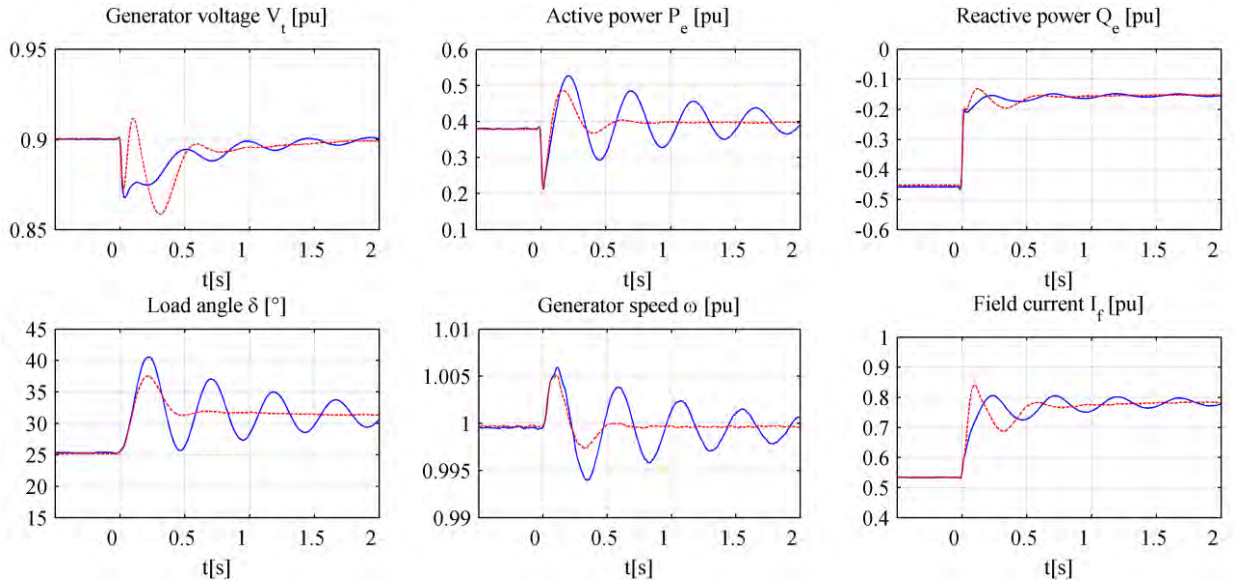


Figure 8. Experimental results for discconection of one transmission line for the system with AVR (blue) and system with PSS2A (red dash)

VI. CONCLUSION

Laboratory model for design and verification of synchronous generator excitation control algorithms was presented. Proposed model can be efficiently used for development and testing of classical linear and modern nonlinear excitation control algorithms. Implementation of control algorithms on the digital control system is accomplished by use of well-known graphically based development software. Furthermore, the graphic based software reduces time for control law development and verification.

Automatic voltage regulator and power system stabilizer were implemented and experimentally tested. Experimental tests of the generator voltage change, mechanical power change, and disconnection of transmission line showed that the PSS2A successfully damps electromechanical oscillations. Furthermore, experimental results were compared to results from Matlab/Simulink model. Developed simulation model can be used for parameter estimation and testing of new control algorithms before prototyping on laboratory model.

APPENDIX A. LABORATORY MODEL PARAMETERS

TABLE III. NOMINAL DATA OF SYNCHRONOUS GENERATOR

Voltage	400V	Speed	600r/min
Current	120A	Power factor (cos φ)	0.8
Power	83kVA	Excitation voltage	100V
Frequency1	50Hz	Excitation current	11.8A

TABLE IV. NOMINAL DATA OF DC MOTORS

Voltage	Current	Power	Speed
220V	192A	44,24kW	600r/min

TABLE V. NOMINAL DATA OF INDUCTORS

Inductor	L1	L2	L3
Inductance	3,5mH	1,35mH	0,45mH
Current	86A	180A	226A

TABLE VI. NOMINAL DATA OF TRANSFORMERS

Transformer	T1	T2
Voltage	380/380±10%	10/0,4kV
Power	145kVA	1MVA
u_k	6%	6%

REFERENCES

[1] A. A. Bayod-Rújula, "Future development of the electricity systems with distributed generation," *Energy*, vol. 34, no. 3, pp. 377–383, Mar. 2009

[2] F. Blaabjerg, Z. Chen, and S. B. Kjaer, "Power electronics as efficient interface in dispersed power generation systems," *Power Electronics, IEEE Transactions on*, vol. 19, no. 5, pp. 1184–1194, 2004.

[3] P. Kundur, N. J. Balu, and M. G. Lauby, *Power System Stability and Control*, 1st edition. New York: McGraw-Hill, 1994.

[4] G. Andersson, P. Donalek, R. Farmer, N. Hatziargyriou, I. Kamwa, P. Kundur, N. Martins, J. Paserba, P. Pourbeik, J. Sanchez-Gasca, R. Schulz, A. Stankovic, C. Taylor, and V. Vittal, "Causes of the 2003 major grid blackouts in North America and Europe, and recommended means to improve system dynamic performance," *IEEE Transactions on Power Systems*, vol. 20, no. 4, pp. 1922–1928, Nov. 2005.

[5] M. J. Gibbard, "Robust design of fixed-parameter power system stabilisers over a wide range of operating conditions," *Power Systems, IEEE Transactions on*, vol. 6, no. 2, pp. 794–800, 1991.

[6] D. Sumina, N. Bulić, and M. Mišković, "Parameter tuning of power system stabilizer using eigenvalue sensitivity," *Electric Power Systems Research*, vol. 81, no. 12, pp. 2171–2177, Dec. 2011.

[7] "IEEE Recommended Practice for Excitation System Models for Power System Stability Studies," *IEEE Std 421.5-2005 (Revision of IEEE Std 421.5-1992)*, pp. 0_1–85, 2006.

[8] T. Idžotić, D. Sumina, and I. Erceg, "DSP based excitation control system for synchronous generator," presented at the EDPE 2005, 2005.

[9] D. Sumina, N. Bulić, and M. Mišković, "Application of a DSP-Based Control System in a Course in Synchronous Machines and Excitation Systems," *International Journal of Electrical Engineering Education*, vol. 49, no. 3, pp. 334–348, Jul. 2012.

[10] I. Erceg, S. Tusun, and G. Erceg, "The use of programmable automation controller in synchronous generator excitation system," in *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, 2012, pp. 2055–2060.

[11] Siemens, Simoreg, DC-MASTER, "6RA70 Microprocessor-Based onverters from 6kW to 2500kW for Variable-Speed DC Drives", 13th ed., Siemens, 11 2007.

[12] T. Stjepan, I. Erceg, and G. Erceg. "Synchronous generator load angle estimation using Parks transformation." *9. savjetovanje HRO CIGRE 2009*. 2009.

[13] National, Instruments. (2009) PID and Fuzzy Logic Toolkit User Manual, aviliable at: www.ni.com/pdf/manuals/372192d.pdf

The European Project SolarDesign Illustrating the Role of Standardization in the Innovation System

W. Brenner*, N. Adamovic*

* Vienna University of Technology, Institute of Sensor and Actuator Systems, Vienna, Austria

werner.brenner@tuwien.ac.at

Abstract - The Framework Program 7 Project SolarDesign is focused on new photovoltaic (PV) integrated product solutions. To achieve this, new materials, flexible production - and business processes in PV powered product design and architecture had to be developed. Promising markets such as sustainable housing, temporary building structures, outdoor activities, electro-mobility, road lighting and mobile computing drive the demand for decentralized, attractive energy solutions. As products have to respect existing standards SolarDesign from the begin decided to proactively integrate standardization into the project's efforts. The example PV driven streetlamp demonstrates the influence of standards on performance requirements of the innovative PV materials.

I. INTRODUCTION

Designers, architects and industrial manufacturers share a common interest in using Photovoltaics (PV) as a decentralized and sustainable source of energy in their designs. Indeed photovoltaic modules are the only viable renewable energy solution that can be integrated directly into objects such as devices, textiles and surfaces of buildings. Photovoltaics (PV) is widely recognized as one of the key technologies for the future energy supply. The cost reduction within the last 15 years is a result of the utilization of learning effects, expansion of production capacities, extensive automation and standardization efforts [1].

SolarDesign reflects the consortium's commitment to actively contribute to a solid and sustainable standardization system that fosters added value for the European industry.

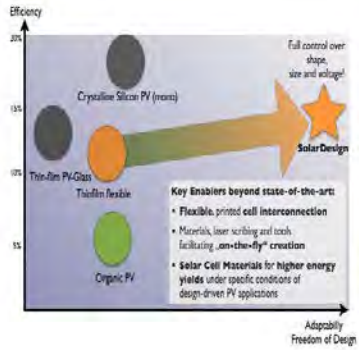


Figure 1. "Technical constraints" versus "design freedom"

II. SOLARDESIGN VISION

By adjusting the compositions and spectral responses of the functional layers of Thin Film Photovoltaics not only the overall efficiency can be improved but also energy yields under diffuse light conditions or at higher operating temperatures are changeable. This uniqueness of CIGS had potential for exploitation. The scientific and technical objectives (STOs) of this project are to develop:

STO1: A flexible scribing and printing technology that allows producing a given photovoltaic module according to specific design requirements "on-the-fly". This flexible interconnection is applied on the solar foil (i.e. an endless solar cell) with a minimum width of 300 mm and allows curved solar cells and interconnection patterns with a minimum radius of 10 mm.

STO2: Novel materials for the underlying flexible solar cell technology to extend the design related degrees of freedom and to optimize the materials used for integrative solar applications.

STO3: Novel materials for satisfying design related requirements on solar module level (the part that is most visible for the beholder). Focus will be laid on materials for the electrical conducting front grid to allow a high design freedom of patterns and colour variations, as well as using of different novel encapsulants allowing custom designed optical appearance.

STO4: A methodological toolbox to provide design rules for the best solar cell super-structure and module design layout for a given application by using numerical modelling and simulation.

STO5: The following applications demonstrate the developed technologies: solar charging (cover for tablet PC and solar powered radio), solar powered light, solar powered sensor networks for detection of fire in forest, urban solar lighting (compact solar street lighting system) for Product Integrated Photovoltaics (PIPV), Building Integrated Photovoltaics (BIPV) and integration of PV in textile support.

III. STANDARDIZATION AND INNOVATION

Driven by increased competition many countries and companies have started efforts focused on implementation of international standardization in an early phase of research and development (R&D) [9].

Extensive research on the economic evaluation of new technology development and international standardization has been conducted [10], [11] which evidences that R&D that takes into consideration standardization can enhance the efficiency of investment and stimulate the introduction of developed technologies and solutions.

Accordingly, world wide economy and mass production are typically associated with standardization. Standardization meanwhile extended its scope towards the research community in order to ease market access of innovations and to provide interoperability between new and existing products, services and processes

Companies and research institutions can profit from participation in a Technical Committee (TC), which enables them to benefit from the ability to internalize the external information resources and thus to increase innovation skills [12]. In this light, the positive effect of participation can foster an enhancement of existing – and a creation of new – competence and innovations. On the other hand, it evidences interest in new technologies and solutions, shows the willingness to co-operate with other prospective users and indicates the intention to enter innovative markets. Early access to information, the degree of the strategic influence on new projects and the influence on technological development can be affected by the participants' traditional or developing positioning (modes of participation, roles overtaken, functions, social rank etc.) [13]. Standards enable companies and research institutions to comply with relevant laws and regulations and help to ensure methodological robustness and wide acceptance [18].

Standards provide a guiding framework for FP and H2020 research projects, ensuring: tests and analytical work are carried out according to established norms, and developed technologies are interoperable with existing technologies and compliant with industry standards. By working to existing standards research projects have a higher chance of their outputs being accepted by scientific and industrial communities. Working with existing standards also enables researchers to recommend and contribute to new standards development, thus increasing their technical knowledge, widening their business networks and strengthening the market exploitation of their results [6].

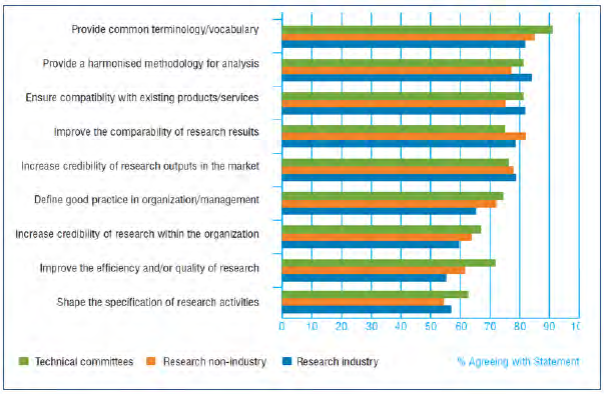


Figure 2. Benefits of Research of Using Standards [5]

SolarDesign as many FP7 and H2020 projects started with an analysis and review of existing standards, either to identify gaps in coverage or as the basis for revising resp. extending them.

There are many thousands of standards of various types. Standards can be categorized into four major types

- Fundamental standards - concerning terminology, conventions, signs, units etc
- Test methods and analysis standards - which measure characteristics such as temperature, size, force and chemical composition
- Specification standards - which define a product's characteristics (product standards), or a service (service activities standards) and their performance thresholds such as fitness for use, interfaces and interoperability, health and safety, environmental protection, etc
- Organisation standards - which describe the functions and relationships of a company, as well as elements such as quality management and assurance, maintenance, value analysis, logistics, project or system management, production management, etc

SolarDesign has decided to integrate standardization into the project's efforts as standards can enhance the economic value of research and innovation projects [3]. An early step was the decision which route answered SolarDesign's needs best. There are several ways in which standards and standardization can be integrated

- Integration of standardization bodies into the project's consortium
- Identification of a national, European or international Standards Body which can be an associate in the project, for example in a Steering Group
- Informal participation of CEN-CENELEC Management Centre as associate in a Steering Group. This is possible in projects with specific work packages on standardization
- Identification of the links SolarDesign has with ongoing standardization (e.g. "Photovoltaics in buildings" prEN 50583)
- requesting a Project Liaison with an existing Technical Committee. As soon as this status has been granted, the project can participate in the Technical Committee's plenary meetings and contribute to the working groups.
- The project requesting the Project Liaison can demonstrate formal collaboration with the European Standardization System (usually a requirement in FP7 calls);
- The project representative can participate in the TC directly thus ensuring synergies between the research and standardization work (avoiding duplication of standardization work);
- The project representative can propose a new work item (standard) directly to the TC without

going through a national delegation (direct impact on the standardization work program.

International standardization is a highly social activity involving specialists, detailled observation of the written standardization process, and a clear strategic agenda. Often it is a highly organizational activity, and in many cases it is a highly individual-dependent activity. Consensus building is an important process which demands the following skills [14]:

- sharing goals, costs, risks, quality requirements, measures, and alternatives
- sharing awareness of each party's positions, expectations and backgrounds
- skills in communications, listening, persuasion and facilitation
- ability to foster compromises

It is impacted by the following issues:

- ongoing explicit or hidden business models
- human networks and trusts among participants
- issues due to external organizational aspects
- technology trends

In the frame of SolarDesign the coordinating institution TU Wien motivates the partners to proactively contribute to the standards development processes. Standardization within Task 5.3 aims at providing a bridge connecting research to industry by promoting innovation and commercialization through dissemination of new ideas and best practice. This comprises circulating of new measurement and evaluation methods, implementation of new processes and procedures created by bringing together all interested parties such as manufacturers, researchers, designers and regulators concerning products, raw materials, processes or services. SolarDesign's consortium puts strong efforts on including sessions on standardization in the frame of meetings for increasing the internal and external awareness of the topic (e.g. a standardization workshop was held at midterm).

Integrating photovoltaics in architecture and industrial design is in its infancy and taking into account different European perspectives is crucial for its success. This is especially true for Building Integrated PV where legal conditions like construction codes or feed-in regimes are differing from one European country to another.

By seeing design from a stylish or sensory perspective regional characteristics exist. Taking product designers and PV experts from different European countries can ensure the incorporation of varying design manifestations.

Dissemination, implementation and standardization cannot be seen as separate activities. They must be fully coordinated and represent two sides of the same plan for success. Networking and documentary standards are promising and vital tools of disseminating SolarDesign's research to the market place.

All documents of standardization relevance are jointly prepared by the partners and ongoingly are bundled and released by the task leader TU Wien.

IV. SolarDesign's Streetlamp - An Example how Innovation and Standardization Influence Each Other

In an early stage of the project the SolarDesign consortium identified where standards can benefit the project. In the next step the participants defined standardization issues.

A typical problem of PV systems is the power loss due to temperature increase, because modules often operate close to the product envelope with low ventilation. SolarDesign's partner institution EURAC therefore evaluated and compared the PV temperature conditions of different PV module categories (in terms of PV technology and material type). A simple linear expression for the evaluation of the PV module temperature is $T_{mod}=T_{amb}+kG$ which links T_{mod} with the ambient temperature T_{amb} and the incident solar radiation flux G. Within this expression the value of the dimensional parameter k, known as the Ross coefficient, depends on several aspects (i.e. module type, wind velocity and integration characteristics). However, dispersed values for this parameter can be found in literature (in the range of 0.02-0.06 K m^2/W) according to different module types [7].

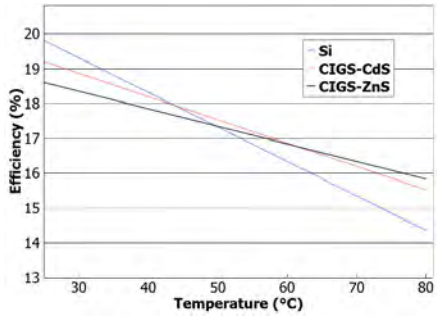


Figure 3. Overview of temperature coefficients of different thin film photovoltaic technologies [9]

As can be seen in Figure 3, the efficiency decrease of ZnS-buffered CIGS (Copper, Indium, Gallium, Diselenide) is smaller than for CdS-buffered CIGS. With further improvement of alternative buffers (like ZnS, ZnMgO, Zn(O,S)), the efficiency values of CdS-buffered CIGS at 25°C can be reached (demonstrated by Helmholtz Zentrum Berlin [8]. Due to lower temperature coefficient even higher efficiencies at operating temperatures (typically ~60°C for midsummer day) can be expected for alternative buffers – without using toxic Cd. For deposition of different buffer layer materials both sputtering and wet chemical process are available. Manufacturers typically rate PV modules at standard test conditions. However, the actual energy production of field installed PV modules is a result of a range of operating temperatures, irradiances, and sunlight spectra. Therefore, there is an urgent need to characterize PV modules at different temperatures and irradiances to provide comprehensive rating information [15]. One of the most relevant PV standards issued by the

International Electrotechnical Commission (IEC) Technical Committee 82 Working Group 2 (IEC/TC82/ WG2) is the IEC 61853 standard titled Photovoltaic Module Performance Testing and Energy Rating (IEC, 2011). This can be seen as relevant for the CIGS PV modules. Why is their high efficiency of relevance in this specific case of autonomous streetlamp? The technical parameters of streetlamps are defined in four standards which guarantee a common base of light conditions in European cities.

PD CEN/TR 13201-1:2014 Road lighting. Guidelines on selection of lighting classes Road lighting, Lighting systems, Roads, Lighting levels, Classification systems, Road safety, Traffic flow

BS EN 13201-2:2015 Road lighting. Performance requirements Road lighting, Lighting systems, Roads, Performance, Classification systems, Luminance, Glare, Lighting levels, Road safety, Environmental engineering, Luminaires, Pedestrian-crossing lights

BS EN 13201-3:2015 Road lighting. Calculation of performance Road lighting, Lighting systems, Roads, Performance, Mathematical calculations, Photometry (light measurement), Light distribution, Lighting levels, Luminance, Road safety, Luminaires

BS EN 13201-4:2015 Road lighting. Methods of measuring lighting performance Road lighting, Lighting systems, Roads, Performance, Photometry (light measurement), Performance testing, Test equipment, Luminance, Lighting levels, Luminaires, Reports.

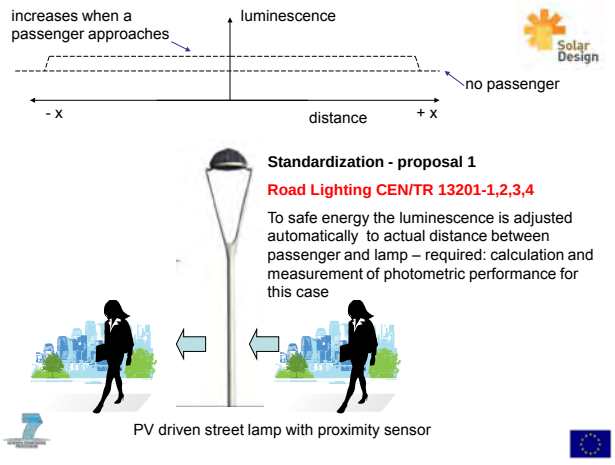


Figure 4. Energy saving performance of a PV driven street lamp as example of need for standards development in the frame of the FP7 NMP project SolarDesign [2]

To meet all these requirements research undertaken in the frame of FP7 NMP project SolarDesign had to concentrate on an appropriate level of CIGS efficiency to guarantee the luminance defined in the respective standard.

To safe electric energy which can be harvested only during day, the luminescence of the street lamp is modulated according to the distance of approaching passengers (by an integrated proximity sensor). This special performance characterized by changing lighting levels makes necessary a new standard covering related calculation procedures too.

V. Characteristics of Standardization Process

International standardization from the viewpoint of social structure shows several characteristics. First, the level of knowledge required to be a leader is significantly high. This leads to the fact that acting as the main editor of a technical standardization requires experience and extensive work, which is difficult to carry out as a side job.

Standardization on European level turns out to be a long process [16], different from national standards which often can be finalized and issued within a few months. At the international level, the full process takes two up to five years. There are six main stages in publishing a new European resp. international standard. Proposal stage**:** originates from a National Committee. Represented countries vote on the interest and nominate experts which define a work programme with target dates. Preparatory stage**:** a Working Draft (WD) is prepared by a project team. Committee stage: a Committee Draft (CD) is submitted to the National Committees for comment. Enquiry stage: a Committee Draft for Vote (CDV) is submitted to all National Committees: a majority of two thirds is requested. Approval stage: a Final Draft International Standard (FDIS) is prepared. Publication stage: if the FDIS is approved by a two thirds majority, the document is published by the International Electrotecnical Commission (IEC) Central Office as an international standard.

To avoid duplication of efforts, speed up standards preparation and ensure the best use of the resources available and particularly of experts' time IEC and CEN/CENELEC [5] agreed: if the results of parallel voting are positive in both the IEC and CEN/CENELEC, the IEC will publish the International Standard, while the CEN/CENELEC Technical Board will ratify the European Standard [17].

In technology-based markets, to absorb external knowledge and conversely to stake their claims, stakeholders participate in the global socio-technical network of standardization [12]. Main target stakeholders of SolarDesign to be addressed are:

- Standardization bodies: International standardization organisations such as: (CEN/CENELEC, IEEE, etc.) and national such as: HZN, AFNOR, DIN, SNV, SIST. They develop and establish product and/or process standards to be followed by producers and application developers
- Certification entities: (EEPCA - Professional Association of the European Certification Bodies, EECC - The European Certification Council, etc.) provide confidence to users that a certain element of the SolarDesign project is produced and/or operated according to a defined set of practices or standards
- Photovoltaic systems suppliers, installers and service providers for energy efficient buildings (Building integration and architecture) and Solar-powered consumer products (e.g. solar lighting,

textile integration) through national, European and international associations related with the technologies developed in the project

- Training providers: provide training to qualify developers and other professionals to work with the SolarDesign project results. Such as: Universities, RTD and devoted training organisations
- Local Authorities & National/Regional Public Bodies are key players as policy makers, favourable legislative framework creation, public procurement
- Architects' associations. Architects need to be provided with appropriate training, tools and guidelines for them to consider the integration of photovoltaic materials
- Solar-powered consumer users. To be provided with appropriate training, tools and guidelines for them to consider the integration of photovoltaic materials in their products
- Construction companies associations and related research associations should be aware of the new technologies that will be installed
- Public and private real estate Promoters. They can offer to their clients the advantage of the developed system therefore they should be informed about it and about new business models
- Clients and users (citizens): key actors interested in cooperative working systems or applications providing their perspectives in the formulation and assessment of the project results
- Network operators: may act as a channel for offering and billing services and/or access devices to users.
- Energy management agencies: regional and/or national energy agencies are promoting efficient and innovative energy technologies

A basic requirement to minimalize the duration of the whole consensus process is a consistently written specification, which requires a skilled and dedicated editor resp. team of editors. This often is characterized by the fact that there are some hidden structures in the consensus process.

The consensus process never will be an equality-based decision-finding and communication process. Some stakeholders contribute significant efforts for editing the specification in order to drive some favoured business goals [14]. But without such contributions, it is difficult to carry out today's international standardization. This situation results in a clear diversification of roles and responsibilities in the complex standardization process as shown in Table I.

TABLE I. ROLES OF PLAYERS IN INTERNATIONAL STANDARDIZATION [14]

Role	Description
Editor	Coordinates all inputs and updates, has to be well prepared to provide editorship and can replace other editors if this should become necessary
Contributors	Experts who create new inputs and support discussions, they do not overtake responsibility as editors
Driver	The workforce that drives and accelerates the standard with collaborative writing.
Registers	Try to delays or obstructs the progress of a standard.
Watcher	The people that watch the trends.
Facilitator	Facilitate consensus process
Process Observer	Watch the consensus process and express opinions on process violation, but they avoid responsibility for consensus finding
Lobbyist	People that negotiate under the table for their own business models.
Technology Observer	Watch the consensus process for technical quality

Barriers and problems encountered when contributing to new or revised standards can be summarized as follows [5]: first, timetabling issues, wherein the timeframes for the research project and the standards development work are not aligned sufficiently and coordination between the two processes becomes difficult because of the differing stages of progress. Difficulties in gaining acceptance for the inputs put forward by the project team, noting competing research resp. industrial interests, lobby groups, or results too 'innovative' to be accepted by the standardization groups. Another problem often is based on resource availability over a longer time, given the short-term nature of project funding, often resulting in a funding gap. In cases of limited experience in standardization additional difficulties might rise from unclear access to standardization bodies and their technical committees, due to membership rules, lack of direct participation, etc.

Researchers often see standards development as a time consuming and difficult process which limits the available time for research and teaching. Project teams might find it difficult to identify suitably qualified experts able and willing to work within the standardization process to implement the project results. Anyway the consortium has to be aware of a learning curve associated with understanding the world of standardization (e.g. how to identify already existing standards, how to propose and make changes to standards, how to gain acceptance, etc.).

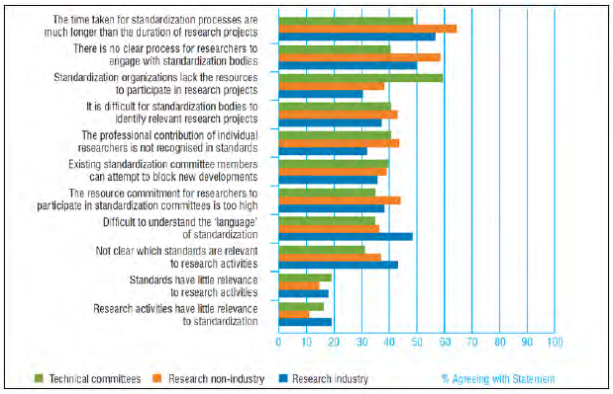


Figure 5. Barriers of Links between Research and Standardization [5]

VI. CONSORTIUM

The SolarDesign consortium is constituted by 11 participants (6 SMEs and 5 research institutions) from 6 countries who gather all the necessary background and expertise to achieve the ambitious research and standardization objectives of the project.

Consortium: Technische Universität Wien (Austria), Sunplugged GmbH (Austria), Faktor 3 ApS (Denmark), Innovatec Sensorisatión y Comunicación S.L (Spain), Studio Itinerante Arquitectura S.L. (Spain), RHP Technology GmbH (Austria), Asociación de Industrias de las Technologias Electrónicas y de la Información del País Vasco (Spain), Munich University of Applied Sciences (Germany), Accademia Europea Bolzano (Italy), Università degli Studi di Milano-Bicocca (Italy), Commissariat à l'energie atomique et aux energies alternatives (France).

VII. CONCLUSION

Integrating photovoltaics into industrial design is in its infancy and taking into account different European perspectives is crucial for its success. Seeing design from a stylish or sensory perspective regional characteristics exist. This is especially true where legal conditions are differing from one European country to another.

Dissemination, implementation and standardization cannot be seen as separate activities. They must be fully coordinated and represent two sides of the same plan for success. SolarDesign from the beginning decided to contribute to the role that standardization plays, such as:

- Ensuring broad applicability of SolarDesigns's outcome
- Fostering an increase of the efficiency of research and development work
- Supporting that interoperability of developed solutions is given with already existing technologies or regulations.
- Contributing to standards development where a need was identified

ACKNOWLEDGMENT

The FP7 NMP SolarDesign Project has received funding from the European Union's Seventh Framework Programme under grant agreement n° 310220.

REFERENCES

[1] B. Parida, S. Iniyan and R. Goic, "A review of solar photovoltaic technologies," Renewable and Sustainable Energy Reviews 15 (2011), pp. 1625–1636.

[2] SolarDesign Newsletter 4, January 2015, http://www.solar-design.eu/ .

[3] CEN (European Committee for Standardization)/CENELEC (European Committee for Electrotechnical Standardization), "Why Standards," http://www.cencenelec.eu/research/WhyStandards/Pages/default.aspx.

[4] S. Ishizuka, T. Yoshiyama, K. Mizukoshi, A. Yamada and S. Niki, "Monolithically integrated flexible Cu(In,Ga)Se2 solar cell submodules," Solar Energy Materials and Solar Cells 94 (2010), pp. 2052-2056.

[5] CEN (European Committee for Standardization)/CENELEC (European Committee for Electrotechnical Standardization), "Research Study on the Benefits of Linking Innovation and Standardization," Dec 2014, Ref: J2572/CEN.

[6] J. Stroyan and N. Brown, "Study on the contribution of standardization to innovation in European-funded research projects," technopolis group, Sept 2013

[7] L. Maturi, G. Belluardo, D. Moser and M. Dell Buono, "BiPV system performance and efficiency drops: overview on PV module temperature conditions of different module types," Energy Procedia 48 (2014), pp. 1311-1319.

[8] H. Papathanasiou, "ROHS compatible buffer layers for CIGS based solar modules – first results from the Photovoltaic Alliance Project," 3rd International Workshop on CIGS Solar Cell Technology, Berlin, 17.-19.04.2012.

[9] S. Y. Sohn and Y. Kim, "Economic evaluation model for international standardization of correlated technologies," IEEE Transactions on Engineering Management, vol. 58, no. 2, May 2011, pp 189-198.

[10] DIN (Deutsches Institut fuer Normung e.V.), (2000), "The economic benefits of standardization," Research Rep., http://www2.din.de/ .

[11] Standards Australia. (2006), "Standards and the economy," Research Report, http://www.standards.org.au/ .

[12] E. N. Filipovic, "How to support a standard on a multi-level playing field of standardization: propositions, strategies and contributions," Proceedings of the 2014 ITU Kaleidoscope Academic Conference: Living in a converged world - Impossible without standards?, 2014, pp: 207-214, DOI: 10.1109/Kaleidoscope.2014.6858464 referenced in: IEEE Conference Publications.

[13] S. Wurster, K. Blind and S. Fischer, "Born global standard establishers identification of a new research field and contribution to network theory," Proceedings of the 8th IEEE Conference on Standardisation and Innovation in Information Technology (SIIT), Sophia-Antipolis, 09/ 2013, doi 10.1109/SIIT.2013.6774583, pp. 1-12

[14] T. Yamakami, "A three-dimensional view model of international standardization using techno-sociological analysis," 4th International Conference on Interaction Sciences (ICIS), 2011, IEEE Conference Publications, pp. 13-18.

[15] G. Tamizh Mani,. K. Paghasian, J. Kuitche, M. Gupta Vemula and G. Sivasubramanian, "Photovoltaic module power rating per IEC 61853-1 standard: a study under natural sunlight," Arizona State University Photovoltaic Reliability Laboratory (PRL), March 2011, www.solarabcs.org/ratingper61853 .

[16] D. Blanquet, P. Boulanger, A. Guerin de Montgareuil, P. Jourde, Ph. Malbranche and F. Mattera, "Advances needed in standardisation of PV components and systems," Proceedings of 3rd World Conference on Photovoltaic Energy Conversion May 11-18, 2003 Osuka, Japan, pp. 1877-1881.

[17] IEC International Electrotecnical Commission, "Inside the IEC – Information," www.iec.ch.

[18] CEN (European Committee for Standardization)/CENELEC (European Committee for Electrotechnical Standardization), "The importance of standards," http://www.cencenelec.eu/research/tools/ImportanceENs/Pages/default.aspx .

Open public design methodology and design process

D. Rembold*, S. Jovalekic**
* IES, Albstadt, Germany, rembold@hs-albstadt.de
** IES, Albstadt, Germany, jovalekic@hs-albstadt.de

Abstract - This paper proposes a design methodology and design process for mechatronic systems incorporating the public community. As a proof of concept, we are building up robot system consisting of the following components: software code files, hardware design files and controller hardware. Every component will be available for review on an open repository and review tool which is GIT/Gerrit. Designers release components into the repository in form of commits. Commits can be downloaded by every community member, examined and reviewed in detail by the public community. At certain time spots we pick a number of commits with positive review results, and review the commits on an overall system perspective. If this review passes, we create a real system from the commits. The steps are: software code compilation, hardware components creation with a 3D printer, controller hardware ordering. Then we build the new robot upon these components. If we can build successfully the robot system (which is expected due to the overall review) and if we are passing all testcases, a new release is created from the picked commits. Every community member can download the latest release for their own usage. The business case for this design methodology and design process, is that we offer to the customer a service to conduct tests from a set of reviewed commits. We provide the resources and the knowledge to this service. We inform the customer about the review and test results and the customer can precede improving the system based on the feedback we provide.

I. INTRODUCTION

Recently a new maker community has developed for machine development. The designs is often free to download, and vice versa everybody can upload his machine design on dedicated websites. The public community is able to put feedback in form of remarks next to the upload. Unfortunately this is not really following a real process in a way we know this from companies. The originator of the upload is free to take a look into the remarks or comments, but there is no requirement to pursue the feedback.

Often the uploaded designs lack of sufficient description, e.g. we see quite often that a build of material (BOM) is missing. So we think, that more effort must be spent into improving the build description.

Our goal of this project is to create a process to build machines (in our case a robot system) consisting of components such as software code, design files, controller hardware and build instructions in form of videos and boms.

Every component will be available for review on a public repository and an open review tool such as GIT/Gerrit, see Figure 1.

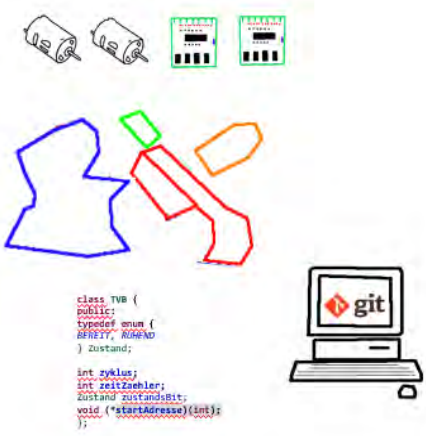


Figure 1. Design Parts in Repository

The public community can review every single component and add its approval or rejection to it, see Figure 2. In detail this means, all components to be reviewed are available in the repository in form of commits, see Figure 3. Commits can be downloaded by every community member, examined and reviewed in detail. At certain time spots we pick a number of commits with positive review results, create a system from them, test it and eventually release the commits if successful, see Figure 4.



Figure 2. Review

During creation of the system, software components are compiled, design files are created by a 3D printer, and controller hardware are ordered automatically. Then we build the new robot upon these components. If we are able to build successfully the robot system and if the robot is passing all testcases, a new release is created from the picked commits. Then the picked commits are moved into the master repository of GIT for the public download.

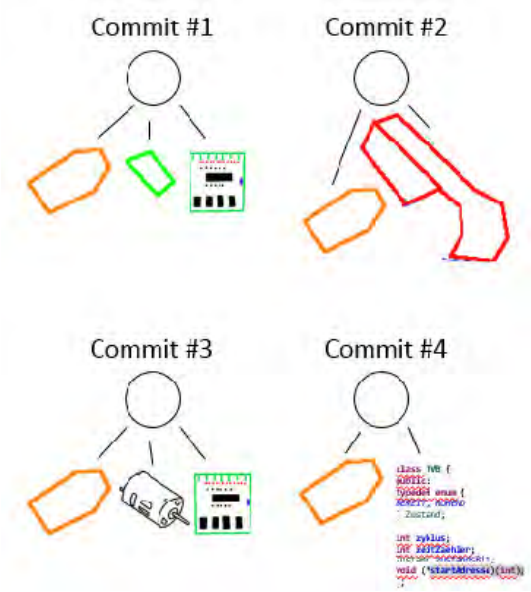


Figure 3. Commits

The business case for this design methodology and design process is to offer a service to the customer to conduct tests from a set of reviewed commits. We provide the resources and the knowledge to this service. After executing the above described process we inform the customer about the test results and the customer can precede improving the system based on the feedback we provide.

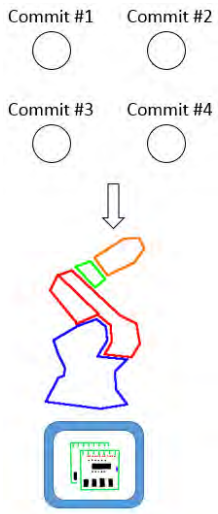


Figure 4. New release from commits

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

II. COMMIT DEPENDENCIES

There is a dependency between design files and software code. Updated design files for the robot can cause an update to new robot geometry layout, therefore the robot's inverse transformation algorithm needs to be adapted inside the software code. So in a case there is a change in robot geometry, but the person who created the commit did not change the inverse transformation, then we do not pick the design file change, because this is a violation of the process rule. The person who commits must indicate the dependency between the design file commit and the software code commit.

Another dependency lies in the choice of the motor driver, motor and the robot layout. The motor driver and motor need to be able to handle the inertia of the robot which is given by the robot layout. In case there is a new commit with a new motor driver, but no commit with a robot layout change, then there must be at minimum a justification that a new robot layout is not needed. Also a new motor driver, might have consequences on the software which drivers the motor driver.

There are dependencies on commits such as software code and testcases. E.g. if a motor driver is updated, we expect a new release of a testcase which reflects the testing of new attributes of the motor driver. Otherwise it would make little sense to introduce a new motor driver, if there are no testcases to verify them.

We offer simulation code for simulation tests of newly committed software code. However new design files can cause a change in robot geometry layout, so the simulation program must be adapted here as well. We are working on providing a configuration file of the robot's geometry. In future we will not pick any design file commits, if there was no adaption of the robot's geometry configuration file.

III. AUTOMATIC TESTING

As continuous integration tool we use Jenkins to gather the commits and to process the components in an automatic fashion. There are several steps Jenkins is working on the commits. The steps depend on the content of the commits:

The first step is that Jenkins extracts reviewed commits containing software driver changes and it compiles a new driver level. Jenkins then extracts the testcases and the robot geometry layout and updates the simulation program. The driver is then tested against the simulation.

In case the commits contain new robot components in form of design files, Jenkins loads the design files automatically into 3D printer and it creates them one by one.

Jenkins passes the content of the commits with purchasable parts such as new robot controller parts to an ordering software and orders automatically the new robot controller parts.

We cannot do the assembly automatically due to its complexity, so this is done manually. The tester of the robot system assembles the robot with the newly created components (design files and robot controller parts), loads

the compiled software to the controllers and tests the system.

IV. BUSINESS IDEA

We are working toward a goal to use this process to create a business opportunity to offer a service to customers. There might be customers who are interested in one of our designs, but in general the designs don't fulfill completely their requirements. The customer can download the design and create their own system from it, but he has often a good understanding of what needs to be done differently. Often the customer does not have the knowledge to build a new system, or he is not willing to spend the effort himself. So he can ask us for a service to conduct tests from a set of reviewed commits. The commits can be existing commits, commits provided by the customer himself or requested design changes we released for the customer.

We provide the resources to do the required design changes and to conduct tests. Every design change will be uploaded to git in form of a commit. We retrieve the requested commits and apply them to a process in three phases:

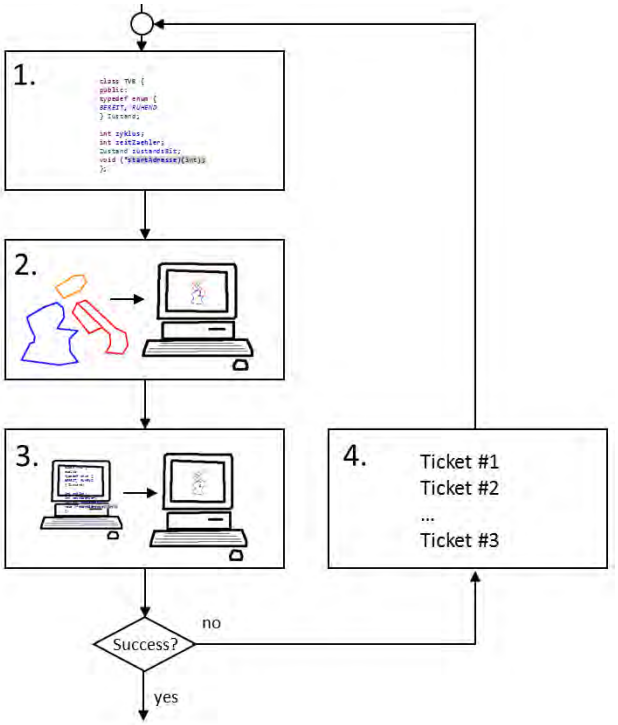


Figure 5. First Phase

A. First Phase

In the first Phase we pick components and solely test them on simulation, Figure 5. depicts this process.

1. The customer selects together with our consultation the committed software code from our repository, and we pull the software code and create a driver from it.
2. We pull simulation software from the repository, if e.g. updated design files require this. This is the case when the design files cause a new robot geometry layout. A new simulation system is built from the new design files.
3. We run and test the driver against the simulation with testcases located in the repository. In some cases the testcases need to be updated, due to new robot layout, or new robot controller parts.
4. In case simulation fails, we will create a list of errors and a list of suggestions for improvement (e.g. with a ticketing system) and hand them back to the customer.

If all tests pass in the first phase, we can move to the second phase:

B. Second Phase

The purpose of the previous phase was to test the software code and the hardware in simulation only. The second phase is concerned about creating the real hardware of the robot system, see Figure 6.

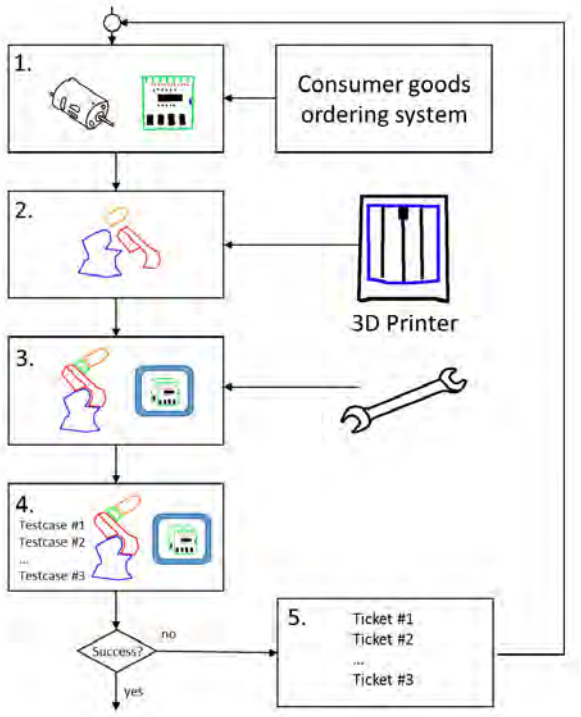


Figure 6. Second Phase

1. In case the customer requires new controller hardware described in the commits, and the commits passed the reviews, then Jenkins will pull the content of the commits from the repository and orders automatically the hardware components using a consumer goods ordering system.
2. If there are new design files, Jenkins pulls the commits with design files from the repository and

loads them into the 3D printer and it prints the updated parts.

3. From all newly created parts, we build manually the new robot system.
4. We run testcases against the real hardware. These are the same testcases, which we ran previously in Phase 1. against the simulation.
5. If the test against real hardware fails, we create a list of errors and a list of suggestions for improvement and hand them back to the customer.

If the second phase passes all tests, then we are moving to the third phase:

C. Third Phase

The second phase was about testing the system in a real hardware environment. If all testcases pass the second phase, then we precede with the repository git itself, see Figure 7.

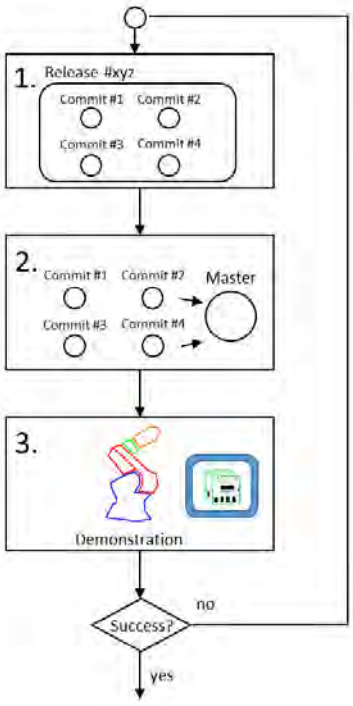


Figure 7. Third Phase

1. We gather all commits contributed in phase 1. and phase 2., and create a new release level from the them.
2. All commits are merged into the master branch of the repository.
3. The system is demonstrated to the customer. Customer accepts the system if he is satisfied and we will bill him.

V. System Components

In the following you find a summary of components which will be released into the git repository and a list of software components to set up the process. All components can be openly pulled. Also everybody can review all components and add comments. The comments are shown publicly.

A. Control Software

The control software is code which drives motors and contains control interfaces to move the robot head from one point to another. There are currently two sorts of movements possible, which is linear movement and point to point movement. To translate the movements to motor signals, we need to transform Cartesian coordinates to motor positions. The software component "inverse transformation" is dealing with this task.

We purposely use UML to generate automatically code from the UML diagrams. The advantage of UML is that graphical representation of code is easier to review than pure C++ code. There are certainly drawbacks with UML, such as fixes in pure C++ code must be phased back to UML which is often quite cumbersome. There are two sorts of diagrams we store into the git repository: UML Class Diagrams and UML Object Diagrams.

Currently we have a linear programming approach for our control software. This means that each task is executed in sequence. We will enhance this to a real time operating system containing scheduler, dispatcher etc. So we will be able to execute tasks concurrently and we will be able to set priorities to tasks. E.g. a task with low priority can be the driver for the display. Below you find a list of software components stored in the git repository.

- UML Class Diagrams
- UML Object Diagrams
- inverse transformation algorithm
- point to point movement algorithm
- linear movement algorithm
- control program
- future topic: scheduler
- future topic: preemptive multitasking code
- testcases

B. Simulation Software

The reason to use simulation software before testing on real hardware is to save the effort to build up the real hardware, just for testing the software. Revealing software problems might lead to the decision to stop building up the hardware, which saves time and effort. Another reason is to continuously testing the software released inside the repository. Single commits can be pulled in an automatic fashion and drivers can be built from them. Tools such as Jenkins are suited for this kind of tasks.

The simulation system we build up with will be a Mathematica model. It offers the feature to generate the source code from it. Therefor we release Mathematica models into the git repository. The model is currently pure static, meaning that the motor dynamics and the robot systems inertia is not taken into account. In future we will implement a dynamic model as well. We will configure the static model of the robot in a configuration file. This means that if we have a design change of the robot, we often do not have to touch the Mathematica model, but just need to update the configuration file.

In future we will also combine the stl design elements with the simulation, so the robot system can be tested visually. The following simulation components are released into the repository.

- mathematic models (Mathematica export files, source code)
- pure static model
- dynamic model (future topic)
- configuration file for the robot's geometry
- future topic: graphical simulation elements

C. Robot Controller

We have set to our self the requirement that we don't design any component of the robot controller, such as the ARM32 board and the step motor driver board, besides the interface connector between the components. We put weight to the fact that these components must be purchasable. Here comes the consumer goods ordering system into effect. Jenkins triggers the ordering process with the ordering data inside the commit. The following descriptions for automatic ordering are stored into the repository

- power supply
- Controller hardware: ARM32 board, step motor controller and the ordering data

Other files going into the repository:

- Interface connector for ARM32 board and step controller board.
- design description
- complete layout plan with pure purchasable standard components and their interface connectors.

D. Robot

The component files (design files) of the robot are released as Sketchup files, files in stl format and files in gcode format. Downloading the gcode format files makes it us easy to automatically move the files to the 3D Printer and print them out. This can be partially triggered by Jenkins again. There are a few parts from the robot, which cannot (or should not) be printed, such as bearings, screws, carriages, step motors etc. Here comes the consumer good ordering system into play again. So the person who releases purchasable parts, must fill in the commit comments with data, where to buy the part. Jenkins triggers the ordering application, gathers the ordering data from the commit and automatically orders the parts. The following is release in git.

- components (design files)
- 3D printer parts
- carriages
- step motors

E. Software

We mostly use public available software to accomplish the system we are proposing. First the git repository, which is more and more becoming a popular repository for code and other design files. We use Gerrit as a review tool, so the community can check about newly available released components for review. Every community member can download the newly components and check them for their own purpose. Gerrit offers a mechanism to evaluate the new component by pressing radio buttons with attributes such as: ok to release or prefer not to release. At the end, the administrator has the right to overrule the comments of the public community if required, but in general, the comments are taken very seriously.

Jenkins is a continuous integration tool using a plugin from Gerrit. E.g. if there is a control software release, Jenkins can download automatically the files, compile them and test them against a simulation. Then even Jenkins can set automatically an evaluation about the released code, depending if the testcases which have been running against the simulation were successful or not.

The following software tools are currently used to accomplish the proposed system:

- git repository
- Gerrit review tool
- Jenkins, a continuous integration tool
- consumer goods ordering system

VI. Outcome and Conclusion

In this paper we presented a new design methodology for machinery which includes the public community in reviewing the machinery parts consisting of software code, controllers, machinery parts and other. We are still in an early stage. GIT, Jenkins etc. are set up, but the team is struggling to get used to the work process. Also the automatic features are still missing, such as downloading code to flash the memory.

In our case we studied mainly the design a delta robot, but we believe that the concept can be extended to any kind of machinery. In Figure 8 you can see an improved version of the robot "Cherry Pi II" released in thingiverse which we have redesigned trying the proposed concept. This is the first robot which is undergoing the proposed

process. The redesign involved the removal of mechanical play of the carriages. Another delta robot is currently created, see Figure 9. There you can see a redesign of the carriage using only low cost parts. Our study does currently not involve the open public, but only a selected number of persons.

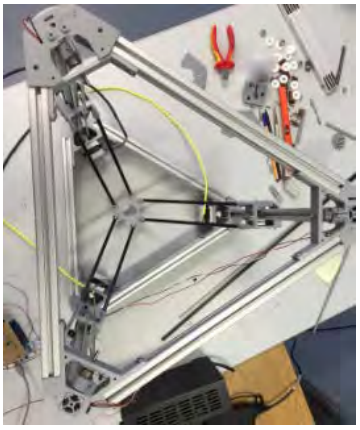


Figure 8. Cherry Pi II robot

In future we will involve the public community too. The incentive of the public community for reviewing the components is that they can download all parts of the machinery for free to build devices for their own purpose. Therefor the open public does have an interest in continuously improving the design located in the public repository.

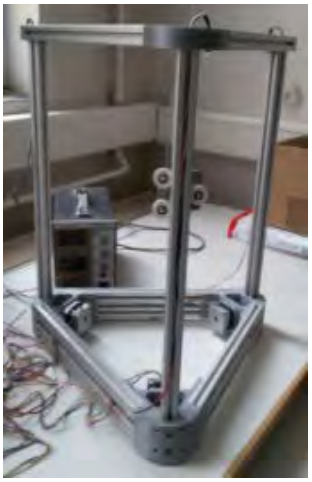


Figure 9. Redesign of the carriage

The administration of the repository needs funding, so we came up with a business idea, that we offer a purchasable service to our partners to test out the machineries according to their demands. The customer can take a look at the machines we offer on the website and convince them, that we have the competence to build such machinery. The service include consulting, simulating and testing.

ACKNOWLEDGMENT

We would like to thank the University of Applied Science Albstadt-Sigmaringen for sponsoring this project within the Program Fit4Research. Without the funding we would have no opportunity to start the project described in this paper. Also we would like to thank Florian Wiest to show so many new research idea within this area. Also we thank our students Sinem Guruplar and Goncagül Albayrak working on their theses providing the outcomes presented in this paper.

REFERENCES

[1] Oracle White Paper, The Department of Defense (DoD) and Open Source Software, 11/2013

[2] G. Madey, V. Freeh and Renee Tynan, The Open Source Software Development Phenomen: An analysis based on social network theory.

[3] Open Sources: Voices from the Open Source Revolution, 1-56592-582-3

[4] J. Asundi, Software Engineering Lessons from Open Source Projects, Position Pater for the 1st Workshop on Open Source Software, ICSE-2001

[5] R. Vallance, S. Kiani, and S. Nayfeh, Open Design of Manufacturing Equipment, MIT

[6] Sören Sonnenburg et. al., The Need for Open Source Software in Machine Learning, Journal of Machine Learning Research, 2007.

[7] S. Maurer, S. Scothmer Open Source Software: The new intellectual property paradigm, National Bureau of Economic Research (NBER) working paper, 2006

[8] P. Rigby et. al., Peer Review on Open-Source Software Projects: Parameters, Statistical Models, and Theory

[9] B. Lampson, R. Sproull, An Open Operating System for a Single-User Machine

[10] NEMA. 1996, Motion/Position Control Motors, Controls, and Feedback Devices, NEMA Standard MG-7 Rev. 1. National Electrical Manufacturers Association (NEMA), Rosslyn, VA. September 1996.

[11] Free Software Foundation, GNU General Public License, http://www.gnu.org/licenses/gpl-3.0.en.html

[12] Open Design Foundation. "Open Design Definition", http://www.opendesign.org/odd.html

[13] Maker culture, https://en.wikipedia.org/wiki/Maker_culture

[14] humanity's first general-purpose self-replicating manufacturing machine, http://reprap.org/

[15] Thingiverse, http://www.thingiverse.com/

[16] Openbuild, http://openbuilds.org/

[17] www.**amazon**.com/oc/dash-button

[18] Pinshape, https://pinshape.com/

[19] 3D Shook, www.3dshook.com/

[20] Trinpy, https://www.trinpy.com

[21] CGTrader, https://www.cgtrader.com/

[22] 3D-File Market, http://3dfilemarket.com/

[23] Cults, https://cults3d.com/

[24] Makershop, https://www.makershop.co/

[25] MyMiniFactory, https://www.myminifactory.com/de/

[26] Redpah, https://www.redpah.com/

[27] Shapeking, http://www.shapeking.com/de/

[28] Yeggi, http://www.yeggi.com

[29] Youmagine, https://www.youmagine.com/

[30] Threeding, https://www.threeding.com/index.php

[31] TurboSquid, http://www.turbosquid.com/

[32] STL Finder, http://www.stlfinder.com/

[33] Florian Wiest, Die Zukunftsmacher, Schwäbisches Tagblatt, 6.3.20

DC VIS

International Conference on

DISTRIBUTED COMPUTING, VISUALIZATION SYSTEMS AND BIOMEDICAL ENGINEERING

Steering Committee

Chairs: Karolj Skala, Ruđer Bošković Institute, Zagreb, Croatia
Roman Trobec, Jožef Stefan Institute, Ljubljana, Slovenia
Uroš Stanič, Biomedical Research Institute-BRIS, Ljubljana, Slovenia

Members: Enis Afgan, Ruđer Bošković Institute, Zagreb, Croatia
Almir Badnjević, International Burch University, Sarajevo, Bosnia and Herzegovina
Piotr Bala, Nicolaus Copernicus University, Toruń, Poland
Leo Budin, University of Zagreb, Croatia
Borut Geršak, University Medical Center, Ljubljana, Slovenia
Simeon Grazio, "Sisters of Mercy" University Hospital Centre, Zagreb, Croatia
Gordan Gulan, University of Rijeka, Croatia
Yike Guo, Imperial College London, UK
Ladislav Hluchý, Slovak Academy of Sciences, Bratislava, Slovakia
Željko Jeričević, University of Rijeka, Croatia
Peter Kacsuk, Hungarian Academy of Science, Budapest, Hungary
Aneta Karaivanova, Bulgarian Academy of Sciences, Sofia, Bulgaria
Zalika Klemenc-Ketiš, University of Maribor, Slovenia
Charles Loomis, Laboratory of the Linear Accelerator, Orsay, France
Luděk Matyska, Masaryk University, Brno, Czech Republic
Željka Mihajlović, University of Zagreb, Croatia
Damjan Miklavčič, University of Ljubljana, Slovenia
Tonka Poplas Susič, Health Centre Ljubljana, Slovenia
Laszlo Szirmay-Kalos, Technical University of Budapest, Hungary
Tibor Vámos, Hungarian Academy of Sciences, Budapest, Hungary
Matjaž Veselko, University Medical Center, Ljubljana, Slovenia
Yingwei Wang, Universitiy of Prince Edward Island, Charlottetown, Canada

INVITED PAPER

Views on the Role and Importance of Dew Computing in the Service and Control Technology

Zorislav Šojat, Karolj Skala
Ruđer Bošković Institute, Centre for informatics and Computing, Zagreb, Croatia
sojat@irb.hr, skala@irb.hr

Abstract - Modern day computing paradigms foster for a huge community of involved participants from almost the entire spectrum of human endeavour. For computing and data processing there are individual Computers, their Clusters, Grids, and, finally, the Clouds. For pure data communication there is the Internet, and for the Human-understandable Information Communication for example the World Wide Web. The rapid development of hand-held mobile devices with high computational capabilities and Internet connectivity enabled certain parts of Clouds to be "lowered" into the so called "thin clients". This led to development of the Fog-Computing Paradigm as well as development of the Internet of Things (IoT) and Internet of Everything (IoE) concepts.

However, the most significant amount of information processing all around us is done on the lowest possible computing level, outright connected to the physical environment and mostly directly controlling our human immediate surroundings. These "invisible" information processing devices we find in our car's motor, in the refrigerator, the gas boiler, air-conditioners, wending machines, musical instruments, radio-receivers, home entertainment systems, traffic-controls, theatres, lights, wood-burning stoves, and ubiquitously all over the industry and in industrial products. These devices, which are neither at the cloud/fog edge, nor even at the mobile edge, but rather at the physical edge of computing are the basis of the Dew Computing Paradigm.

The merits of seamlessly integrating those "dew" devices into the Cloud - Fog - Dew Computing hierarchy are enormous, for individuals, the public and industrial sectors, the scientific community and the commercial sector, by bettering the physical and communicational, as well as the intellectual, immediate human environment.

In the possibility of developing integrated home management/entertainment/maintenance systems, self-organising traffic-control systems, intelligent driver suggestion systems, coordinated building/car/traffic pollution control systems, real-time hospital systems with all patient and equipment status and control collaborating with the medical staff, fully consistent synaesthetic artistic performances including artists and independent individuals ("active public") from wide apart, power distribution peek filtering, self-reorganisation and mutual cooperation systems based on informed behaviour of individual power consumption elements, emergency systems which cooperate with the town traffic, etc., etc., the Dew-Computing paradigm shows the way towards the Distributed Information Services Environment (DISE), and finally towards the present civilisation's aim of establishment of a Global Information Processing Environment (GIPE).

It is therefore essential, through Research, Innovation and Development, to explore the realm of possibilities of Dew Computing, solve the basic problems of integration of the "dew" level with the higher level Dew-Fog-Cloud hierarchy, with special attention to the necessity of information (not only data) processing and communication, and demonstrate the viability and high effectiveness of the developed architecture in several areas of human endeavour through real life implementations. The present scientific and technological main objective is to provide the concepts, methods and proof-of-concept implementations that are moving Dew Computing from a theoretical/experimental concept to a validated technology. Finally, it will be necessary to define and standardise the basics of the Dew Computing Architecture, Language and Ontology, which is a necessity for the seamless integration of the emerging new Global Information Processing Architecture into the Fog and Cloud Paradigms, as a way towards the above mentioned civilisation goals.

I. INTRODUCTION

From the very begin of the "computer era" our civilisation tends to become a global information society, and the early developments were already aimed towards computer and data/information interconnectedness. The science fiction of 1960's/1970's, as well as many scientists predicting possible future, actually envisioned the present day informatisation. The primary predisposition for that development was achieving substantial speed improvements of processors and their interconnections, and the ability to process increasing amounts of data in memory. Through considerable advances of computer science High-performance distributed computing systems were founded on the Grid computing paradigm, while scalable distributed computing systems evolved through the Cloud, later Fog and now the new computing paradigm called Dew computing. The exponential growth of Internet of Things (IoT), Internet of Everything (IoE) and Big Data processing led to the necessity of horizontal and vertical scalability of distributed resources at multiple levels or layers. To facilitate the rapidly developing complex distributed computer systems and meet the performance, availability, reliability, manageability, and cost adapted for requirements of today's, and specially tomorrow's users, a three-layer scalable distributed platform evolved (Cloud - Fog - Dew computing).

Dew computing is a new computing paradigm, which appeared after the widespread acceptance of Cloud Computing. The initial research of Dew computing started from the Cloud Computing architecture, and the natural metaphor of a cloud lead to the metaphor of the dew. Dew computing can be seen as a ground, or physical, edge of Cloud Computing.

II. The Story of the Dew

The spread of data processing devices into all segments and all layers of modern life in our civilisation has to be, naturally, followed by specific system architectures. By realising that there is a vast amount of processing power and storage spread around the world into quite minute hardware systems, it became obvious that this widespread distribution of resources could be used on a more global scale. Actually this type of system architectures development, the development of architectures which are not computer oriented, but oriented towards reasonably effective exploitation of distributed resources, can be seen to have proceeded in several stages.

The first of these stages we can already recognise in the development of the environment of the first large mainframes and supercomputers during the 1960-ies. A typical supercomputer (often as a multiprocessor setup, specially from the second half of the 1960-ies) would actually use quite a few other computers to perform the input/output and telecommunications functions, as well as to prepare data and analyse the results. The cost and speed of the supercomputer was such that it, although generally able to stand as a fully self-standing computer system, would not be used for menial tasks, except during installation and testing, and therefore would actually never be used without it's distributed computers processing environment. The other important distribution was through the data communication channels towards other computing systems. A typical large university, institution or multinational company would have had all their major computing resources always communicational interconnected. Those were the starting points of development of mentioned higher level communication and processing distribution (i.e. not computer-oriented) architectures from which probably the Internet is the most generally known.

The Era of Supercomputers lasted till the second half of the 1990-ies. In the meantime the interconnection of computers in the world already lead to the Internet, and a lot of distribution problems were tackled and solved - like the distribution of storage, through the Network File System (NFS), presently probably the most used distributed file system; the distribution of tasks and requests, like for example in the Remote Procedure Call protocol (RPC); or the distribution of time, through the community developed Network Time Protocol (NTP); and the distribution of hypertext human-readable information through the World Wide Web (WWW), invented by Tim Berners-Lee in 1989. With the vast spread of computing (i.e. mostly model, data and less information processing), the availability of faster and faster individual hardware processing devices, and the availability of high-quality inter-computer communication links and protocols, and finally, the huge availability of extremely cheap "personal" computers, naturally lead to a lot of investigation into the possibility of using a bunch of the so-called "of the shelf" "personal" computers in parallel, as to enable high-speed parallel processing.

The task of organising a bunch of heterogeneous interconnected computers into a stable, usable, fast and easily programmable and maintainable system has shown to be quite complex. Though the interconnection of the cluster elements was solved through the long-time stabilised Ethernet standards and internet protocols, the major problem, not solved to this very day, was and is the already mentioned organisation of a heap of individual computers into a consistent and easily controllable, programmable and fault-tolerant, reconfigurable environment of sometimes hugely heterogeneous individual elements.

To get a reliable network of computer clusters on which computational jobs could be executed at the best place when it is most convenient, the Grid Computing paradigm was developed. A lot of scientific areas benefit from the Grid paradigm, as much of their jobs can be Workflows of quite different data sizes and processing complexity either overall or at the level of workflow elements. Huge experimental data sets, the general cycle of scientific exploration and computer use (e.g. experimentation, thinking, writing, programming...), etc. generally do not require the computer results interactively, therefore the Grid architecture enables them much higher processing equipment utilization than it would be possible on any local system.

Unfortunately, outside some scientific and other complex computing environments (like e.g. statistics, business information processing, film rendering etc.), the Grid paradigm, due to its "batch-processing" general approach, is not easily or at all directly applicable. Obviously a non-interactive environment is neither appropriate for the general public, nor for artists and scientists which deal with real-time phenomena.

Out of that situation a new notion in computing paradigms and appropriate architectural structures have been born. A Cloud. The Cloud Paradigm (or, if you prefer, actually the Cloud Allegory) means organisation of such a processing structure that computing, data retention, data retrieval etc. is done somewhere-anywhere in the world by systems of clusters, and then provided to anybody needing data storage and/or processing for any reason whatsoever, by the so-called "service providers". The user interfaces changed and adapted themselves to the needs of a more general public, and a lot of cloud-compatible programmes are written to accustom many differing end-user needs. However, the Clouds themselves, as opposed to the Grids, are not user-level controllable or programmable, and therefore there is a lot of proprietary solutions, though, in the manner of a Cloud, this huge heterogeneity of the individual computers, clusters and grids of clusters inside any particular available Cloud (provided by a cloud-provider) is, or at least tends to be, much hidden from the end-user.

As with all the nomenclature of different generations or system architectures mentioned, all of them started

being implemented and used often much earlier than their names have been invented. The development process in science has in this sense a two-sided push. Something is invented, starts being used, becomes a seed for more global thought, a name and paradigm is invented, scientifically investigated and than applied to generalise the usage of such inventions. Solving problems of generalisation of a specific paradigm leads to a new circle of invention... in a kind of spiralling development.

After seeing the Cloud we realised that we have to bring processing also closer to the Earth, to the user/consumer edge. So the Cloud slowly settled into the Fog.

Probably the first actual fully fledged Fog paradigm idea was the development of the Postscript language - a language specifically designed for Printers, not for Computers. Actually there is no reason whatsoever (except speed) not to use your printer to calculate for example appropriate sine waves and musical envelopes for your computer resident audio generation programme. So we here immediately see the Fog Paradigm - using "lowest" level equipment to process part of the wider processing need. The adoption of Postscript for real-time calculations and rendering in the SUN NeWS and the NeXTStep display systems, which allowed to run Display Postscript applications on remote viewing foreground stations (the Display Servers) controlled and programmable from the background servers (the Display Clients), enabled splitting the processing of certain applications seamlessly into distinct computers, the central ones (the Clients) and the peripheral ones (the Servers)[1].

This idea, combined with the, in that time already ubiquitous, HyperText Markup Language (HTML) led to the development, again by SUN Microsystems, of the Java programming language, and the JavaScript branch. To mention is also the standardisation development done for the Wireless Application Protocol (WAP), and the associated Wireless Mark-up Language (WML), which, though in intensive use quite short lived, did define some important points for further development. Many other such user-station execution languages/environments have also been developed, like e.g. Flash etc.

So nowadays the situation is that we do have a kind of primitive "operating system based on device-distributed 'intelligence'" common to all modern web-browsers, and which allows a lot of processing to be done on the end-user (i.e. hierarchically lowest processing level) equipment - from rendering properly human-readable text, through decoding video and audio streams, up to executing fully functional full-fledged important applications, or even parts of scientific crowd-supported distributed computing environment jobs.

Now we are in the situation mentioned above that we have to scientifically develop this which we already perceive as a computing paradigm, and gave it the name of Clouds Laying Low - the Fog. In this effort we will have to solve a lot of problems to make the usage of Things (as predicted by the buzz-name Internet of Things) seamlessly coordinated through the Fog Architecture with the Clouds, Grids, Clusters and individual Computers, as to be really usable in the emerging Global Information Processing Environment.

So, with the Fog we came down from the Clouds to the "intelligent" Things in our everyday environment, like our mobile communicators, or our home-entertainment television-sets, or even our refrigerators and ovens, many of which show the tendency to be connected to the Internet. The real challenge now is to find a common communication framework, information structure and language to be able to have a grasp over the uncountable possibilities of such an integrated global system - the addition of the Internet of Things and the Fog paradigm onto the already existing huge global processing infrastructure. And one could imagine a highly refined home oven helping, in free time, to calculate the ephemeris of some just discovered asteroid, or to help with its own cooking experience some newly opened restaurant on the other side of the world. Believe it or not, this is the goal our present day development is aiming at.

III. WHAT DEW CAN DO

However, as all cells in our body are not primarily information processors, like neurons, but the vast majority of them are physically controlling their global environment - our common human body (although also processing and exchanging information), the same applies to all of our Environment, natural or human made.

The vast majority of things we see all around us have no processing power at all: the house, the car motor, the bathtub, the road or the chair. Or do they?

As opposed to your internet-connected oven, which has to have quite a high processing power already due to the complexity of internet communication itself, your car motor can work without any computing equipment at all - although we use presently more and more processors to fine-tune the mechanical behaviour of the motor based on environmental conditions. However, there is actually no possibility for this kind of processors to calculate PI for some scientist on the other side of the globe. But they do process a lot of environment information and generate a lot of control and status data.

The road does not by the first sight seem as something which would process information. However, a lot of modern roads do sense traffic, know their surface temperature and even some of them can sense if they are covered by rain, snow, ice or mud. That is, in other words, even our modern roads collect environment data and distribute it to higher levels, primarily for traffic counting and direct traffic control if necessary.

"Intelligent homes", massage chairs, hydro-massage bathtubs, all of those have more and more integrated control processors. The water and heating boilers, the air-conditioners, the washing-machines... are all controlled by

[1] Please beware the terminology. A display terminal is a Server, which serves the remote requests of a possibly larger number of compute-server-type display Clients. The human looks at the output of the Server, and is mostly interested in the calculations of the Clients.

processors gathering environment information and controlling the same environment in a specific way.

This is the very Dew of Computing. And these simple computers influence our everyday physical life in an extremely strong way, with the aim to keep our immediate environment as close to our wishes as possible. Their physical influence is truly immediate and tangible, an influence on the physical quality of human life which no Computer, Cluster, Grid, Cloud or Fog device could ever dream to have.

The viability and high effectiveness of the Dew computing paradigm and emerging architecture is quite obvious in many important areas of real life implementations (weather/soil conditions, traffic, medical emergency, smart home, distribution balancing, environment protection etc., etc.).

The merits of seamlessly integrating "dew" devices into the Cloud - Fog - Dew Computing hierarchy are enormous, for individuals, the public and industrial sectors, the scientific community and the commercial sector, by bettering the physical and communication, as well as the intellectual, immediate human environment.

In the possibility of developing integrated home management/entertainment/maintenance systems, self-organising traffic-control systems, intelligent driver suggestion systems, coordinated building/car/traffic pollution control systems, real-time hospital systems with all patient and equipment status and control collaborating with the medical staff, load balancing energy and water distribution systems, emergency systems which cooperate with the town traffic, fully consistent synaesthetic artistic performances including artists and independent individuals ("active public") from wide apart, power distribution peek filtering, self-reorganisation and mutual cooperation systems based on informed behaviour of individual power consumption elements, etc., etc.[2], the Dew-Computing paradigm shows the way towards the Distributed Information Services Environment (DISE), and finally towards the present civilisation's aim of establishment of a Global Information Processing Environment (GIPE).

IV. THE DEW

The mission of Dew computing is to fully realize the potentials of personal/body and human environment-control computers in symbiosis with Cloud services. However, the complexity of interconnectivity, and even more the heterogeneity of equipment used through these paradigms is drastically growing as we approach the Dew computing level.

Whereas the Cloud and Fog Computing paradigms address the principles of operation on complex computations using massive amounts of data, which is context-free, Dew computing is context-aware, giving the meaning to data being processed. Data is context-free, while information is data with accompanying meta-data. The meta-data places the data in a specific context. Information processing (data + meta-data) enables the application of self-organisation on all levels, providing a much wider scope of problems which can be solved, and also allows more comprehensive analysis and deeper knowledge discovery. The Cloud and Fog Computing paradigms operate on huge quantities of raw data generated by specific Things, via predefined services. Since the raw data is out of context, the services need to be tailored and application specific, requiring data driven decisions. Building an integrated scalable heterogeneous distributed computing environment from the level of Cloud or Fog is currently not plausible (or viable), as the lack of contextual information disallows generic integration of all processing elements.

To solve the problems of everyday human/computer interaction (communication) through ergonomic and human-oriented end-user interfaces, and therefore also to allow the adaptability of the whole user environment system to human needs and wishes, Dew computing has to be based on Information-oriented processing rather than being Data-oriented.

In the Dew computing scenario individual Things are responsible for collecting/generating the raw data, and are thus the only components of the computing ecosystem which are completely aware of the context the data were generated in, therefore dew-devices actually must produce and exchange Information.

The idea behind the Dew computing is primarily to use the resources at the lowest level as self-organising systems solving everyday human and industrial environment problems. Significant advances and savings can be achieved by using such physical-edge computing in the areas of, for example, traffic control, power distribution (balancing, peek protection, etc.), integrated home systems, medical systems, emergency public services systems, industrial control systems (specifically those necessitating high level, e.g. Cloud, services) etc.

On the physical-edge level a vast majority of Dew-devices will not be connected primarily or at all to the Internet or to the mobile phone network for communication with other global computing environment elements. As the impact of Dew-devices is primarily space-oriented, i.e. the device's actions are directly connected, by action or communication, with their immediate physical and informational surrounding, i.e. environment, for communication those devices will mostly use near-communication means as e.g. direct cabling, free line optical communication, short to medium distance radio (e.g. Bluetooth) or even electricity distribution line communication (e.g. X-10).

Actually the "dew" devices are exactly on the opposite side of the High Performance Computing (HPC), we could call them Low Performance Information Processing (LPIP). HPC presupposes high data and computing rates, whereas LPIP presupposes low information and decision rates. However, just to mention, the sum of the processing power of all LPIP devices on our planet highly supersedes

[2] It is essential to understand that on the level of Dew-Computing the Human is the prime and final decision maker (it is up to a person, the human being, if it will head the warnings of the car, or if it will necessitate water even if the system is trying to reduce overall consumption!).

the processing power of all HPC systems, though their specialisations are on opposite sides of the human endeavour spectrum, therefore they are actually incommensurable.

It is obvious from the above that Dew computing in the sense of information processing actually presupposes extremely fine granulation of parallelism. On the lowest level there are devices with slow serial processing speeds and small to extremely small data/information retention and processing possibilities. Therefore the Dew paradigm is radically different from the Fog/Cloud or Cluster/Grid paradigms, as its lowest and low level devices cannot be used in a "conventional" programmable way, but they have to cooperate on that lowest level to solve human environment problems or needs, and be able to pass (and consume) information from all hierarchical levels.

Dew computing must be a generic and general architecture, which must be able to include all of the present and future hybrid and extremely heterogeneous information gatherers, information distributors, information processors, information presenters and information consumers, at the lowest level directly connected to everyday machines which are part of the common modern human physical environment, and at the highest level interconnected into the global information processing and distribution system.

To achieve this it is essential to develop and define a generic communication *language* and a full *ontology*, and not just communication protocols, which are applicable in data processing environments, but not in information processing systems. Only in such a way we will be able to cope with the complexity and heterogeneity of future equipment, and be able to harness the vast possibilities which a Global Information Processing System can offer.

V. CONLUSION

The Dew-Computing Paradigm may well be the final missing ingredient to the computing development, transforming the all-pervading clusters, grids, clouds and fogs of computers into a human-helping Global Information Processing Environment. Solving properly many of the envisioned problems at this very beginning of a New Information Processing Era may well spare our children of coping with myriads of incompatible and pseudo-compatible systems which necessitate "higher magical skills" to get them to do what is necessary, needed or wanted.

ACKNOWLEDGEMENTS

This work was, in part, supported by the European Commission through the Horizon 2020: EGI Engange and INDIGO DC projects.

REFERENCES

[1] John Backus “Can Programming Be Liberated from the von Neumann Style? A Functional Style and Its Algebra of Programs”, Communications of the ACM, Vol 21, No. 8, pp. 613-641, August 1978.,

[2] Borna Bebek, Zorislav Šojat: "Optimal Distribution of Organisational Resources", Diplomatic Academy Year–Book, Edition: Central European Initiative – International Conference "Diplomacy for the twenty-first century: knowledge management", Vol. 2/2000, No. 1, HR: Ministarstvo vanjskih poslova i europskih integracija; 2000

[3] R. Buyya, C. Yeo, S. Venugopal, and J. B. I. Brandic: “Cloud Computing and Emerging IT Platforms: Vision, hype, and reality for delivering computing as the 5th utility”, Future Generation Computer Systems, vol. 25, no. 6, pp. 599–616, 2009.

[4] EGI Glossary; European Grid Initiative WIKI, https://wiki.egi.eu/wiki/Glossary_V1, 22/2/2014, 21:56

[5] Roland N Ibbett & Nigel P Topham (1996): “HIGH PERFORMANCE COMPUTER ARCHITECTURES - A Historical Perspective”, http://homepages.inf.ed.ac.uk/rni/comp-arch/index.html

[6] Sinisa Marin, Robert J. Thorpe, Zorislav Sojat: “A Modular Robot Programming System”, 1989; London, GB: RDP Technology.

[7] Sinisa Marin, Mihajlo Ristic, Zorislav Sojat: “An Implementation of a Novel Method for Concurrent Process Control in Robot Programming”, Third International Symposium on Robotics and Manufacturing: Research, Education and Application, ISRAM ‘90, Burnaby, BC/CA; 1990

[8] Janko Mršić-Floegel, Derek Reynolds, Zorislav Šojat, Marco Biancessi, Stefano Sala: Data Communications, Patent WO2002025897 A1, priority 13/9/2000, http://www.google.com/patents/WO2002025897A1, retreived 29/2/2016.

[9] John Owens "Data Level Parallelism (2)" EEC 171 Parallel Architectures, UC Davis, http://www.nvidia.com/content/cudazone/cudau/courses/ucdavis/lectures/dlp2.pdf 1/3/2014, 6:13

[10] Karolj Skala, Zorislav Sojat: "Towards a Grid Applicable Parallel Architecture Machine.", Computational Science - ICCS 2004, 4th International Conference, Kraków, Poland, June 6-9, 2004, Proceedings, Part III; 2004

[11] K. Skala, Z. Sojat: "Image Programming for Scientific Visualization by Cluster Computing", Autonomic and Autonomous Systems and International Conference on Networking and Services, ICAS-ICNS; 2005;

[12] Karolj Skala, Davor Davidovic, Enis Afgan, Ivan Sovic, Zorislav Sojat: “Scalable Distributed Computing Hierarchy: Cloud, Fog and Dew Computing”, Open Journal of Cloud Computing (OJCC), 2(1), Pages 16-24, 2015, https://www.ronpub.com/publications/OJCC_2015v2i1n03_Skala.pdf

[13] Zorislav Sojat: "Operating System Based on Device Distributed Intelligence", 1st Orwellian Symposium, Baden Baden, Germany; 1984

[14] Zorislav Sojat, Sinisa Marin: “ISOCOM 20 Filter System User Documentation: Flow Through Filter Language and Graphics Organisation Language”, 1992; BTS, Purley, GB.

[15] Zorislav Šojat: "Nanoračunarstvo i prirodno distribuirani paralelizam", Glasilo Instituta Ruđer Bošković. 3(7/8):20-22.; 2002

[16] Zorislav Sojat, Karolj Skala: "Multiple Programme Single Data Stream Approach to Grid Programming", Hypermedia and Grid Systems, Opatija, Croatia; 2004

[17] Zorislav Šojat, Tomislav Ćosić, Karolj Skala: "Virtue — A different approach to human/computer interaction" Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014, 37th International Convention on. IEEE, 2014, http://grgur.irb.hr/Library/Shoyat.Cyosicy.Skala.Virtue_A_different_approach_to_human_computer_interaction.pdf

[18] Y. Wang, “Cloud-dew architecture,” International Journal of Cloud Computing, vol. 4, no. 3, pp. 199–210, 2015.

PAPERS

DISTRIBUTED COMPUTING AND CLOUD COMPUTING

Parameters that affect the parallel execution speed of programs in multi-core processor computers

Valon Xhafa*, Flaka Dika**
* University of Prishtina/Department of Computer Engineering, Prishtina, Kosovo
**University of Vienna/Faculty of Computer Science, Vienna, Austria
valon.xhafa@uni-pr.edu

Abstract - The speed of the computers work depends directly on the number of processor cores used in the parallel execution of the programs. But, there are some other parameters that have impact on the computer speed, like maximum allowed number of threads, size of cache memory inside the processors and its organization. To determine the impact of each of these parameters, we have experimented with different types of computers, measuring the speed of their work during solving a certain problem and during image processing. To compare impacts of particular parameters on the computers speed, the results we have presented graphically, through joint diagrams for each parameter.

I. INTRODUCTION

Currently, on the market there are different types of computers with multi-core processors. Use of computers with these features has an important impact on their parallel work. Users have difficulties to distinguish which is the best solution for them, knowing clock speed of processor, or the information for the type of processor, like Core i3, Core i5 or Core i7 (in this case often it is assumed that we have to do with the processor that have 3, 5 or 7 cores) [5].

Programs that are written to be executed on a single processor or in a processor core without having the possibility to execute more instructions immediately, are known as serial programs, or sequential programs [1]. On the other side, programs that are written for simultaneous execution in more cores, or more logical processor, are known as parallel programs [2][13]. Parallel execution of programs in a computer is based on the parallel work of processor cores. But, using the Intel's technology that is known as Hyper-Threading it is possible for each core to work as two logical processor [5], respectively it is possible for the processor with 4 cores to work parallel with 8 logical processors.

The parallel work of microprocessors has a great influence on the calculation speed of computers which is important for the massive use of computers in communication through the Internet, also searching on the Web, in the processing of various biometric information, monitoring the climate changes in real time, scientific researches, in complicated three-dimensional games on the computer etc.

Measuring the performance of computers during parallel execution of programs is not an easy task, because in the speed of their work, except clock speed, indicates cores number, model, the support of Hyper-Threading technology, organization of cache memory in separated levels [7] etc.

II. EXPERIMENTAL RESULTS

In order to get relevant conclusions, initially we have experimented with numerous computers that have different processor types of Intel producer. In order to measure the speed of the computer work during experimentation we have used an ordinary algorithm for multiplication of two square matrices with different dimensions, from 100 x 100 to 1000 x 1000 (10 cases). For parallelism of loops are used methods of programming language C# [3][10][12], which are included in Microsoft Visual Studio 2010 package.

We have started the experimentation with computers that are built using previous processor 2 Duo, which have cache memory only on the level L2, then with Core i3, Core i5 and Core i7 processors (short i3, i5, i7) [16], whose memory are on the levels L2 and L3 and processor used for servers, such as Xeon type [15]. Finally from all the computers that are used for experimentation, we have singled out only the results of 15 computers with Intel processor, whose characteristics are given in TABLE I, without differentiating as desktop or notebook computer.

TABLE I. LIST OF USED COMPUTERS

Computer	Model	Clockspeed (GHz)	L2 (KB)	L3 (KB)	Cores	Logical processors
C1	i5	2.90	1024	6144	4	4
C2	i5	3.10	1024	6144	4	4
C3	E3	3.10	1024	8192	4	4
C4	i5	1.70	256	3072	2	4
C5	i3	2.40	256	3072	2	4
C6	i5	3.33	512	4096	2	4
C7	i7	2.20	256	6144	4	8
C8	i3	3.30	512	3072	2	4
C9	2 Duo	2.93	3072	0	2	2
C10	i5	1.60	256	3072	2	4
C11	i7	3.60	1024	8192	4	8
C12	2 Duo	2.66	3072	0	2	2
C13	i5	2.50	256	3072	2	4
C14	i7	2.00	256	4096	2	4
C15	i3	2.40	256	3072	2	4

From the given table we can see that the model of the processor does not automatically determines the number of cores, neither the number of logical processors. Thus,

for example, computer C7 belongs to model i7 and and contains 4 cores and 8 logical processors [8]. But, his memories L2 and L3 are smaller, e.g. than at the model i5 processor of computer C2. On the other side, the computer C14 has model i7 processor, but with two cores and four logical processor [9]. At the same time, processors of earlier generation 2 Duo, which are included in the computers C9 and C12 have relatively large cache memory on the level L2, but don't have the level L3 memory and in parallel work they have shown lower speed than computers with processors of next generation.

A. Influence of number of cores

To have a clearer overview in how much does the parallel execution of programs indicates, initially we have experimented with serial execution testing program for matrices multiplication, image horizontally flipping, and Object tracking by activating only one core of the processor. Then, we have measured the execution time of a parallel program when all cores are active, respectively all logical processors. Finally, using the results of measurements, we found the execution speedup and the level of parallelism of the program. Having no chance to present all the results obtained, in the following section we have given only the graphically results for four computers from the list given in TABLE I, who have clock speed approximately the same level (above 3 GHz).

1. Serial execution time

Results of serial execution time (Ts) for four specific computers are shown in the diagram given in Fig.1.
The diagram shows that computer C11 is faster, which uses a processor that is ranked in the group of processors that are actually fastest [8], the i7 model, with maximum clock speed of 3.60 GHz, and with 4 cores (short 4-c) and 8 logical processors (short 8-lp). Computer C2, whether it is the i5 model, has competitive speed with computer C11. This is seen in the daily ranking of Intel processors [9], where he has high position in the list. Simultaneously, the computer C6 with i5 processor is slower than computer C8 who have i3 processor. From the given diagram is clear that the curves are ranked based on the clock speed of processors.

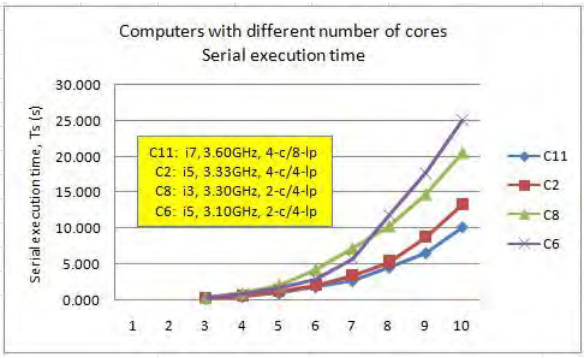


Figure 1. Serial execution time of the program

2. Parallel execution time

During the parallel execution of the testing program in separate computers, parallel execution time (Tp) have big difference from serial execution time (Ts), which can be seen from the diagram given in Fig.2.

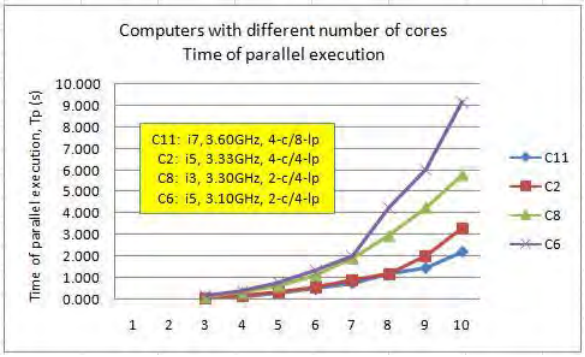


Figure 2. Parallel execution time of testing program

Here it is seen that computer C11 has higher speed, and computer C6 is slower, which is consistent with their speed in serial work. But, in this case execution times are significantly shorter. Thus, for example, computer C11 multiplication of matrices with dimensions 1000 x 1000 performs for more than 2 seconds, and during serial execution is needed more than 10 seconds. Parallel execution time in computers C11 and C2 do not differ much between them, despite being different models of processors (i7 and i5). During the parallel work of computer C11 are 8 logical processor active, and in computer C2 - 4 logical processors. On the other side, as it is shown in the diagram, with increased dimensions of matrices, the speed of the computer C6 drops dramatically, despite having the i5 processor and its clock speed of 3.10 GHz. From what was said and by experimentation that we had with other computers, we can conclude that the speed of parallel execution of the program can't be determined based on the model of the processor (i3, i5 or i7), or number of cores, because it is also influenced by the clock speed of processor. Same results are achieved during image processing. In our case, we have taken an image to process by using image flipping algorithms. First of all, we have measured the execution time of the serial image processing. After that, the algorithm used for image flipping is parallelized in order to process the image in more cores. Later on, we have measured the execution time of parallel image processing too. Achieved results are shown in Fig. 3. We have experimented with computers with 1, 2, and 4 cores,

Cores	Time of executions (ms)
1	783
2	456
4	234

Figure 3. Execution time for horizontally image flipping

also in the figure are shown the average time of execution of horizontally image flipping.

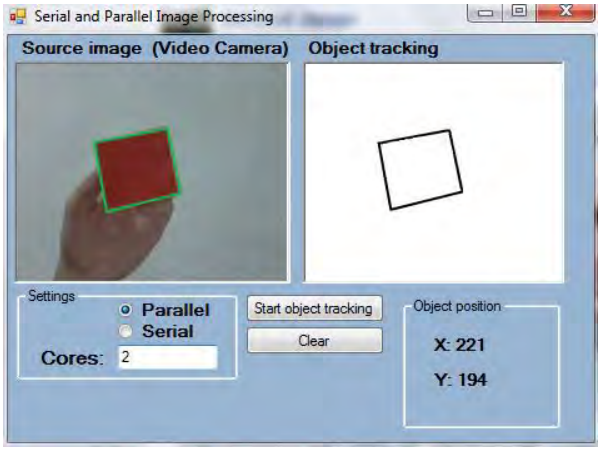


Figure 4. The application interface used in measuring the execution speed of Contour tracking algorithm

In computers with processor with 1 core, we have used the serial algorithm for image flipping, thus the average time of execution is 783 ms. In computers with 2 or 4 cores processor, the parallel algorithm for image flipping is used. The execution speed
are 456 ms and 234 ms, when 2 core and 4 core processors are used. Based on this, the execution speed is boosted when processors with more cores are used. The interface of the application used for testing and measuring the execution time of horizontally image flipping algorithms is shown in Fig.5. Part of the interface is also the option to choose the number of processor cores to be used during the processing. Choosing the execution mode between Parallel and Serial is also part of the interface. In another experiment we have taken the case of object tracking application which is a time-sensitive application. The application frames the video camera into images, than for every image uses Contour Tracking algorithms to detect objects contours. The application interface is shown in Fig. 4. The detected contours of an object are transferred and displayed in the Object tracking panel of the application from which is measured the relative position of the object center. During the experiments we have measured the serial execution of the Contour Tracking algorithm and the parallel execution as well. The results are shown in Fig. 6. Based on the diagram, there is as an increase of speed during parallel

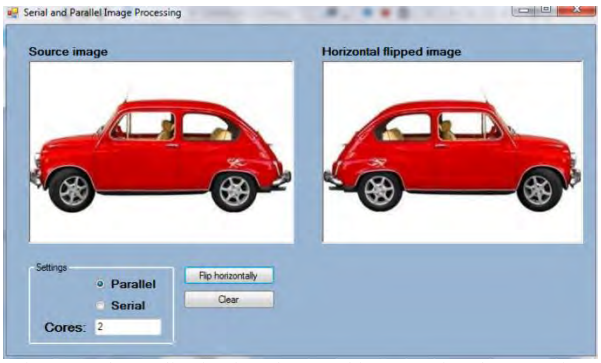


Figure 5. The application interface used in measuring the execution speed of image flipping algorithms

execution in finding the object contours. As mentioned above, object tracking applications are time-sensitive because their aim is to detect and find object position in a specific time, therefore delays between framing and contour tracking of an object should be decreased. From the diagram we can conclude that delays decrease with the parallel processing of framed image using multi-core processors. But the rate of the delay decrees can't be determined based on the model of the processor (i3, i5 or i7), or number of cores, because it is also influenced by the clock speed of processor, this is the same as during matrices multiplication.

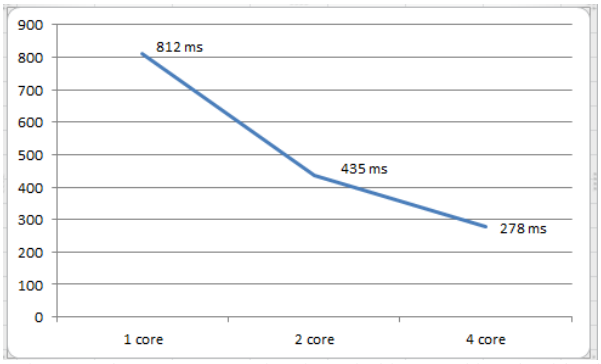


Figure 6. The average execution time of Object Tracking algorithm in processors with different number of cores

3. *Execution speedup*

If it's known the serial execution time (Ts) and parallel execution time (Tp) of the testing program, can be calculated execution speedup (S) through the expression [6]

$$S = \frac{Ts}{Tp} \qquad (1)$$

Using the experimental results mentioned above, execution speedup of 4 computers taken into consideration, it will look like is given in Fig. 7. Even the execution speedup maintain the sequence of curves for specific computers as well as in diagrams for execution time. But curves have variations, with increases and decreases. At the end of their, obvious that only the computer C11 continues accelerating growth (finally

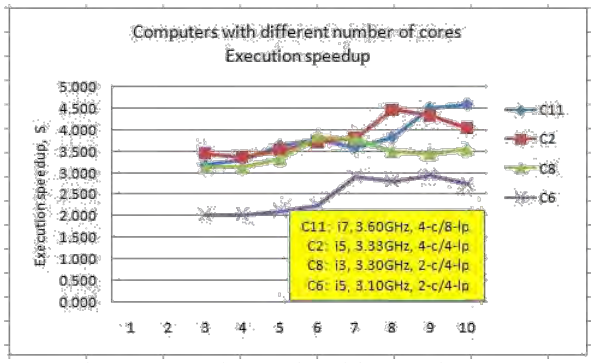


Figure 7. Execution speedup of the execution process

reaches the value of 4.5 times), while the three other computers values stay the same, or decrease. As it is seen from the diagram given, and even from experimentation that we had with other computers too, the execution speedup is greater if the number of operations that are executed (e.g. multiplication of matrices with large dimensions) increases up to a certain thresholds. During this there is no continual growth execution speedup, which shows that parallelism is not perfect.

B. The impact of cache

In the speed of the computer work it is important even the size of the cache memory. Models of the last generations of processors contain cache memory on the three levels. Initially we have experimented with computers of prior generations, like 2 Duo processor cores. In TABLE I we've included computers C9 and C12, whose processors have only memory L2 level and we have noticed that compared with other computers in parallel work are extremely slow, despite having relatively higher clock speeds. Then we experimented with other 6 computers also included in the TABLE I and the results obtained for them are presented in the diagram given in Fig. 8.

From the given diagram it is shown that computer C7 with i7 processor is faster and has large cache memory on the third level whether that it has the lowest clock speed. Computers C2 and C1 have approximately the same speed, with processor of model i5, clock speed much greater than computer C7 and have a maximum cache memory at two levels. In the overall ranking of

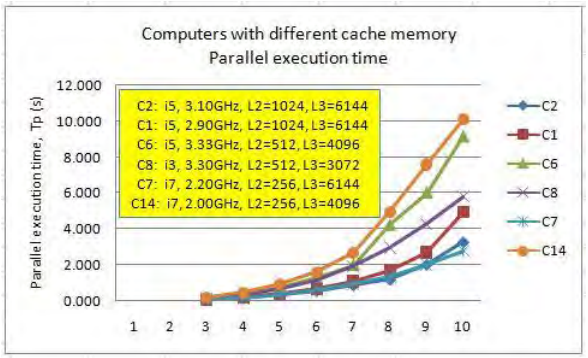


Figure 8. Parallel execution time of computers with processors that have different cache memory

processor of computer C1 have higher rank than the computer C2 [9]. Speed of computers C8 and C6, despite that they have processors i3 and i5, in their speed ranking have higher speed clock and cache memory size. The Computer C14, with i7 processor, have the highest execution time.

C. The level of parallelism

The maximum execution speedup of program can be determined by Amdahl's Law [4][6]

$$S = \frac{1}{(1-P) + \frac{P}{N}} \qquad (2)$$

Here, N is the number of processors, and P represents the relative percentage of the program parallelism. Starting from the given Law, through mathematical operations can be derived expression for the calculation of the level of parallelism of a program:

$$P = \frac{N(S-1)}{S(N-1)} \qquad (3)$$

Since in the measurement of execution speedup values are calculated using (1), and knowing the number of cores, i.e. the number of logical processors (N), the level of parallelism can be calculated through (3). Thus, for computers with i7 processor, the level of parallelism will appear as is given in Fig.9. From the diagram provided and experimentation that we have done with other computers, it is shown that the relative level of parallelism of the program execution is over 0.65 and just a bit under 1.00, which is directly linked with the speed of parallel execution of the program.

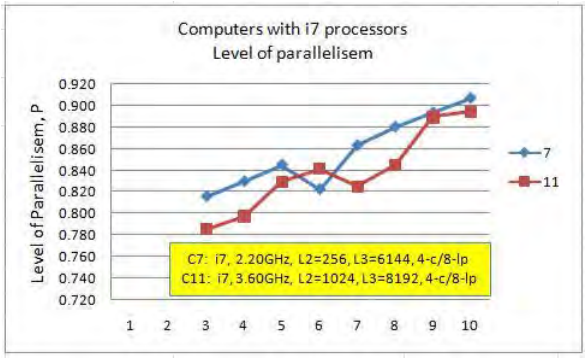


Figure 9. Parallel execution time of computers with processors that have different cache memory

D. The maximum rate of parallelism

During the parallel work of computer, the operating system determines the number of threads that are activated [10], or being allocated as needed certain number of threads to programs that run in parallel [11]. However, this can be done by the user too, by limiting the number of threads. If parallel execution relates only to a program, computer work can be optimized if the number of threads deal as far as the number of cores, i.e. the number of logical processors.

To show this we experimented with several computers. But in Fig. 10 have just given diagram of

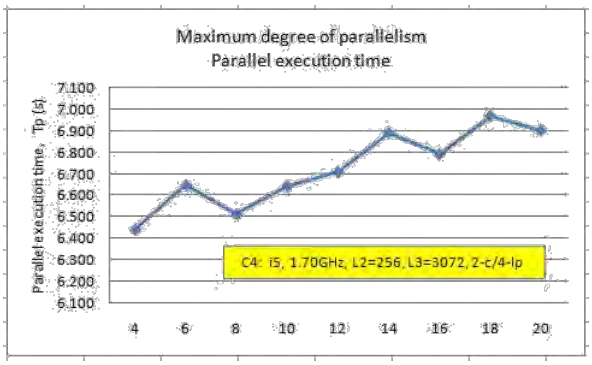


Figure 10. The rate of parallelism

parallel execution time of testing computer C4, limiting the number of threads between the values 4 and 20. The given diagram clearly shows the execution time is minimal if the maximum degree of parallelism obtained 4, which corresponds to the number of processor cores of the computer where the program is executed. Execution time increases if it's allowed a higher degree of parallelism, which is observed in most of the computers with which we experimented. From this we can conclude that the higher speed of parallel program execution is usually achieved when the maximum degree of parallelism is taken as far as the number of processor cores.

III. CONLUSION

The speed of the serial execution of the program for different computers mainly is indicated directly from the processor clock speed. But the speed of parallel execution of the program can't be determined based on the model of the processor, or the number of cores, because the cache memory size and clock speed have an important impact. Execution speedup of the computer, depends on the number of processor cores, as well as the level of parallelism of the program being executed, these conclusions are achieved during the experiments of matrices multiplication as well as during image processing. But the level of a parallelism of a program execution depends directly on the speed of the parallel work of the computer. In computers with which we experimented, level of parallelism is over 0.65 and just under 1. In order to optimize the parallel speed of computer we can manipulate with the allowed number of threads. Experimental results show that if a single program is executed, the speed of computer work is usually better if the number of threads deal as far as the number of processor cores.

REFERENCES

[1] Blaise Barney. Intoduction to Parallel Computing. Lawrence Livermore National Laboratory. https://computing.llnl.gov/tutorials/parallel_comp/ [Accessed on 10.9.2014]

[2] Ananth Grama, Anshul Gupta, George Karypis, Vipin Kumar. Introduction to Parallel Computing, Second Edition. Addison Wesley, January 16, 2003

[3] Parallel Programming with .Net. http://blogs.msdn.com/b/pfxteam/ . [Accessed on 17.8.2014]

[4] Frank Willmore. Parallel Programming. The University of Texas at Austin https://portal.tacc.utexas.edu/c/document_library/get_file?uuid=d7f742a0-42c8-4407-b3c6-9fe1cbda14f1&groupId=13601 [Accessed on 25.11.2014]

[5] http://www.pcworld.idg.com.au/article/386100/what_difference_between_an_intel_core_i3_i5_i7_/ [Accessed on 18.12.2014]

[6] Jan Zapletal. Amdahl's and Gustafson's laws. VSB - Technical University of Ostrava, 2009

[7] Shared Memory Multiprocessors. http://www.inf.ed.ac.uk/teaching/courses/ pa/Notes/lecture04-multi.pdf [Accessed on 23.12.2013]

[8] http://www.pcworld.idg.com.au/article/386100/what_difference_between_an_intel_core_i3_i5_i7_/ [Accessed on 16.12.2014]

[9] http://www.cpubenchmark.net [Accessed on 18.1.2015]

[10] Optimal number of threads per core. http://stackoverflow.com/questions/1718465/optimal-number-of-threads-per-core [Accessed on 12.3.2013]

[11] https://msdn.microsoft.com/en-us/library/system.threading.tasks.paralleloptions.maxdegreeofparallelism%28v=vs.110%29.aspx [Accessed on 23.1.2014]

[12] Gaston C. Hillar. Proffesional Parallel Programming with C#, Wiley Publishing, Inc., Indianapolis, Indiana, 2011

[13] Peter S. Pacheco. An Introduction to Parallel Programming, University of San Francisco, Morgan Kaufmann Publishers, 2011

[14] Hans-Wolfgang Loidl. Parallel Programing in C#. Heriot-Watt University, Edinburgh, 2012

[15] http://www.intel.com/content/www/us/en/processors/xeon/xeon-processor-e7-family.html [Accessed on 18.1.2015]

[16] http://www.intel.com/content/www/us/en/processors/core/5th-gen-core-processor-family.html [Accessed on 15.7.2014]

Federated Computing on the Web: the UNICORE Portal

Maria Petrova-El Sayed*, Krzysztof Benedyczak†, Andrzej Rutkowski† and Bernd Schuller*
*Jülich Supercomputing Centre
Forschungszentrum Jülich GmbH, Germany
Email: {m.petrova, b.schuller}@fz-juelich.de
†Interdisciplinary Centre for Mathematical and Computational Modelling
Warsaw University, Poland
Email: golbi@icm.edu.pl, rudy@mat.umk.pl

***Abstract*—As modern science requires modern approaches, vast collaborations comprising federated resources on heterogeneous computing systems rise to meet the current scientific challenges. Due to their size and complexity these computing systems become demanding and could further complicate the scientific process. As a result, scientists are overrun with the necessity of additional technical experience that lies outside their domain.**

UNICORE is a middleware which serves as an abstraction layer to mask the technical details and ensure an easy and unified access to data and computation over federated infrastructures. The Portal is the newest client in the UNICORE portfolio providing web access to data and computing systems. With the rising demand of having an up-to-date user-friendly graphical interface and access to computational resources and data from any device at any point in time, the Portal meets the contemporary needs of the dynamic world wide web.

This paper describes the functionality of the Portal and its advantages over other clients. It also discusses security and authentication methods offered by the Portal and presents use cases and client customisation from our practice. It concludes with ideas about future work and extension points for further simplification for the scientific use.

I. Introduction

Complex scientific projects involve multiple computer simulations and calculations. In order to achieve their goals, scientists need to benefit from heterogeneous computing systems and resources, which can be distributed all over the world. The working process in such an environment is very challenging for scientists with limited technical background. UNICORE is a middleware software that ensures a smoother path for researchers, allowing them to concentrate on the actual scientific problem by reducing the amount of required technical knowledge. The development of UNICORE [1] dates back to 1996 and was initiated in an effort to provide access to the three largest German High-Performance Computing (HPC) centres. Its core design principles comprise: abstraction of resource-specific details, openness and extensibility, security, operating system independence and autonomy of resource providers. The software is available as open source from the SourceForge repository [2] under a commercially friendly licence.

The traditional end-user clients in the UNICORE portfolio are the UNICORE Commandline Client (UCC) and the desktop UNICORE Rich Client (URC). However, in the current era of technological advancement the world wide web is the main means to enable access to data. In an answer to contemporary demands, the UNICORE Portal provides a web-based Graphical User Interface (GUI). It offers not only a user-friendly GUI for work with distributed computing systems, but also serves as an access point to compute centres. Web technologies have multiple advantages, for instance, constant availability from an arbitrary location.

The remainder of this paper is structured as follows. Section II describes the functionality of the Portal. Supported authentication methods are covered in section III, which includes a concise presentation of the new Unity [3] component that plays a crucial role in many of the new development and integration scenarios. Section IV focuses on client customisations and extensions from our practice. In section V we discuss some advantages of the Portal by making a high-level comparison between the different UNICORE clients as well as by comparing the Portal to other existing web solutions. Finally, in section VI we summarise and conclude with future work.

II. Functionality of the UNICORE Portal

The UNICORE Portal facilitates computation in federated systems and makes the grid more accessible. It is also used as a convenient access point to a particular compute centre.

The functionality of the Portal covers generic use cases. Before being able to compute or manage data via the Portal, the user needs to authenticate himself at login. More information on available authentication methods is presented in section III. After successfully logging in, the user is directed to the home page of the Portal. The home page is to a great extent configurable by an administrator and can contain either a Really Simple Syndication (RSS) feed by choice, or a personalised design from a local Hypertext Markup Language (HTML) file. The Portal allows multiple customisations (done by an administrator only) like hiding certain views, branding the GUI, adding and exchanging logos and much more.

A. Job Submission

Focus points of the Portal functionality are execution of jobs, fetching of results and presentation of outcomes. By clicking on *New Job* from a flat navigation menu, the user is forwarded to a view for job submission (Fig. 1). Here he can select an application from a pre-filled list, which is being dynamically updated. The information about the applications in this list is fetched from the configuration files of the UNICORE sites which are accessible to the user.

In this view, job descriptions and application specific input parameters can be entered in a graphical form. The latter would have normally been put into a script and manually submitted for execution. A GUI client eliminates the need to look into a detailed job script by automatically generating it in a ready to be submitted state. The user has to only prepare the necessary data and set the desired arguments. He can as well upload files from a remote or local storage. When clicking on the corresponding button, a window with the accessible storages will be open. Then the user can browse a selected storage and navigate through its folders in a manner resembling work within a file system. He also has the possibility to create new files where he can write necessary algorithms or other required input information. Most of the local or remote files can be changed on the fly through a file editor. However, restrictions on the size prohibit opening and editing of too large files from within the Portal.

A stage-out destination of objects to a desired output location can also be specified for every job. For example, the user can define that the job output be copied to a remotely located folder of choice.

Already in this view the user can see which sites he has access to and can choose where to send his job for execution. If none selected, the UNICORE server will take the decision based on a thorough brokering algorithm, which takes into account the needed resources, the load of the sites, etc. The required resources for the successful completion of his job can also be specified by the user in a comfortable GUI form. Furthermore, multiple jobs can be edited at the same time in multiple windows. Full job definitions can be saved and restored, thus allowing for parameter fitting and re-submission.

B. Monitoring

The monitoring of the job results is represented in a dedicated job table on a separate view. The table contains the job name, status, submission time, tags and other details. The update for each status change or a new job submission happens automatically. The user can sort, filter and search the table entries, as well as re-submit, abort or delete jobs. He can also browse the working directory. If the job's execution completed successfully, the working directory contains all job related files including the fetched output results. These can normally be viewed in the browser or downloaded to the local machine. As expected, the user can monitor and operate only with his own data and only after a successful authentication. In case of an unsuccessful job execution, the working directory stays empty as no output has been produced. However, the user can view the reason for failure by clicking on a button which opens a window with more detailed information.

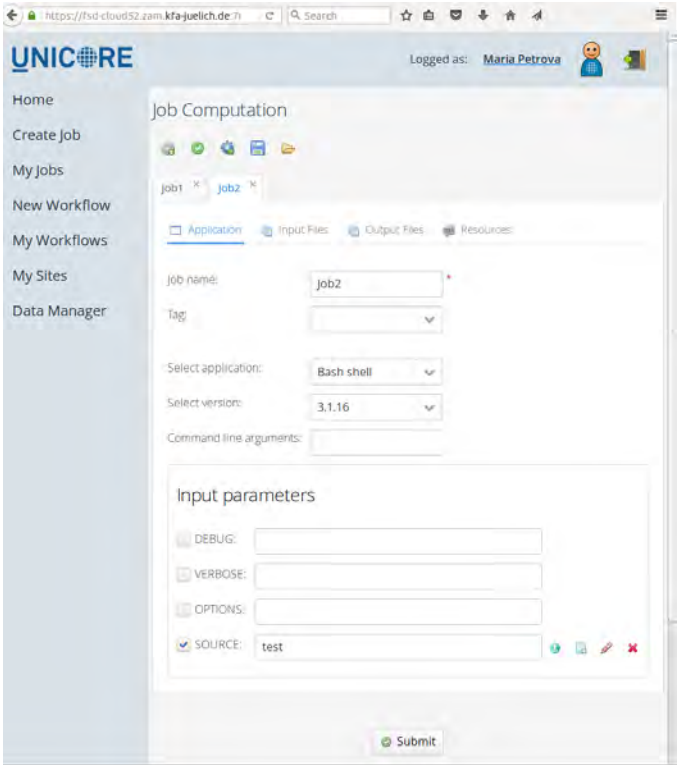


Fig. 1. The view for preparation and submission of generic jobs.

C. Data Manager

The Portal offers a powerful data manager which assists users in data-related tasks. The intuitive look & feel ensures a comprehensive GUI for transfer, upload or download of data within different locations. Users are able to create, preview and delete files in the browser as well as edit their content on the fly. The data manager supports both client-server upload/download and server-server transfers with several available data transport protocols. High throughput protocols such as UNICORE File Transfer Protocol (UFTP) [4] are also included.

D. User Workspace

Another typical characteristic is the so called *user workspace*. It enables the storage of data directly in the Portal in order for them to be accessible from anywhere on the web. For example, the data from submitted jobs are being preserved in the workspace. Each user has his own separate workspace, which is always created by default. However, for security reasons the Portal administrator has the possibility to hide the workspace from the user, thus restricting his access. If access is granted, the user has full control over the data from his workspace with the help of the data manager.

E. Workflows

The Portal supports the creation and submission of elementary workflows. With a few mouse clicks the user can compose a graph of jobs with the required connections, parameters, imports and exports. The workflow editor is meant to be simple and intuitive and serves only the basic needs, i.e. only task sequences. All the advanced features, such as conditional and loops, are currently omitted.

Identical to the job table, there is a separate table, containing all submitted workflows and their details. The user can browse through a workflow's working directory in a manner resembling browsing through a file system.

F. Sites view

The sites view contains information about all the accessible to the user sites. The user can track each site's availability, the number of its cores, its installed applications and more. If the geographic location is configured, it will be marked on Google Maps for a clear representation of the infrastructure. The sites view is just informative and has no influence on computing.

G. Configurations

The Portal is designed to be flexible and allows for multiple configurations. All settings including authentication modes, accessible compute centres, graphical customizations of logos, home page, main menu entries and more are defined in a properties file. This file is managed by a so called Portal administrator who is responsible for the initial set up of a Portal instance. The end user normally does not have access to the configuration file and does not need to know or manage the Portal instance itself. His working process starts from the authentication point.

H. Internationalisation

Last but not least, the Portal supports internationalisation. The default language is English. However, adding a new language to the GUI is uncomplicated and does not require any programming effort. All the messages are defined in properties files which can be translated. With a simple configuration from a Portal administrator, the new language will appear for selection together with all available languages at the login screen.

III. Security and Authentication

The certificates of the X.509 Public Key Infrastructure system are commonly used by grid middleware as a base of authentication and trust delegation, needed to coordinate distributed job execution. This approach turned out to be a major obstacle in grid adoption: certificates are difficult to obtain, distribute, refresh and manage. Conversion between different formats, installation in clients, etc. are problematic as well.

Web scenarios are even more troublesome. The Portal is a middleware client and is accessed itself by the user's browser. Therefore, the user's private key cannot be retrieved by the grid client—i.e. the Portal. Even if the user installs a certificate in the browser, it can only be used for the authentication step. As browsers do not support grid-specific trust delegation, which is needed by the grid, this has to be handled differently.

The UNICORE Portal approach to authentication and trust delegation addresses the aforementioned issues. It preserves backwards compatibility and does not scarify an overall security. Authentication in the Portal is designed in a modular way. Each authentication module provides support for a different authentication mechanism. In general these mechanisms can be grouped into two categories: *local mechanisms* and *remote mechanisms*.

The local authentication is handled fully by the Portal. Currently implemented local authentication modes comprise:

- An X.509 certificate—pre-installed in the user's web browser.
- User name and password—for simple installations as it lacks advanced features such as password reset.
- Demonstration account—a shared pseudo account that can be enabled for demonstrative access.

In order to use the UNICORE infrastructure after any of the local authentication mechanisms, the user is prompted to upload a so called trust delegation token to the Portal. The only exception is the demonstrative account. This is a solution to the delegation problem, as the private key stays at the user's machine. The drawback is that it requires the user to start a Java Web Start application (served by the Portal) and provide X.509 credentials to this application. The application generates a time-limited delegation token and uploads it to the Portal. This extra step is executed roughly once a month depending on the user's trust in the Portal. Nevertheless, it can be seen as an obstacle for portal users as it requires possession of a certificate and a potentially difficult action.

To offer first class authentication, the UNICORE Portal also supports a remote authentication over the Security Assertion Markup Language (SAML) protocol. The SAML authentication using the web POST profile is described in the OASIS SAML Standard [5]. Besides the authentication itself, support for an additional trust delegation assertion was added. Such assertion must be generated by a universally trusted third party service and returned to the Portal together with the SAML authentication assertion.

The role of a trusted delegation and authentication authority is fulfilled by a Unity service [3]. Unity is a powerful identity provider, featuring support for outsourcing authentication to different external identity providers, using both OAuth and SAML protocols. Through Unity, a middleware can be configured with an arbitrary login mechanism. Unity can enable one or more from the standard solutions such as password, authentication with data centre Lightweight Directory Access Protocol (LDAP), home Identity Provider (IdP) of a SAML federation, etc. It can even use a social identity provider like Google, Facebook or Github to name a few.

It should be underlined that the configuration of the selected authentication mechanisms is done by a Portal administrator and is transparent for the end user. Thus the choice of allowed authentication sources is fully controllable and the Portal's

authentication uses UNICORE trust delegation without any tricks like generation of short lived certificates or storing user's private key in a web server.

IV. Use cases and user customisation

A. Generic use case

Already in its default configuration, the UNICORE Portal can be used for a wide variety of use cases. Access to a federated infrastructure of HPC resources is one of the primary design goals.

As a concrete example for this use case, we would like to mention the Human Brain Project (HBP) [6], a European FET flagship project. The HBP targets a wide range of topics, aiming at a deeper understanding of the human brain and attempting to leverage this understanding for new technological advancements. The HBP operates on HPC resources, integrating major supercomputing sites in Jülich, Barcelona, Bologna and Lugano as well as cloud storage and other resources. The HBP uses a single-sign-on system based on OpenID Connect (OIDC) [7]. With UNICORE deployed as underlying middleware, the UNICORE Portal is used as a simple way to access the HPC and data resources. The Portal is configured to use Unity for authentication, and is thus integrated with the single-sign-on functionality. The user benefits already in this simple scenario from the abstraction provided by UNICORE and the single-sign-on functionality; it is no longer required to know and understand how to log into the various resources, or be aware of the various batch schedulers that are used at the various sites (e.g., Slurm and IBM LoadLeveler). The Portal is used to prepare and submit so called *generic jobs*. The user interface for these generic jobs is generated at runtime based on metadata provided by the UNICORE server.

B. Custom plugins

The generic approach is sufficient for the majority of applications, especially those whose users were traditionally preparing input files in some well-established format. However, there are cases when a more sophisticated interface is welcomed. For example, some applications require special resource allocations such as CPU, number of nodes, used memory or architectures, which might all depend on other application parameters. A specialised application interface can assist with these complex resource settings. Customisations to the interface can also improve the preparation of input. For instance, selecting parts of the image for a visual analysis is much more convenient when clicking on the image itself than by providing numeric pixel coordinates. Even though the generic application description is quite powerful and allows for expressing the most common dependencies, some complex relationships between input settings may require special implementation. Furthermore, for many applications output visualisation is a key feature and its proper placement in the Portal is crucial for effective research.

As the above aspects were known from the very beginning, the core Portal is designed as a base for development of applications or domain specific plugins. A plugin is free to implement an arbitrary user interface as well as to contribute with one or more entries to the main menu. The administrator has control over all menu entries and can select which of them to be shown on the GUI. This allows for tailoring of the Portal to particular domain needs even to a degree where a generic job interface is not presented at all.

The development of a portal plugin is restricted to the base technology of the Portal, which includes Vaadin [8] as the underlying framework for building user interfaces in Java and Spring [9] as a container for objects. These restrictions are counterbalanced by the possibility to use internal portal components and features to speed up the plugin development. For example, multiple graphical elements such as the job table can be re-used from the existing implementation. Already present are also the various user authentication methods as well as a grid security context for grid interaction like asynchronous discovery of resources and jobs. Thus, the development can be focused purely on application domain aspects.

We present here two examples of such custom modules developed in PL-Grid Plus and PL-Grid NG projects. The first one is named *SinusMed*. *SinusMed* processes an input, that being a series of computed tomography (CT) images, which form a 3D image of a human head. It detects all air-filled head areas. Areas are subsequently marked and categorised using reference data sets prepared by medical doctors. Results of the simulation comprise several sets of layered masks on the input image, which must be presented in a visual form.

Another example is the *AngioMerge* application. It is used to synchronise several angiograms, which are taken periodically with an interval of several minutes. The output is supposed to be displayed in 3D, optionally as an animation.

The portal modules for *SinusMed* and *AngioMerge* allow for a simplified job preparation, which focuses on the actual application input and automatically pre-sets the required resources for each application. Therefore, the user is not burdened with additional knowledge about the computing infrastructure being used. What is more, applications submit the jobs with the help of a UNICORE broker so that site selection is fully automated.

The most attractive features of the above application specific interfaces are found in the output visualisation part. The *SinusMed* directly shows the original CT images in the browser (Fig. 2), enabling the overlay of synchronised masking layers, computed by the application. The *AngioMerge* module embeds a WebGL 3D interface (Fig. 3) allowing for interactive viewing of the output and even playing the resulting animation.

V. Comparison with other solutions

A. With Other UNICORE Clients

The UCC is a commandline tool that enables job and workflow management from a shell or scripting environment [10]. It covers all features of the UNICORE service layer including data transfers, job and workflow submission and monitoring, results fetching, etc. Among its characteristic qualities is the possibility to submit multiple jobs in an automated fashion via a batch mode. To speed up the work, new customised commands can be defined by the user. However, the UCC is

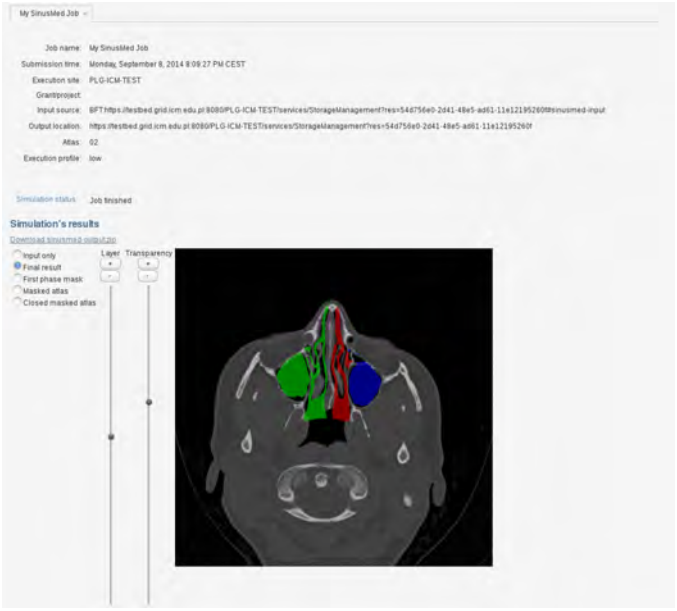


Fig. 2. Results of sinuses detection embedded in the Portal interface after a simulation in a low quality (fast) mode.

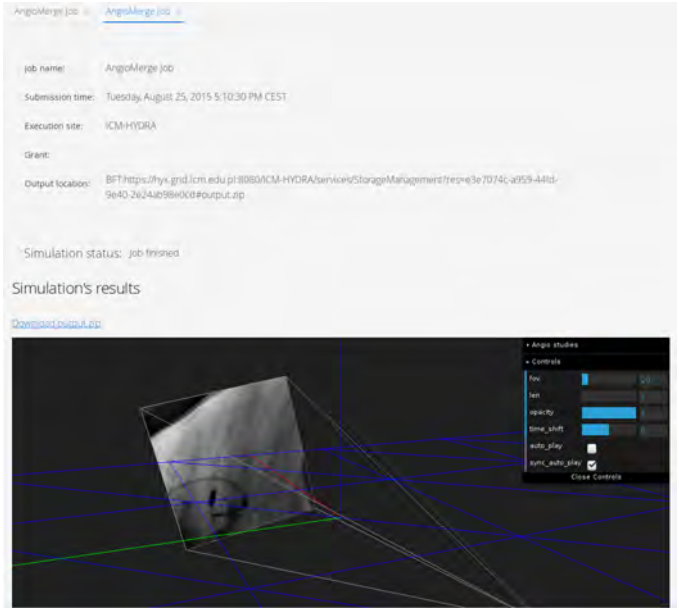


Fig. 3. Output visualisation of the *AngioMerge* module directly in the portal web interface.

oriented towards experienced computer users, who find working with a terminal faster and more convenient. In contrast, there are scientists who welcome an easy to use GUI.

The URC [11] is a standalone client that provides a detailed graphical interface for the UNICORE functionality. Similar to UCC, the URC supports all features like job and workflow creation and submission, retrieval of results, data transfers and grid browsing. In its essence the software is based on Eclipse [12], which makes it easily extensible via the plugin mechanism. Specific to the URC are the different perspectives, which hide or show features from the UI in accordance to how advanced the user is. The major strength of this client is its powerful workflow editor with constructs like loops and conditional statements for steering the control flow and enabling a fully automated work process.

In comparison to the URC, the functionality of the Portal's workflow editor is elementary. We have taken this approach due to feedback from our users. In order to use the advanced structures, many users still need the assistance of a UNICORE specialist. This is one of the main reasons why we decided to keep the design of the Portal's workflow editor simple and light-weight. Thus, we ensure a comprehensive and intuitive interface that covers the basic need, whereas the complex use cases can benefit from the well-developed and well-established editor of the URC.

Federated computing on the web has multiple advantages to desktop clients. Unlike the UCC and the URC, the Portal is available from any location and can be accessed by various devices. While standalone clients require installation, administration and regular updates done by the user himself, a web portal releases the user of this responsibility, thus saving time and effort.

B. With Other Existing Web Clients

There are several existing web portals that concentrate on HPC. One such solution is the InSilicoLab [13]. It is an environment built to assist researchers in their specific domain. In principle it has been designed to be generic, but it is currently used only by projects in chemistry [14]. Similar to the UNICORE Portal, this tool offers access to distributed resources on the web, job preparation, submission and monitoring as well as fetching of results and data management. However, the InSilicoLab has rather specialised into domain-specific interfaces. In contrast, the Portal is more generic with the intention to be an easily extensible solution that can be beneficial in various research domains. The Portal is also more flexible with its numerous authentication possibilities through the integration with Unity.

Another web client for HPC is Adaptive Computing's Moab Viewpoint [15]. The Moab Viewpoint is solely concentrated on job and data management and does not cover any workflow features. However, it offers an interface for creating application templates, which helps in optimizing application run times and reduces errors. Another enticing feature, which is not present in the UNICORE Portal, is the existence of two GUI modes: an end user and an administrator one. The administrator mode enables supplementary functionality like reporting on resource utilisation, troubleshooting of the workload, editing or cancellation of jobs and tracking of node-usage. The Portal can be customized by an administrator through a properties file but this should not be confused with the administrator mode from Viewpoint. The administration in the Portal is not job based and offers no GUI. It concerns the enabled options like languages, authentication methods, menu entries, etc. for a particular Portal installation. One drawback of

the Moab Viewpoint is that it is a commercial software, which means extra costs incurring for the users. Furthermore, Viewpoint does not operate in a distributed environment, but only supports work on a singular site.

VI. Conclusion

The UNICORE Portal is an open source solution that facilitates scientific work by enabling federated computing on the web. It offers a convenient and friendly GUI for job and workflow submission, monitoring and data management. The authentication to the Portal is done either locally or with the help of Unity, which allows for using numerous authentication mechanisms, including social and federated ones, without the need to possess an X.509 certificate.

The default Portal implementation covers the generic use case, which is sufficient for most projects. We presented as an example the integration of the Portal in the HBP platform. The Portal is also flexible and, due to its modular design, it is easily extensible by self-developed plugins for specific applications. The creation of a new plugin requires solely the effort that is necessary to add the particularities of the targeted domain, whereas the base portal components and features can be easily re-used. We described two use cases with custom modules, namely the *SinusMed* and the *AngioMerge* applications.

We also made a concise comparison between the Portal and the other clients in the UNICORE portfolio with their strong points and disadvantages. We discussed related web solutions for HPC environments.

The Portal's development does not stop here. It faces multiple milestones on its future path. One of them, which is under current development, is the workflow template feature. The Portal will offer the option to import a ready template file. The latter will be parsed and a familiar graphical form will be automatically generated. Then the user will need to fill the data and submit the updated workflow without having to compose a new one from scratch. This approach will ensure a comfortable parameter fitting, simple re-use of workflows and uncomplicated workflow submission.

Another idea that is also in progress is the integration with data sharing solutions. Scientists will be able to share with others their files, such as workflow templates, from within the Portal's interface.

Acknowledgement

This work was made possible with assistance of the PL-Grid Plus project, contract number: POIG.02.03.00-00-096/10, and the PL-Grid NG project POIG.02.03.00-12-138/13, website: www.plgrid.pl. The projects are co-funded by the European Regional Development Fund as part of the Innovative Economy program.

The research leading to these results has also received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 604102 (Human Brain Project).

References

[1] A. Streit, P. Bala, A. Beck-Ratzka, K. Benedyczak, S. Bergmann, R. Breu, J. M. Daivandy, B. Demuth, A. Eifer, A. Giesler, B. Hagemeier, S. Holl, V. Huber, N. Lamla, D. Mallmann, A. S. Memon, M. S. Memon, M. Rambadt, M. Riedel, M. Romberg, B. Schuller, T. Schlauch, A. Schreiber, T. Soddemann, and W. Ziegler, "UNICORE 6 – recent and future advancements," *Annals of telecommunications-annales des télécommunications*, vol. 65, no. 11-12, pp. 757–762, 2010.

[2] "UNICORE Open Source project page," [accessed: 2016-02-01]. [Online]. Available: http://sourceforge.net/projects/unicore/

[3] "Unity project website," February 2016, [accessed: 2016-02-01]. [Online]. Available: http://www.unity-idm.eu

[4] B. Schuller and T. Pohlmann, "UFTP: High-performance data transfer for UNICORE," in *Proceedings of 7th UNICORE Summit 2011*, ser. IAS Series, no. 9. Forschungszentrum Jülich GmbH, 2011, pp. 135–142.

[5] J. Hughes, S. Cantor, J. Hodges, F. Hirsch, P. Mishra, R. Philpott, and E. Maler, "Profiles for the OASIS Security Assertion Markup Language (SAML) V2.0 OASIS Standard," 3 2005, [accessed: 2016-02-01]. [Online]. Available: http://docs.oasis-open.org/security/saml/v2.0/

[6] "Human Brain Project," [accessed: 2016-02-01]. [Online]. Available: http://www.humanbrainproject.eu/

[7] "OpenID Connect," [accessed: 2016-02-01]. [Online]. Available: http://openid.net/connect

[8] "Vaadin framework website," [accessed: 2016-02-01]. [Online]. Available: https://vaadin.com/

[9] R. Johnson, J. Höller, K. Donald, C. Sampaleanu, R. Harrop, T. Risberg, A. Arendsen, D. Davison, D. Kopylenko, M. Pollack *et al.*, "The spring framework–reference documentation," *Interface*, vol. 21, 2004.

[10] "UNICORE commandline client: User Manual," [accessed: 2016-02-01]. [Online]. Available: http://unicore.eu/documentation/manuals/unicore/files/ucc/ucc-manual.html

[11] B. Demuth, B. Schuller, S. Holl, J. Daivandy, A. Giesler, V. Huber, and S. Sild, "The UNICORE Rich Client: Facilitating the automated execution of scientific workflows," in *e-Science (e-Science), 2010 IEEE Sixth International Conference on*, Dec 2010, pp. 238–245.

[12] "The Eclipse Foundation open source community website," [accessed: 2016-02-01]. [Online]. Available: https://eclipse.org/

[13] J. Kocot, T. Szepieniec, D. Harężlak, K. Noga, and M. Sterzel, "Insilicolab – managing complexity of chemistry computations," in *Building a National Distributed e-Infrastructure–PL-Grid*. Springer, 2012, pp. 265–275.

[14] A. Eilmes, M. Sterzel, T. Szepieniec, J. Kocot, K. Noga, and M. Golik, "Comprehensive support for chemistry computations in PL-Grid infrastructure," in *eScience on Distributed Computing Infrastructure*. Springer, 2014, pp. 250–262.

[15] "Adaptive computing: Viewpoint portal," [accessed: 2016-02-01]. [Online]. Available: http://www.adaptivecomputing.com/products/hpc-products/viewpoint/

Problem-Oriented Scheduling of Cloud Applications: PO-HEFT Algorithm Case Study

E.A. Nepovinnykh and G.I. Radchenko
*South Ural State University, Chelyabinsk, Russia
nepovinnykhea@susu.ru, gleb.radchenko@susu.ru

Abstract - Today we see a significantly increased use of problem-oriented approach to the development of cloud computing environment scheduling algorithms. There are already several such algorithms. However, a lot of these require that the tasks within a single job are independent and do not account for the execution of each task and the volume of data transmitted. We propose a model of problem-oriented cloud environment. Using this model, we propose a list-based algorithm of problem-oriented planning of execution of applications in a cloud environment that considers the applications' execution profiles, based on a Heterogeneous Earliest-Finish-Time (HEFT) algorithm.

Keywords – scheduling, execution planning, cloud computing, grid computing, HEFT

I. INTRODUCTION

Today a lot of complex e-Science tasks are solved using computer simulation which usually requires significant computational resources usage [1]. Moreover, the solutions, developed for such tasks are often characterized by structural complexity, which causes different resources (informational, software or hardware) to be integrated within a single solution. The complexity of the solutions grows as the multidisciplinary tasks are considered.

Today's common approach for building composite solutions is based on Service-Oriented Architecture [2] which forms the basis from interconnection of services and hiding their complexity behind their interfaces. Interconnection of the services within complex tasks is usually implemented in a form of workflow structures, which exploits graph-based structures to describe interconnection of used services. On the other hand, today the Cloud Computing concept is developed as a business framework for providing on-demand services supporting computing resources' consolidation, abstraction, access automation and utility within a market environment. The service-oriented architecture in the cloud is best implemented using the microservice approach. The microservice model describes a cloud application as a suite of small independent services, each running in its own container and communicating with other services using lightweight mechanisms. These services are built around separate business capabilities, independently deployable and may be written by different development teams using different programming languages and frameworks [3].

To provide scientists and engineers a transparent access to the computing resources a "Problem Solving Environment" (PSE) concept is commonly used. A PSE is a system that provides all the computational facilities necessary to solve a target class of problems. It uses the language of the target class and users need not have specialized knowledge of the underlying hardware or software [4]. At present, PSE researchers are investigating a variety of fields, e.g., Cloud computing support, education support, CAE usage support, document generation support, and so on.

Today most of the systems that provide a problem-oriented approach to e-science problems on the basis of high performance computing resources use workflows to organize a computational process [5]. Nodes of such workflows represents separate tasks implemented by individual services, and the edges define the data or control flow. In this paper, under the "Problem Solving Environment" term we would understand a set of services, software and middleware focused on the implementation of workflows to solve e-Science problems in a specific problem domain, using resources of cloud computing system [6].

Within a problem domain of PSE, a set of tasks, forming the workflow, is predetermined. Those tasks can be grouped into a finite set of classes. Task class is a set of tasks that have the same semantics and the same set of input parameters and output data. On the one hand, this imposes restrictions on the class of problems that can be solved using the PSE. On the other hand, such restriction allows to use a domain-specific information (such as task execution time on one processor core, scalability limits, and the amount of generated data) during resources allocation and scheduling, increasing the efficiency of use of available computational resources.

So, in order to increase efficiency of distributed problem-oriented computer environments it is feasible to use problem-oriented task scheduling methods that use domain-specific information in order to predict computational attributes of a particular workflow.

The main goal of the research is to develop a scheduling algorithm for a workflow-based problem-solving environment, which would effectively use a domain-specific information (such as task execution time, scalability limits, and the amount of data transfer) for prediction of cloud computing environment resources load.

This paper is organized as follows. In section II we present the concept and the basic idea of scheduling applications in cloud environments. In section III we

The reported study was partially supported by RFBR, research project No. 14-07-00420-a and by Grant of the President of the Russian Federation No. MK-7524.2015.9

describe the cloud-based problem solving environment model. In section IV we describe HEFT and PO-HEFT cloud scheduling algorithms complete with a mathematical task model. In section V we describe the implementation of PO-HEFT algorithm in Workflow Sim's cloud environment simulation package. In section VI we summarize the results of our research and give further research directions.

II. SCHEDULING APPLICATIONS IN CLOUD ENVIRONMENTS

Analysis of the main trends in resource scheduling research in distributed problem-oriented environments shows that the theme of the problem-oriented scheduling and prediction of environment load is an urgent task.

In the cloud computer data centers, Holistic Model for Resource Representation is used in virtualized cloud computing data [7]. This model is designed to represent physical resources, virtual machines, and applications in cloud computing environments. The model can be applied to represent cloud applications, VMs, and physical hosts. Each of these entities is described by multiple resources: computing, memory, storage, and networking. A holistic model increases the precision of cloud environment simulation and enables a number of new simulation scenarios focused on heterogeneity of the hardware resources and virtualization. The model distinguishes between computing, memory, storage, and networking types of resources. However, the model can easily scale to include other types of resources as well, e.g., additional GPGPU units.

New cloud-related techniques for resource virtualization and sharing and the corresponding service level agreements call for new optimization models and solutions. Computational Intelligence proves to be applicable to multiple resource management problems that exist at all layers of Cloud computing. Standard optimization objectives for scheduling are to minimize makespan and cost, but additional objectives may include optimization of energy consumption or communications. Solutions to this multi-objective optimization problem include but are not limited to: Improved Differential Evolutionary Algorithm combined with the Taguchi method, Multi-Objective Evolutionary Algorithm based on NSGA-II, Case Library and Pareto Solution based hybrid GA Particle Swarm Optimization, Auction-Based Biobjective Scheduling Strategy etc. [2]. The main drawback of mentioned algorithms is the fact that they do not use information about previous executions.

The main reason that traditional cluster and grid resource allocation approaches fail to provide efficient performance in clouds is that most of cloud applications require availability of communication resources for information exchange between tasks, with databases or the end users [10]. CA-DAG model for cloud computing applications, which overcomes shortcomings of existing approaches using communication awareness. This model is based on Directed Acyclic Graphs that in addition to computing vertices include separate vertices to represent communications. Such a representation allows making separate resource allocation decisions: assigning processors to handle computing jobs, and network resources for information transmissions. A case study is given and corresponding results indicate that DAG scheduling algorithms designed for single DAG and single machine settings are not well suited for Grid scheduling scenarios, where user run time estimates are available.

For practical purposes quite simple scheduler MaxAR with minimal information requirements can provide good performance for multiple workflow scheduling [11]. In real Grid environments this strategy might have similar performance comparing with the best ones when considering approximation factor, mean critical path waiting time, and critical path slowdown. Besides the performance aspect the use of MaxAR does not require additional management overhead such as DAG analysis, site local queue ordering, and constructing preliminary schedules by the Grid broker. It has small time complexity. This approach is related with offline scheduling which can be used as a starting point for addressing the online case. Online Grid workflow management brings new challenges to above problem, as it requires more flexible load balancing workflows and their tasks over the time.

Nowadays the shifting emphasis of clouds towards a service-oriented paradigm has led to the adoption of Service Level Agreements (SLAs) [12]. The use of SLAs has a strong influence on job scheduling, as schedules must observe quality of service constraints. In terms of minimizing power consumption and maximizing provider income Min-e outperforms other allocation strategies. The strategy is stable even in significantly different conditions. The information about the speed of machines does not help to improve significantly the allocation strategies. When examining the overall system performance on the real data, it is determined that appropriate distribution of energy requirements over the system provide more benefits in income and power consumption than other strategies. Min-e is a simple allocation strategy requiring minimal information and little computational complexity. Nevertheless, it achieves good improvements in both objectives and quality of service guarantees. However, it is not assessed its actual efficiency and effectiveness.

One of the most popular algorithms is scheduled list-based algorithm Min-min [13]. Min-min sets high scheduling priority to tasks which have the shortest execution time. The main drawback of scheduled list-based algorithms is that they do not analyze the whole task graph.

One of the important classes of computational problems is problem-oriented workflow applications executed in distributed computing environment [14]. A problem-oriented workflow application can be represented by a directed graph whose vertices are tasks and arcs are data flows. Problem-oriented scheduling (POS) algorithm is proposed. The POS algorithm takes into account both specifics of the problem-oriented jobs and multi-core structure of the computing system nodes. The POS algorithm is designed for use in distributed computing systems with manycore processors. The algorithm allows one to schedule execution of one task on several processor cores with regard to constraints on scalability of the task.

Cloud computing can satisfy the different service requests with different configuration, deployment condition and service resources of various users at different

time point. With the influence of multidimensional factors, it is unreality to test with different parameters in actual cloud computing center. Typical Tools for Cloud Workflow Scheduling Research are CloudSim and WorkflowSim [9]. CloudSim is a toolkit (library) for simulation of cloud computing scenarios. It provides basic classes for describing data centers, virtual machines, applications, users, computational resources, and policies for management of diverse parts of the system (e.g., scheduling and provisioning).

WorkflowSim extends the CloudSim simulation toolkit by introducing the support of workflow preparation and execution with an implementation of a stack of workflow parser, workflow engine and job scheduler. WorkflowSim is used for validating Graph algorithm, distributed computing, workflow scheduling, resource provisioning and so on. Compared to CloudSim and other workflow simulators, WorkflowSim provides support of task clustering that merges tasks into a cluster job and dynamic scheduling algorithm that jobs matched to a worker node whenever a worker node become idle.

In the following sections, we would present a new problem-oriented resource-scheduling algorithm for distributed computing environments, which uses heuristic score-based approach based on the HEFT algorithm for the task of the problem-oriented scheduling in cloud environments.

III. CLOUD-BASED PROBLEM SOLVING ENVIRONMENT MODEL

Let us define a model for cloud problem solving environment, so we could simulate the task scheduling algorithm. We would work in the conditions, where a set $\mathfrak{M}$ of virtual machines $\mathfrak{m} \in \mathfrak{M}$ was distributed between all available nodes $\mathfrak{n} \in \mathfrak{N}$ of cloud platform.

Let us define a virtual machine image performance factor as:

$$\pi: \mathfrak{m} \to \mathbb{Z}_{>0},$$

where $\mathfrak{m}$ is a virtual machine image.

Numerical characteristics of a virtual machine image, synthetic tests results or existing functions' test execution results can serve as examples of such performance characteristics [15, 16].

In order to maximize the quality of task characteristics prediction on specified machine, we need to take into account several performance characteristics, including such characteristics as the number of available processors and memory; CPU frequency; hard drive data exchange speed; LINPACK test results and so on. Thus, let us define a vector Π of the performance characteristics of virtual machines deployed in the cloud-based PSE:

$$\Pi = [\pi_1, \pi_2 \dots \pi_r].$$

Each machine $\mathfrak{m} \in \mathfrak{M}$ in a cloud-based PSE is comparable to the performance characteristics of the vector Π, which reflects the values of the performance of the machine:

$$\Pi: \mathfrak{M} \to \mathbb{Z}_{>0}^{r}.$$

Let's define a set of tasks, that can be executed in a PSE as a set $\mathcal{F}$ of functions $f \in \mathcal{F}$. Each function $f: \mathcal{C}^{in} \to \mathcal{C}^{out}$ receives n information objects $\mathcal{I}^{in} = (I_1^{in}, \dots, I_n^{in})$ of classes $\mathcal{C}^{in} = (C_1, \dots, C_n)$. The result of the function is m new information objects $\mathcal{I}^{out} = (I_1^{out}, \dots, I_m^{out})$ of classes $\mathcal{C}^{out} = (C_1', \dots, C_m')$. We assume that in our model, each task of a workflow is allocated to one virtual machine. Direct access to components of the computing system is not provided.

One particular feature of a problem-oriented computing environment is the fact that said environment uses information about task classes' features during scheduling and resource provisioning. We require that every task class should have these functions defined for prediction of task execution process depending on input parameters:

1) output data volume estimation function;
2) task execution time estimation function on a machine with a given performance characteristics values vector.

To implement the problem-oriented scheduling, let us define two operators, which should be implemented in the PSE:

1) *The operator of the expected output* $\nu(f, \mathcal{I}^{in})$ – it is the operator that returns the expected total size in bytes of output data objects $\mathcal{I}^{out}$ for the function $f: \mathcal{C}^{in} \to \mathcal{C}^{out}$:

$$\nu(f, \mathcal{I}^{in}) = |\mathcal{I}^{out}|.$$

2) *The operator of the expected function's execution time* $\tau(f, \mathcal{I}^{in}, \Pi)$, that returns the estimated run time (in seconds) of a function $f: \mathcal{C}^{in} \to \mathcal{C}^{out}$ for a given set of input data $\mathcal{I}^{in}$ on a given machine, with the performance characteristics vector Π:

$$\tau: (f, \mathcal{I}^{in}, \Pi) \to \mathbb{N}.$$

Execution time of a function $f: \mathcal{C}^{in} \to \mathcal{C}^{out}$ on a given machine with a performance values vector Π can be defined as an operator that takes input information objects vector $\mathcal{I}^{in}$. Unfortunately it is impossible to estimate a function execution time with absolute accuracy due to the fact that the computations involved in output information objects preparation $\mathcal{I}^{out}$ might indirectly depend on multiple factors that our model does not account for, including, but not limited to, background processes, available cache volume, branch prediction rate, etc. In order to take into account this inherent inaccuracy, execution time estimate can be modelled as a random value that is a sum of two parts:

$$\chi(f, \Pi, \mathcal{I}^{in}) = \tau(f, \Pi, \mathcal{I}^{in}) + \alpha,$$

where $\tau(f, \Pi, \mathcal{I}^{in})$ – a deterministic function that represents a dependency of execution time of the function f that is running on a computer with a performance values vector Π on input information objects vector $\mathcal{I}^{in}$, α – a stochastic value with the expected value ($M[\alpha] = 0$), that represents factors that our model does not account for.

In conditions, when a specific function (task) of PSE is implemented on the basis of a virtual machine with a pre-defined performance characteristics vector Π_0, to evaluate

the expected value $\hat{E}[\chi(f, \Pi_0, \mathcal{I}^{in})]$ we can use the k-nearest neighbor method, based on records of execution time of the previous launches of function f on the same machine with close values of input parameters:

$$\hat{E}[\chi(f, \Pi_0, \mathcal{I}^{in})] = \hat{\tau}(f, \Pi_0, \mathcal{I}^{in}) = \frac{1}{k}\sum_{i=1}^{k} W_i(\mathcal{I}^{in}) t^{f}_{\mathcal{I}_i^{in}, \Pi_0} \quad (1)$$

where $\hat{\tau}(f, \Pi_0, \mathcal{I}^{in})$ is a weighted average execution time estimation for the function f with the $\mathcal{I}^{in}$ input parameters on the basis of k previous observations of the execution time $t^{f}_{\mathcal{I}_i^{in}, \Pi_0}$ of the function f with the input parameters $\mathcal{I}_i^{in}$ on a machine with performance characteristics vector Π_0. The weighting function $W_i(\mathcal{I}^{in})$ assigns greater weight to the function execution time records, where the values of the input parameters are closer to the $\mathcal{I}^{in}$.

To take into account a possibility of execution of the function f on the virtual machine with the characteristics vector that is different from Π_0, we can extend the definition of the evaluation parameter vector, adding to the vector $\mathcal{I}^{in}$ of the input parameters the virtual machine performance vector $\Pi = [\pi_1, \pi_2 \ldots \pi_r]$. Thus, we assume that as a characteristic we use the vector P of dimension $n + r$, where n - number of input parameters of the function f, r - the number of virtual machine performance characteristics, defined as follows:

$$P = [I_1^{in}, \ldots, I_n^{in}, \pi_1, \ldots, \pi_r].$$

In this case, (1) can be transformed into the following form:

$$\hat{E}[\chi(f, \Pi, \mathcal{I}^{in})] = \hat{\tau}(f, \Pi, \mathcal{I}^{in}) = \frac{1}{k}\sum_{i=1}^{k} W_i(P) t^{f}_{P_i},$$

where $t^{f}_{P_i}$ is the value of the previous observation of execution time of the function f with the evaluation parameter vector P_i.

Similarly, we can estimate the amount of output data for the function f:

$$\hat{v}(f, \mathcal{I}^{in}) = \frac{1}{k}\sum_{i=1}^{k} W_i(\mathcal{I}^{in}) v^{f}_{\mathcal{I}_i^{in}},$$

where $\hat{v}(f, \mathcal{I}^{in})$ is an estimate value of the volume of output parameters of the function f on the basis of information on output volumes of k previous runs of the same function with the $\mathcal{I}_i^{in}$ input parameters, using the weighting function W.

IV. HEFT ALGORITHM FOR THE PROBLEM-ORIENTED SCHEDULING

We offer a list-based algorithm for problem-oriented scheduling in cloud environments based on their computing profiles. List-based scheduling involves the definition of computational units' priorities and starting the execution according to the received priority. The binding to high-priority tasks resources takes place first. The proposed approach allows us to take into account the costs of transmission of data between nodes, thereby reducing the total time of execution of the workflow. The proposed algorithm is based on an algorithm of Heterogeneous Earliest-Finish-Time (HEFT), but contains modifications during the node level computation phase, and takes into account the problem of calculating the incoming communication value of its parent task [17].

Let $|T_x| -$ be the size of the problem T_x, and the R be the set of computing resources with an average processing power $|R| = \sum_{i=1}^{n} |R_i| / n$. Then, the average time to complete the task with all available resources is calculated as

$$E(T_x) = \frac{|T_x|}{|R|} \quad (2)$$

Let $\overline{T}_{xy}$ be the amount of data transferred between tasks T_x and T_y, and R be the set of available resources with an average capacity of data transfer $\overline{R} = \sum_{i=1}^{n} \overline{R}_i / n$. Then the average score on data transfer costs between tasks T_x and T_y for all pairs of p.

$$D(T_{xy}) = \frac{\overline{T}_{xy}}{\overline{R}} \quad (3)$$

Thus, the priority calculation unit may be defined as

$$rank(T_x) = E(T_x) + \\ + \max_{T_y \in succ(T_x)} (D(T_{xy}) + rank(T_y)) \quad (4)$$

where $succ(T_x)$ is the set of tasks that depend on the task T_x.

Thus, the task priority is directly determined by the priority of all its dependent tasks. Assign tasks to the resources as follows: a task with a higher priority if all the tasks on which it depends, is appointed to the computing resource, providing less time for the task [18].

Taking into account the specifics of the problem-oriented cloud computing environment following modifications apply to this algorithm.

Let F be the set of all functions that can be implemented in the subject area. Then a separate problem T_x is a function $f_x \in F$ with a set of input data objects $\mathcal{I}^{in} = (I_1^{in}, \ldots, I_n^{in})$:

$$T_x = f_x(\mathcal{I}_x^{in}),$$

We define R as the set of available for the deployment virtual machines with mean production capacity

$$|R| = \sum_{i=1}^{n} \frac{|R_i|}{n} = \sum_{i=1}^{n} \frac{|\Pi_i|}{n}.$$

In this case, for evaluating the execution time we can apply the following formula:

$$E(T_x) = \hat{\tau}(f_x, |R|, \mathcal{I}_x^{in}) \quad (5)$$

where $\hat{\tau}(f_x, |R|, \mathcal{I}_x^{in})$ is an average execution time estimation for the function f_x on a set of machines with mean production capacity $|R|$ with the set of known values of input parameters $\mathcal{I}_x^{in}$.

The model of problem-oriented services should take into account the amount of data returned by each task T_x. This may be used by the operator of the expected output $v(f, \mathcal{I}^{in})$, which returns the expected total size in bytes of output data objects $\mathcal{I}^{out}$. Consequently, within the framework of problem-oriented model for the evaluation of

data transmission time between two tasks the following estimation can be used:

$$D(T_{xy}) = \hat{v}(f_x, J_x^{in}) * \overline{R}_{xy}, \quad (6)$$

where $\overline{R}_{xy}$ is the bandwidth of data transmission channel in the cloud computing system. During the execution of task it can be estimated as one of the following values:

1) $R = 0$, when the data transmission channel consists of a single node;
2) $R = \beta_{group}$, when the data transmission channel is shared by a group of nodes;
3) $R = \beta_{cluster}$,, when the data transmission channel is shared by a cluster of compute nodes.

Figure 1 shows the pseudo-code for algorithm of problem-oriented workflow scheduling in a cloud-computing environment based on computing profiles.

```
PROCEDURE: PO-HEFT
INPUT: TaskGraph G(T, E),
TaskDistributionList, ResourcesSet R
BEGIN
       for each t T from task graph G
           Approximate task execution
           time according to (5)
       for each e E from task graph G
           Approximate data transfer time
           according to (6)
       Start the width-first search in
reverse task order and calculate a rank
for each task according to (4)
       while T has unfinished tasks
       TaskList <- get completed tasks
from task graph G
       Schedule Task (TaskList, R)
       Update TaskDistributionList
END
PROCEDURE: Schedule Task
INPUT: TaskList, ResourcesSet R
BEGIN
      Sort TaskList in reverse task rank
      order
      for each t from TaskList
              r <- get resource from R
              that can complete t earlier
              schedule t on r
              update status of r
END
```

Fig. 1. Problem-oriented heuristic scheduling algorithm PO-HEFT

V. Algorithm Implementation and Performance Evaluation

In order to assess the proposed algorithm's efficiency, we had to develop a benchmark using Workflow Sim cloud environment simulation platform. We have implemented the PO-HEFT algorithm itself, as well as a naive brute force algorithm that finds and ideal scheduling solution.

The algorithm was implemented as a number of Java classes so that Workflow Sim can use it as the simulated cloud environment's scheduler. We have implemented both a custom DatacenterBroker in order to schedule VMs in a data center and a custom CloudletScheduler in order to schedule tasks (cloudlets in CloudSim's and Workflow Sim's terminology) in a single VM.

The algorithm was tested in a simulation in which virtual machines with homogeneous characteristics have been deployed. The simulated system was given the same work flow 60 times, which greatly exceeds the capacity of the system. For the distribution of the workflow we have used: a scheduler that does not use the information about the previous system runs that is built in Workflow Sim itself, the perfect scheduler, which implements the ideal scheduling through complete search space enumeration and a scheduler based on the PO-HEFT algorithm, which uses information about previous runs. The computational complexity of the perfect scheduler does not allow its usage in any non-trivial simulation and, therefore, this algorithm is not present in this comparison. We have also tested several algorithms such as plain HEFT, particle swarm optimization and genetic algorithm and this will be a topic for further research.

We plan to implement the developed cloud system and model DAG, POS and Min-min algorithms behavior in order to assess their efficiency.

VI. Conlusion

In this article, we assessed current scheduling algorithms and defined a model that allows to evaluate various cloud computing environment metrics for problem-oriented scheduling. Next, we described the PO-HEFT scheduling algorithm, which aims to provide workflow scheduling in heterogeneous cloud environments. The main distinctive feature of this algorithm is it's ability to adapt the solution based on previous runs, which allows this algorithm to provide better resource utilization.

The algorithm's efficiency was assessed in the CloudSim with help of Workflow Sim extension cloud environment simulation software. As a benchmark we used CloudSim's built-in scheduler called "space-shared scheduling policy" which uses round-robin for resource provisioning and virtual machines creation. Our proposed algorithm have shown significant efficiency gains over this simple scheduler.

As a further development we will investigate the possibility of deploying this algorithm at a real cluster in order to assess its real-life, non-simulated performance. We will also compare this algorithm against different algorithms that do not use information about previous runs in order to give an empirical prove that this is a viable heuristic in workflow scheduling. We plan to extend the algorithm in order to schedule not only tasks on machines but also to schedule machine provisioning on virtual nodes.

References

[1] S. V. Kovalchuk, P. A. Smirnov, K. V. Knyazkov, A. S. Zagarskikh, and A. V. Boukhanovsky, "Knowledge-based expressive technologies within cloud computing environments," *Adv. Intell. Syst. Comput.*, vol. 279, pp. 1–11, 2014.

[2] D. I. Savchenko, G. I. Radchenko, and O. Taipale, "Microservices validation: Mjolnirr platform case study," *2015 38th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2015 - Proc.*, pp. 235–240, 2015.

[3] J. Thones, "Microservices," *IEEE Softw.*, vol. 32, no. 1, pp. 116–116, 2015.

[4] H. Kobashi, S. Kawata, Y. Manabe, M. Matsumoto, H. Usami, and D. Barada, "PSE park: Framework for problem solving environments," *J. Converg. Inf. Technol.*, vol. 5, no. 4, pp. 225–239, 2010.

[5] E. DEELMAN, D. GANNON, M. SHIELDS, and I. TAYLOR, "Workflows and e-Science : An overview of workflow system features and capabilities," *FGCS. Futur. Gener. Comput. Syst.*, vol. 25, no. 5, pp. 528–540.

[6] A. Shamakina, "Brokering service for supporting problem-oriented grid environments," *UNICORE Summit 2012, Proc.*, vol. 15, pp. 67–75, 2012.

[7] M. Guzek, D. Kliazovich, and P. Bouvry, "A Holistic Model for Resource Representation in Virtualized Cloud Computing Data Centers," *IEEE Int. Conf. Cloud Comput. Technol. Sci.*, pp. 590–598, 2013.

[8] P. Bouvry and B. Service, "Review Article A Survey of Evolutionary Computation for Resource Management of Processing in Cloud Computing," no. may, pp. 53–67, 2015.

[9] C. Chen, J. Liu, Y. Wen, and J. Chen, "CCIS 495 - Research on Workflow Scheduling Algorithms in the Cloud," *Ccis*, vol. 495, pp. 35–48, 2015.

[10] D. Kliazovich, J. E. Pecero, A. Tchernykh, P. Bouvry, S. U. Khan, and A. Y. Zomaya, "CA-DAG: Modeling Communication-Aware Applications for Scheduling in Cloud Computing," *J. Grid Comput.*, 2015.

[11] A. Hirales-Carbajal, A. Tchernykh, R. Yahyapour, J.-L. Gonzalez-Garcia, T. Roblitz, and J. M. Ramirez-Alcaraz, "Multiple workflow scheduling strategies with user run time estimates on a Grid," *J. Grid Comput.*, vol. 10, no. 2, pp. 325–346, 2012.

[12] A. Tchernykh, L. Lozano, U. Schwiegelshohn, P. Bouvry, J. E. Pecero, S. Nesmachnow, and A. Y. Drozdov, "Online Bi-Objective Scheduling for IaaS Clouds Ensuring Quality of Service," *J. Grid Comput.*, 2015.

[13] J. Yu, R. Buyya, and K. Ramamohanarao, "Workflow Scheduling Algorithms for Grid Computing," *Springer Berlin Heidelb.*, vol. 146, pp. 173–214, 2008.

[14] L. B. Sokolinsky and A. V. Shamakina, "Methods of resource management in problem-oriented computing environment," *Program. Comput. Softw.*, vol. 42, no. 1, pp. 17–26, 2016.

[15] J. J. Dongarra, P. Luszczek, and A. Petite, "The LINPACK benchmark: Past, present and future," *Concurr. Comput. Pract. Exp.*, vol. 15, no. 9, pp. 803–820, 2003.

[16] "WPrime Systems. Super PI. 2013." [Online]. Available: http://www.superpi.net/ . [Accessed: 14-Nov-2015].

[17] H. Topcuoglu, S. Hariri, and I. C. Society, "Performance-Effective and Low-Complexity," *Parallel Distrib. Syst. IEEE Trans.*, vol. 13, no. 3, pp. 260–274, 2002.

[18] S. Chen, Y. Wang, and M. Pedram, "Concurrent placement, capacity provisioning, and request flow control for a distributed cloud infrastructure," p. 279, 2014.

Towards a Novel Infrastructure for Conducting High Productive Cloud-Based Scientific Analytics

Peter Brezany*, Thomas Ludescher† and Thomas Feilhauer‡
* *Research Group Scientific Computing, Faculty of Computer Science, University of Vienna, Vienna, Austria and SIX Research Centre, Brno University of Technology, Brno, Czech Republic*
Email: peter.brezany@univie.ac.at
† *Department of Computer Science, University of Applied Sciences, Dornbirn, Austria*
Email: thomas@ludescher.at
‡ *Department of Computer Science, University of Applied Sciences, Dornbirn, Austria*
Email: thomas.feilhauer@fhv.at

***Abstract*—The life-science and health care research environments offer an abundance of new opportunities for improvement of their efficiency and productivity using big data in collaborative research processes. A key component of this development is e-Science analytics, which is typically supported by Cloud computing nowadays. However, the state-of-the-art Cloud technology does not provide an appropriate support for high-productivity e-Science analytics. In this paper, we show how productivity of Cloud-based analytics systems can be increased by (a) supporting researchers with integrating multiple problem solving environments into the life cycle of data analysis, (b) parallel code execution on top of multiple cores or computing machines, (c) enabling safe inclusion of sensitive datasets into analytical processes through improved security mechanisms, (d) introducing scientific dataspace–a novel data management abstraction, and (e) automatic analysis services enabling a faster discovery of scientific insights and providing hints to detect potential new topics of interests. Moreover, an appropriate formal productivity model for evaluating infrastructure design decisions was developed. The result of the realization of this vision, a key contribution of this effort, is called the High-Productivity Framework that was tested and evaluated using real life-science application domain addressing breath gas analysis applied e.g. in the cancer treatment.**

***Keywords*-productivity model, scientific analysis, scientific studies, cloud computing, security, breath gas analysis**

I. Introduction

e-Science refers to the modern large scale science that is increasingly being carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they require access to very large data collections and very large scale computing resources back to the individual user scientists. A key component of this development is *e-Science analytics*, a dynamic research field, that includes rigorous and sophisticated scientific methods of data preprocessing, integration, analysis, and visualization. Unlike traditional business analytics, e-Science analytics has to deal with huge, complex, heterogeneous, and very often geographically distributed datasets, the volume of which is already measured in petabytes. Because of the huge volume and high-dimensionality of data, the associated analytics tasks are typically data and compute intensive; therefore, they are characterized together as a *resource-intensive analytics*. Moreover, a high level of security has to be guaranteed for many analytical applications, e.g., in finance and medical sectors.

So far, the main focus has been on the functionality of Cloud-based analytics systems, and not on the *productivity* aspects associated with their development, use and impacts on expected scientific discoveries. In the high performance computing domain, a special US DARPA program and national programs in several countires have been focused on providing a new generation of economically viable high productivity computing systems for national security and for the industrial and research user communities. We believe that a similar effort is needed in the Cloud research domains, especially, e-Science analytics.

In scientific analysis, when time-consuming calculations are performed, the execution time can be improved by using more and more CPU power and calculating the results in parallel, but the costs for execution will often be increased with the reduced time consumption. Evaluating the productivity helps companies and research institutes to achieve the best possible output with minimal costs and time. Our research presented in this paper addresses exactly this issue. Its main contribution is a formal model enabling to evaluate and predict key productivity paramenters associated with the development and execution of modern scientific analysis software systems. Furthermore, we address several other potential sources enabling to increase the productivity and provide appropriate original solutions. For better understanding, we provide several example scenarios. The productivity model and the different approaches are combined in the *High-Productivity Framework (HiProF)* that we designed and implemented. Its prototype software, documentation, users guide, and a test infrastructure are available on the Internet.

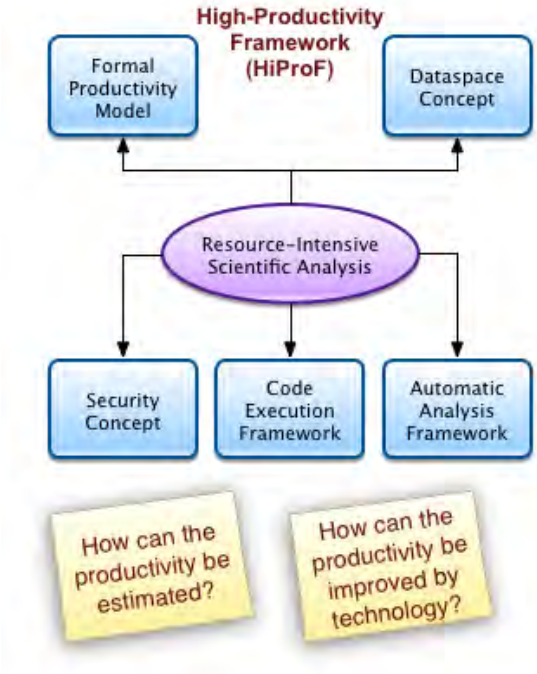


Figure 1. Research Questions and the HiProF Response.

During our research, while developing a high productivity infrastructure for breath gass analysis called the ABA-Cloud[1] [1], we came across two major research questions that are shown in Figure 1 in the HiProF context. The first question "How can the productivity be estimated?" is very important to evaluate the effect of decisions (e.g., increase number of worker nodes to achieve the results in less time). Without a generally usable model, these changes cannot be measured or used to make the best decision in terms of increasing the overall productivity. The second question "How can the productivity be improved by technology?" deals with several different ways to increase the overall productivity in scientific analysis. The *Security Concept*, *Code Execution Framework*, *Dataspace Concept*, and *Automatic Analysis Framework* are four of the most promising contributions, which are used in different sample scenarios introduced in the subsequent paragraphs. The *Security Concept* compares the productivity influence with and without using suitable security mechanisms. The *Code Execution Framework (CEF)* enables researchers to execute different Problem Solving Environment (PSE) codes in parallel in a cloud-based infrastructure, even if the corresponding PSE is not installed locally. We proposed and realized the *concept of scientific dataspace* that involves and manages numerous and often complex types of primary, derived and background data in an intelligent way that is based on the Semantic Web principles. The *Automatic Analysis Framework* (AAF) can be used to pre-analyze newly collected data to detect dependencies or data quality issues. As a next step, the AAF can be used to find new insights into existing research areas or even find new topics of interest. All these components (Formal Productivity Model, Security Concept, Dataspace Concept, Code Execution Framework, Automatic Analysis Framework) are part of HiProF.

Developers of high-productivity infrastructures can select the HiProF technologies and integrate them in their own infrastructures. All provided technologies can be used optionally. If private data (e.g., patient data) are involved and user interactions in the infrastructure must be traced down to the lowest level of system (e.g., database), the provided cloud-enabled security concept simplifies the development of the infrastructure to fulfill such a requirement. Installing and using the CEF in a high-productivity infrastructure enables researchers to execute problem solving environment codes (MATLAB, R, and Octave) in a cloud-based infrastructure. The CEF focuses on saving time and cost to increase the productivity. The AAF supports researchers during their daily work if continuous data is collected. Existing algorithms should be performed at these new data. Furthermore, the developer or administrator of the infrastructure can use the provided formal productivity model to make infrastructure decisions (e.g., size of the private cloud infrastructure). Depending on the requirements of the new infrastructure, the model can be simplified by reducing unnecessary cost variables (e.g., remove the license costs if only open source software is used or the required licenses are already available and do not generate additional costs).

The rest of the paper is organized as follows. Section *Background and Related Work* contains a well known definition of productivity and the related work. Section *Formal Productivity Model* presents a general formula to calculate the productivity. The expected productivity improvements of the most promising approaches are discussed theoretically afterwards in Section *Productivity Model Usage Scenarious*. Finally, we briefly summarize the results achieved in Section *Conclusions*.

II. Background and Related Work

This section provides a general definition of the term productivity and the related work addressing productivity in the field of computer science.

A very common and general definition of productivity is listed at [2]: *"Productivity is a measure relating a quantity or quality of output to the inputs required to produce it."*

Productivity plays an important role in almost every discipline. It can be used to generate a better result, to reduce the accruing costs, to evaluate changes of infrastructure used,

[1]It primarily supports the development and reproducibility of scientific studies in the context of Problem Solving Environments, such as MATLAB, R, and Octave and Cloud-enabled workflows.

to evaluate the impact of the quality of produced final results (e.g., scientific discoveries) on the appropriate domains.

Productivity has been originally informally defined in economics[2] [3] as the amount of output per unit of input, or, in other words, productivity is a ratio of production output to what is required to produce it (input). E.g.: In a computer factory the productivity can be calculated by the number of computers divided by the working hours needed, or inversely the number of working hours to assemble a computer. The productivity can be increased by rising the output or decreasing the working hours needed. This can exemplarily be done by (a) increasing the skills or motivation of the employees, (b) providing better working conditions (e.g., equipment), or (c) providing an automated plant.

In the year 2004, a series of ten papers about the topic High Performance Computing (HPC) productivity was published in the International Journal of High Performance Computing Applications [4]. The different authors share their point of views on this topic. H. Zima arranged the workshop "High-Productivity Programming Languages and Models" [5] in 2004. The goal of this workshop was to design a new high-productivity language system and to find a consensus on a set of common research strategies.

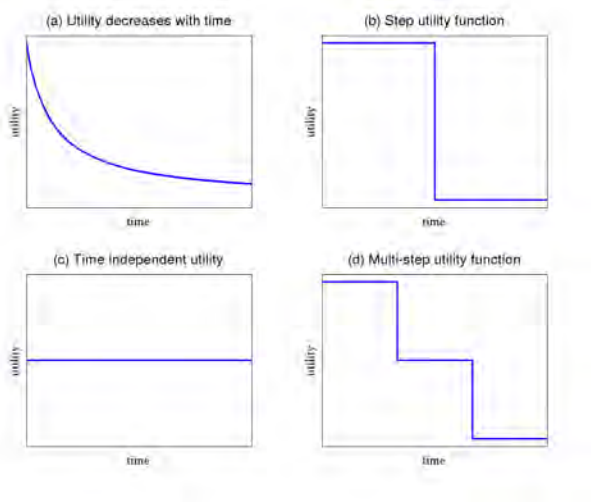


Figure 2. Different utility functions depending on time; modified from [6] This figure shows different time-dependent utility functions. The users of the productivity model can choose the best fitting function for their need.

III. Formal Productivity Model

J. Kepner [6] provides the following productivity formula:

$$\text{Productivity} = \Psi = \frac{\text{Utility}}{\text{Costs}} = \frac{\text{U(T)}}{\text{C}} = \frac{\text{U(T)}}{\text{c(T)}} \qquad (1)$$

[2]The terms productivity and economic purpose partly the same goals. The term economic mainly tries to reduce the costs. The term productivity additionally includes the required time and the result/output.

$U(T)$ represents the time dependent utility function, which is described in Section *Utility*. C stands for the total costs. The total costs can equally be described by a function with the required time ($c(T)$). The costs are further described in Section *Costs*. T is the vector of all relevant time values (e.g., maintenance time, execution time, etc.). This formula is used in this paper to illustrate its support for decision making and improving the overall productivity.

Utility

The Utility is the value a specific user or organization defines on getting a certain answer in a certain time [6]. In general, the utility function is time dependent. A computer manufacturer has a higher benefit if the production process of a computer takes a couple of minutes instead of several hours. Furthermore, the benefit increases if the quality is improved or a more powerful computer is produced.

Figure 2 shows different time dependent utility methods. Its part Figure 2(a) shows a continuously decreasing utility function. This function can be used for many different application areas. Let's just imagine that you are waiting for a bus. If the bus arrives exactly at the same time as you arrive at the bus station you have the highest benefit — if you must wait for a while, the benefit will continuously decrease. Figure 2(b) shows a step utility method. This kind of curve can, for example, be used for a weather forecast simulation. If the simulation takes longer than the time for which the weather should be predicted, your output is useless, because you already have seen the real weather. Figure 2(c) shows a constant function that can exemplarily be used in a non-time critical scientific analysis. Figure 2(d) shows a multi step curve, for example a 3D movie production needs to render their scenes after all changes have been made. The highest benefit is, when the user can see the result right away or after a short coffee break. If he/she needs to wait over night, the utility decreases, but at least the work can be continued at the next working day. It is useless if the user must wait for example one week to see the output of his/her changes and in this case the utility is zero. A step function can be the right choice for a researcher who works with scientific analysis. If the analysis process needs only a couple of seconds, the researcher has the highest benefit. If the computation needs some hours, the utility decreases, but at least the researcher has access to the results on the next working day. Nevertheless, the benefit is probably zero if the complete analysis needs much longer (e.g., years).

Depending on the given problem, it is very important to choose the right utility function.

Costs

The total costs can be divided into total software costs C_{S}, ownership costs C_{O}, hardware/machine costs C_{M}, personnel costs C_{P}, and data costs C_{D}. We can express the same costs dependent on the time $(c_{\text{S}}+c_{\text{O}}+c_{\text{M}}+c_{\text{P}}+c_{\text{D}})\times T$, where c_{x}

are the costs per unit of time and T is a vector of all relevant time values. The software costs can further be divided into license costs C_L for licensed software and full development costs C_D if it is a self developed software ($C_\mathrm{S} = C_\mathrm{L} + C_\mathrm{DEV}$). The ownership costs contain the energy costs C_E, costs for buildings C_B, and maintenance costs C_M ($C_\mathrm{O} = C_\mathrm{E}+C_\mathrm{B}+C_\mathrm{MA}$). Dealing with data usually implies additional costs. These costs accumulate during (a) data gathering C_GA (e.g., collecting sensor data), (b) data generation C_GE (e.g., generating a random time series), data storage C_ST (e.g., store and backup data), data transfer C_T (e.g., transmit data to the Amazon S3), and data transformation or conversion C_C (e.g., convert a HDF5 file [7] to a CSV file), that result in $C_\mathrm{D} = C_\mathrm{GA} + C_\mathrm{GE} + C_\mathrm{ST} + C_\mathrm{T} + C_\mathrm{C}$.

$$C_\mathrm{total} = C_\mathrm{S} + C_\mathrm{O} + C_\mathrm{M} + C_\mathrm{P} + C_\mathrm{D} \quad (2)$$

$$c_\mathrm{total} = c_\mathrm{S} + c_\mathrm{O} + c_\mathrm{M} + c_\mathrm{P} + c_\mathrm{D} \quad (3)$$

$$C_\mathrm{total} = c_\mathrm{total} \times T \quad (4)$$

The formula for the total productivity:

$$\Psi = \frac{U(T)}{C_\mathrm{S} + C_\mathrm{O} + C_\mathrm{M} + C_\mathrm{P} + C_\mathrm{D}} \quad (5)$$

$$\Psi = \frac{U(T)}{(c_\mathrm{s} + c_\mathrm{O} + c_\mathrm{M} + c_\mathrm{P} + c_\mathrm{D}) \times T} \quad (6)$$

The last two overall formulas can be used to calculate the productivity in computer science. These costs contain all accumulated costs (e.g., development, usage, maintenance, license, data). Depending on the given scenario several particular costs are zero or can be neglected. This model can be used to calculate the productivity in resource-intensive scientific analysis (data-intensive and/or compute intensive).

IV. Productivity Model Usage Scenarios

The following sections describe how productivity can be measured and increased by the means of several example scenarios.

Using a Security Concept

The utility method does not only depend on the time, as shown in Section *Utility*. The quality of the output depends on the quality, volume, and availability of the input data as well. The poor quality results in a lower valuable output of the data analysis (utility function) and the additional costs increase the total costs [8]. Both affect the overall productivity negatively.

In many e-Science studies personal related data is involved. The privacy of this data must be assured and all data has to be stored in a secure place during its whole lifetime. Without security, either this sensitive data must be anonymized or the researcher is not allowed to use this private data for the analysis. The original sensitive data is not available for the analysis procedure, which possibly affects the volume of the available input data. The anonymization process can be almost impossible (e.g., if you need the date of birth and the patient's address for further analysis) or it can be time-intensive which increases the total costs. If a researcher is not allowed to access all relevant data, the availability of the input data decreases which determines the quality of the output as well. Both disadvantages decrease the total productivity.

The data quality, data volume, or data availability related utility function can exemplarily be a continuously increasing function or normal step function (Figure 3). The step function can be selected, for example, in the following scenario. If security is available, all data can be used and the utility value is maximized, otherwise only a subset of the data is allowed to be used and the utility value is reduced.

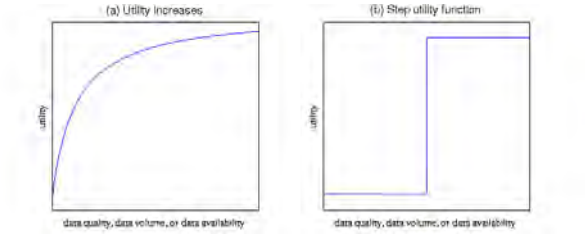


Figure 3. Different utility functions, depending on data quality, data volume, or data availability This figure shows different data-dependent utility functions. These functions depend on data quality, data volume, or data availability.

It is hard to determine a precise rate for the utility in practice. The utility method of e-Science studies can be defined depending on the availability (with security mechanism) of the input data and the time ($U(Time, Data) = U(t, D)$). To simplify the diagram, a continuously decreasing utility function for the time dependencies (see Figure 2(a)) and a step function for the data quality (see Figure 3(b)) is used. Figure 4 shows the total utility function of the supposed scenario. The highest output can be reached if the result is available as soon as possible and if the complete data is available.

Nevertheless the security overhead increases the total costs as well ($\Psi = \frac{U\uparrow}{C\uparrow}$[3]). The security overhead (e.g., Kerberos overhead) mainly depends on the infrastructure used (e.g., network connection, execution time). The following example shows how the productivity can be compared with and without security in general. To calculate the productivity threshold of a system with and without security, we defined formula 7. For simplicity, we assume that the security overhead influences all costs equally (e.g., personnel costs, energy costs). $U(t, D)$ is the utility function without security features, $U_\mathrm{S}(t, D)$ describes the function with security. t_S represents the time additionally required for the security (overhead).

[3]The up arrows indicate that the expected value increases. On the other hand, a down arrow would indicate that the expected value decreases.

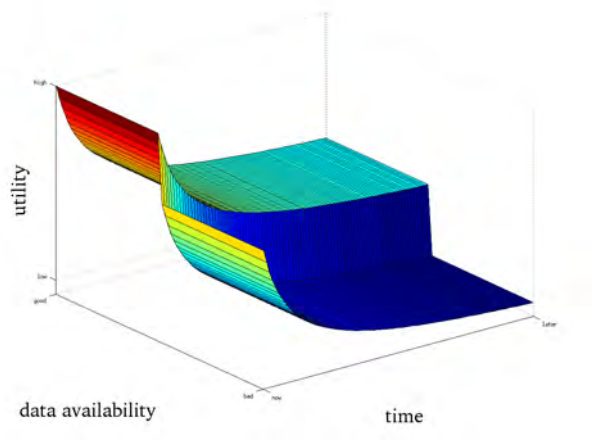


Figure 4. Total utility function In this figure the utility function depends on time (continuously decreasing) and on the data (step function).

$$\frac{U(t,D)}{c_{\text{total}} \cdot t} \quad < \quad \frac{U_{\text{S}}(t+t_{\text{S}},D)}{c_{\text{total}} \cdot (t+t_{\text{S}})} \tag{7}$$

Formula 7 can be rewritten as:

$$t_{\text{S}} \quad < \quad t\left[\frac{U_{\text{S}}(t+t_{\text{S}},D)}{U(t,D)} - 1\right] \tag{8}$$

If we assume that (a) the utility value with security is larger than the utility function without security $U_{\text{S}}(t+t_{\text{S}}) > U(t)$ and (b) the execution time is much higher than the overhead of the security $t >> t_{\text{S}}$, the productivity of the secure system is always better than the productivity of the insecure system. In e-Science studies, where sensitive data (e.g., personal related data) is involved, the productivity increases when security is provided and more available data can be used for analysis. Without security concept, the data must be copied by hand which is time-intensive and error-prone. The extra time increases the personnel costs and invalid data decreases the utility function. Both disadvantages affect the productivity in a negative way. More details of a security concept we have developed are presented in [9].

Dataspace Concept

Our dataspace system [10] involves and manages the primary data captured from the application, data derived by curation and analytics processes, background data including ontology and workflow specifications, semantic relationships between dataspace items based on scientific research ontologies, and available published data. Additionally, it provides advanced querying mechanisms about the contents and relationships in the dataspace and enables productive reproducibility of scientific studies. Our formal productivity model can be applied to this HiProF component in a similar way as we showed it in the previous subsection.

Using a cloud-based data analysis infrastructure

A cloud-based data analysis infrastructure enables researchers to execute different problem solving environment (PSE) codes (e.g., MATLAB, R, and Octave code) in parallel in the cloud. The Code Execution Framework (CEF) [13] component of the HiProF is a kernel system, which is used in such an infrastructure. The same PSE method can be executed with different parameter sets in parallel (embarrassingly parallel problems). The CEF can be installed at a local Eucalyptus infrastructure, Amazon EC2, or every other Amazon compatible cloud-based infrastructure. Different infrastructures can be combined to use a hybrid cloud (e.g., local Eucalyptus for general usage, for high workloads additional EC2 worker nodes are used). Furthermore, it is possible to combine different PSE types within one single analysis (e.g., MATLAB code can be transmitted for execution at the CEF with the R-client and vice-versa). In [12] various scenarious are analyzed in details using our formal productivity model. It is proven there that using the CEF results in a faster execution and the productivity increases.

Automatic Analysis System

An automatic analysis system helps researchers to automatically receive information about existing data. For example, the Automatic Analysis Framework (AAF) [11] uses classification, prediction, and clustering algorithms to automatically analyze research data. The AAF is built upon the workflow management system Taverna, which is widely used in many different domains and uses the data management system, based on the Dataspace Concept [10]. It helps researchers to (a) find new topics of interest or (b) gain new insights in their existing research area. Figure 5 shows a general workflow from this automatic analysis. The AAF can be continuously or in regular intervals executed by the CEF (e.g., once per week).

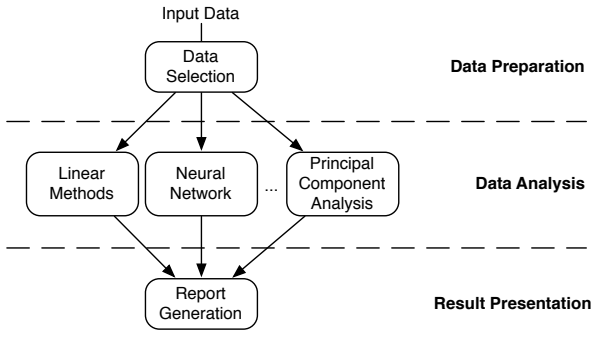


Figure 5. General AAF Workflow This data mining workflow is divided into (a) *Data Preparation*, (b) *Data Analysis*, and (c) *Result Presentation* section.

For calculating the productivity, it is important to know that (a) all automatic analysis calculations have lowest priority, meaning that they will only be executed at the CEF

if no other calculation is waiting, and (b) the output will be evaluated and ranked automatically, depending on the percentage of the highest number of correctly classified test samples.

The productivity of the AAF can be calculated as follows:

$$\Psi = \frac{U(T)\uparrow}{C_{\mathrm{S}} + C_{\mathrm{E}}\uparrow + C_{\mathrm{B}} + C_{\mathrm{MA}}\uparrow + C_{\mathrm{M}} + C_{\mathrm{P}}\uparrow} \quad (9)$$

The AAF uses the CEF only when no other calculation is waiting. Otherwise it waits until the CEF is running idle. Therefore the software costs (C_{S}), building costs (C_{B}), and machine costs (C_{M}) are zero. The energy costs C_{E} and the maintenance costs of the AAF increase, because of the additional usage and further AAF configurations. After the analysis, a researcher must evaluate the result which means that the personnel costs (C_{P}) are rising. The system automatically ranks the result which allows the researcher to skip non-promising results.

$$\Psi = \frac{U(T)\uparrow}{C_{\mathrm{E}}\uparrow + C_{\mathrm{MA}}\uparrow + C_{\mathrm{P}}\uparrow} \quad (10)$$

To increase the productivity, only the utility value must be larger than zero. The project leader must predict the utility function. If the project leader expects a valuable output of the system, the productivity will be increased when using the AAF in comparison to using no automatic analysis.

$$U(T) > 0 \implies \Psi > 0 \quad (11)$$

V. Conclusions

This paper introduces a novel productivity model that can be used in resource-intensive scientific analysis software development processes. Productivity plays an important role to achieve the best results in terms of time, costs, and quality. Our model is an extended productivity formula from J. Kepner with additional cost variables (personnel costs, license costs, energy costs, building costs, maintenance costs, and data costs). These adaptions are necessary in order to be able to calculate the productivity of time-intensive and data-intensive scientific analysis. We defined different approaches to increase the overall productivity for these types of application. Furthermore, we explained the usage of the productivity model with several example scenarios. The scenarios are mathematically described and the threshold formulas are provided. With these formulas and examples, an administrator can evaluate system adaptions (e.g., change the infrastructure by adding additional worker nodes, using faster computers, using a hybrid Code Execution Framework, etc.) and determines how the changes affect the productivity.

The presented framework was tested and evaluated using real life-science application domain addressing breath gas analysis applied e.g. in the cancer treatment.

Acknowledgment

The research leading to these results has received funding from the Austrian Science Fund (Project No. TRP 77-N13) and the Czech National Sustainability Program supported by grant LO1401.

References

[1] I. Elsayed *et al.*, "ABA-Cloud: support for collaborative breath research," *Journal of Breath Research*, vol. 7, no. 2, pp. 026 007–026 007, 2013. [Online]. Available: http://eprints.cs.univie.ac.at/3981/

[2] About.com Economics, "Productivity - Dictionary Definition of Productivity," http://economics.about.com/od/economicsglossary/g/productivity.htm, Accessed July 2013.

[3] OECD, "Measuring Productivity - OECD Manual," http://www.oecd-ilibrary.org/content/book/9789264194519-en, 2001.

[4] Internation Journal of High Performance Computing Applications, "Table of contents, winter 2004, 18 (4)," http://hpc.sagepub.com/content/18/4, 2004.

[5] H. Zima, "Workshop on high-productivity programming languages and models," http://www.cs.illinois.edu/homes/wgropp/bib/reports/hpl_report.pdf, 2004.

[6] J. Kepner, "High Performance Computing Productivity Model Synthesis," *International Journal of High Performance Computing Applications*, vol. 18, no. 4, pp. 505–516, 2004. [Online]. Available: http://hpc.sagepub.com/content/18/4/505.abstract

[7] The HDF Group, "ADF Group - HDF5," http://www.hdfgroup.org/HDF5, Accessed Dec 2012.

[8] D. M. Strong *et al.*, "Data quality in context," *Commun. ACM*, vol. 40, no. 5, pp. 103–110, May 1997.

[9] T. Ludescher, T. Feilhauer, and P. Brezany, "Security Concept and Implementation for a Cloud Based E-science Infrastructure," *2012 Seventh International Conference on Availability, Reliability and Security*, vol. 0, pp. 280–285, 2012.

[10] I. Elsayed and P. Brezany, "Dataspace support platform for e-science," *Computer Science*, vol. 13, no. 1, 2012. [Online]. Available: http://journals.agh.edu.pl/csci/article/view/14

[11] T. Ludescher *et al.*, "Towards a high productivity automatic analysis framework for classification: An initial study," in *Advances in Data Mining. Applications and Theoretical Aspects*, P. Perner, Ed. LNCS, Springer, 2013, vol. 7987, pp. 25–39.

[12] T. Ludescher, "Towards High-Productivity Infrastructures for Time-Intensive Scientific Analysis," Ph.D. Thesis, Faculty of Computer Science, University of Vienna, Austria, 2013.

[13] T. Ludescher, T. Feilhauer, and P. Brezany, "Cloud-Based Code Execution Framework for scientific problem solving environments," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 2, no. 1, p. 11, 2013. [Online]. Available: http://www.journalofcloudcomputing.com/content/2/1/11

An OpenMP runtime profiler/configuration tool for dynamic optimization of the number of threads

Tamara Dancheva, Marjan Gusev, Vladimir Zdravevski, Sashko Ristov
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering
1000 Skopje, Macedonia
Email: tamaradanceva19933@gmail.com, {marjan.gushev, vladimir.zdraveski, sashko.ristov}@finki.ukim.mk

***Abstract*—This paper describes the implementation and the experimental results of a tool for dynamic configuration of the number of threads used in the OpenMP environment based on the current state of the runtime at the time of the call. For this purpose, we use a mix of profiling and machine learning techniques to determine the number of threads. The decision to set the number of threads is made at the time of the call.**

The proposed approach is designed to be cost effective in the scenario of a highly dynamic runtime primarily when running relatively long tasks consisting of a number of parallel constructs.

I. Introduction

The OpenMP API was designed with a goal to provide a simple, fast, compact and portable way of parallelizing programs [1]. Computer architectures have diversified and evolved a great deal since the first release requiring OpenMP to evolve as well. Today, considering the prevalence of SMPs and SMT on the market, the massive virtualization and ccNUMA architectures, the default schemes that have sufficed in the past are no longer producing the desired results with the same consistency as before.

Multi-cache architectures, hyper threading and multi-core techniques helped prevent the potential stagnation in performance due to the limits imposed by physics with a cost. The intensive exploitation of virtualization techniques especially lately (of memory, computing power, application, operating systems) only enhance the complexity introduced.

In order to get a good performance, programmers as well as scientists have had to start studying the hardware characteristics, cache associativity, swapping, virtualization and many other low-level hardware and software characteristics. The main problem is the diversity of resource sharing mechanisms in different architectures, most notably the SMT [2]. OpenMP has a default scheme that is implementation specific, meaning that OpenMP implementations use different default policies to assign and interpret the internal control variables (ICVs). Additionally, in virtual machines, the number of threads is affected by the way different hypervisor present and assign processors and virtual cores to the virtual machine. The default scheme is static and tends to use the number of physical cores as the default number of threads. This does not exploit the hyper threading automatically and whether it is a good or bad decision depends on the interconnection, location, resource sharing policies and mechanisms of the processors involved as well as how the program utilizes them.

Having a tool that can automatically learn the formula for the optimal number of threads without previously gathering explicit user feedback on the program to be ran can significantly automate the process of producing better performing algorithms.

Initially, a suite of benchmarks is ran concurrently using pThreads. Using profiling to capture the state of the runtime while running the suite, an output dataset is created. This dataset is used as an input for the creation of random trees [3] (the forest tree algorithm [4]). This is a machine learning algorithm that uses a supervised approach to train a number of decision trees that all vote on the number of threads i.e classifying the current runtime state, designed in such a way as to avoid overfitting (generate bad predictions because the tree is too specific to a subset of all the possible inputs or runtime states).

In this paper we aim at testing the validity of the following hypotheses:

- H1 There exists a subset or multiple subsets of dataset attributes for which the predictions generated by the random forest/tool does not degrade performance in any of the test cases relative to the performance achieved by assigning the number of cores to the num_threads OpenMP variable.
- H2 There exists a subset or multiple subsets of dataset attributes for which the tool boosts performance in comparison with the default OpenMP scheme either by increasing or decreasing the number of threads.

Further details are disclosed in the following sections. Implementation details of the proposed approach are given in Section II. The experiments are described in Section III and results in Section IV. Relevant discussion is given in Section V, related work in Section VI and conclusions in Section VII.

II. Implementation Details

This section describes the approach used to optimize the number of threads for an OpenMP program.

Retrieval of the runtime parameters of an OpenMP program is done using the SIGAR API [5]. The SIGAR API is an open source cross-platform library that offers helps in parsing and locating major system parameters including CPU usage, ram usage, virtual memory IO traffic (total and by process) as well as more detailed information about the processes, the hardware

TABLE I
LIST OF ATTRIBUTES USED FOR PROFILING

num_threads	number of threads for one run
cpu_usage	cpu usage for all processes
pid_cpu_usage	cpu usage for the current process
ram_usage	ram usage for all processes
pid_ram_usage	ram usage for the current process
pid_page_faults	page faults for the current process
vm_usage	virtual memory usage for all processes
pid_vm_usage	virtual memory usage for the current process
pid_vm_read_usage	virtual memory reads percentage for the current process
pid_vm_write_usage	virtual memory writes percentage for the current process
processes	# of processes and threads created
procs_running	# of running processes and threads
procs_blocked	# of blocked processes and threads

features and the operating system running on the machine. An indicator for its quality is that it is a framework suitable for use in cloud environments for measuring hypervisor performance [6]. It offers an extensive set of utilities to monitor the state of a highly dynamic runtime which makes it a suitable tool for the kind of measurements needed for this tool. A sample of all potential dataset attributes is given in Table I.

PThreads is a an implementation of thread-level parallelism [7]. It is used to concurrently run a suite of OpenMP benchmarks. The benchmarks are ran for a range of threads [1, max_num_threads] constituting one run. This is done 50 times per benchmark by default. The optimal number of threads from the specified range is found using the best performance time yielded during each consecutive run. The corresponding runtime state is then written to disc.

The OpenMP algorithm which utilizes the tool has to call the omp_set_num_threads function before each parallel block to invoke the wrapper function. The number of threads is predicted using the random forest and used as a parameter in the new handle which is returned.

III. DESCRIPTION OF EXPERIMENTS

The goal of running the experiments is to simulate as many states of the runtime as possible so that a representative decision tree can be constructed out of the provided dataset. This is closely related to the parameters that are used to represent the state and is essential to be carefully configured so that performance boost can be obtained. Random delays are enforced in between runs in order to accomplish this better.

The dataset is boosted by generating additional dataset records using the existing dataset records with repetition. Basically the experiment is repeated a number of times, and this helps to break the independency, or decrease its negative impact on the decision trees that are to be generated out of the dataset by adding a variance to the output that helps remove some of the initial bias. This works for small biases only. Additionally, boosting enables us to get p-values and provides a good basis for running statistical tests and deriving conclusions about our data. Therefore, the parameters have to be very carefully selected. Additionally, random forest does not require any normalization of the dataset, since it considers each attribute independently [4].

Each forest is created the first time when the call to set_num_threads is made.

After the initial configuration, the forest is used to predict the number of threads for the parallel region in the OpenMP algorithm. The overall results are 50-150 obtained values of elapsed times in both cases by using the random forest for prediction, and by utilizing OpenMP default schemes.

TABLE II
EXPERIMENTAL ENVIRONMENT

Benchmark programs	Dijkstra, Multitask, Poisson [8]
Hardware	HP Probook 650 i5-4210M RAM 4GB
OS	Debian, GNU Linux 6.0

TABLE III
LIST OF ATTRIBUTES INCLUDED IN THE DATASETS

Dataset 1	all attributes listed in Table I
Dataset 2	num_threads, cpu_usage ,pid_cpu_usage, lst_cpu_usage, ram_usage, pid_page_faults, procs_blocked, user, sys, idle, irq
Dataset 3	num_threads,pid_cpu_usage,lst_cpu_usage, nice, idle, wait, irq, soft_irq
Dataset 4	num_threads, pid_ram_usage, ram_usage, pid_cpu_usage, cpu_usage, procs_blocked nice, idle, user, sys, wait, irq, soft_irq, lst_cpu_usage

The experimental environment is presented in Table II.

The experiments are based on using 4 subsets of the parameters listed in Table I. The corresponding parameters are listed in Table III.

The overall performance is analyzed in two environments. *No stress environment* is the one where the experiments are executed with no additional workload. A *stress environment* is simulated using the stress program package for POSIX systems that imposes stress on the system (CPU, memory, I/O, disk stress) by starting various number of threads concurrently to the experiments.

Three experiments are conducted as OpenMP default schemes and four more experiments using the tool with different dataset. To evaluate the speedup of using the tool and the default configuration we compare the corresponding elapsed times.

Algorithm I conducts exhaustive search to find a combination of input gates to produce logical circuit with output 1 [8]. It has a repetitive structure without complex data dependencies and exploits a can high degree of parallelism. *Algorithm II* is an implementation of the Fast Fourier Transformation (FFT) algorithm with a lot of data-flow dependencies which do not allow a high level of parallelism.

IV. RESULTS

A. Algorithm I in a No Stress environment

Fig. 1 presents the results (elapsed times) obtained with the default OpenMP scheme setting the number of threads variable to 1, 2 and 4. The X-axis represents the number of conducted

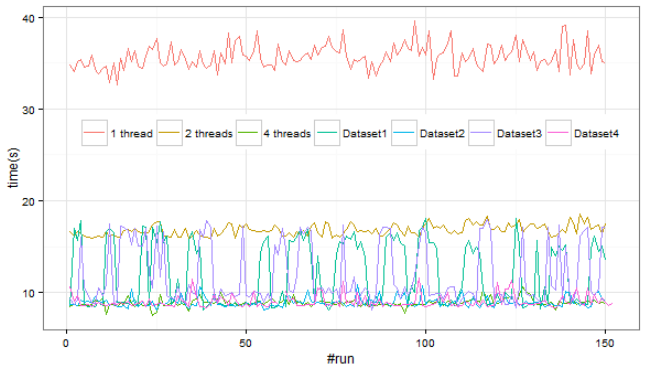


Fig. 1. Performance of executing the Algorithm I in a no stress environment

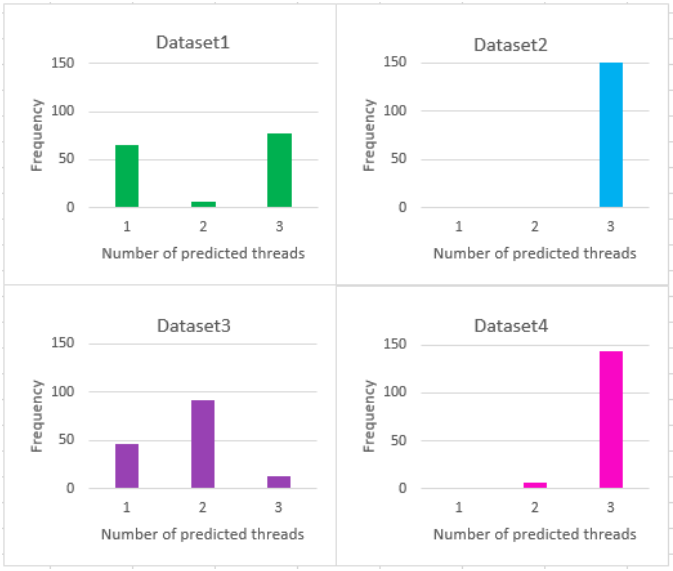


Fig. 2. Histograms for the 4 datasets for executing Algorithm I in a no stress environment.

test runs. Four separate measurements are conducted to evaluate the tool performance results. Each measurement uses a specific dataset defined in Table III to construct the forest. It is used to predict the number of threads, labeled in the graph with Dataset 1-4.

A more detailed information about the output number of predicted threads with the tool is presented in Fig. 2. A total of 150 test runs of Algorithm I is utilizing the same 4 forests used for the performance measurements in Fig. 1 labeled with Dataset 1-4 in the decision making.

The default OpenMP scheme yields best results with 4 threads (value of the OpenMP num_threads variable). Using this reference value for the optimal choice of the number of threads, Dataset 2 and Dataset 4 result in a forest that most accurately predicts the optimal number of threads. This is presented in Fig. 2 pointing out the 100% accuracy for Dataset 2 and 96% accuracy for Dataset 4.

For machines with the same hardware configuration that by default use the number of physical cores (2 in this case it is

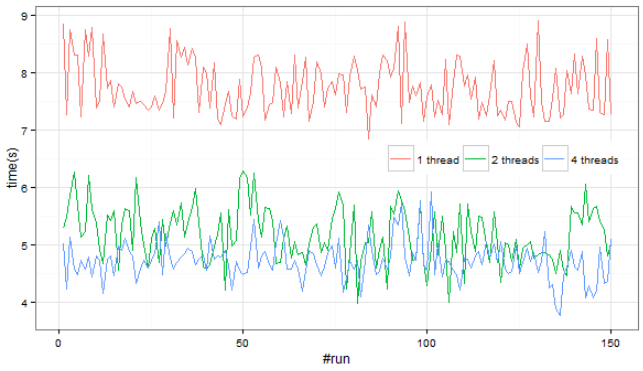


Fig. 3. Performance of executing the Algorithm II in a no stress environment with default OpenMP schemes

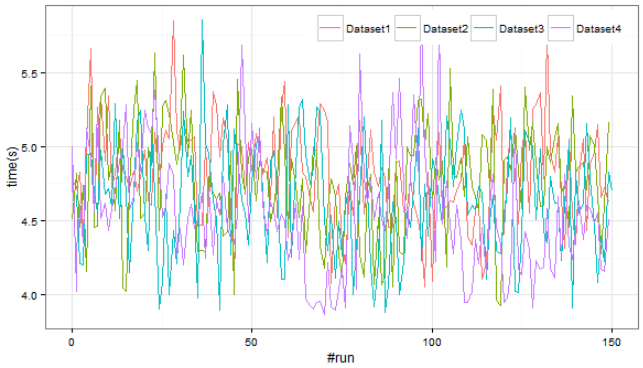


Fig. 4. Performance of executing the Algorithm II in a no stress environment using the tool with Datasets 1 - 4

2 Threads), the best speedup is $Sp(2T, D2)$ achieved using Dataset 2, with an average value of 1.876241, which confirms the Hypothesis H2.

Additionally, the predictions do not include a number of threads less than the physical number of cores nor for the best performing Dataset 2 nor for any of the other datasets, confirming the Hypothesis H1.

B. Algorithm II in a No Stress environment

The results obtained executing the Algorithm II using the OpenMP default scheme are presented in Fig. 3 and for executing the tool with Datasets 1 - 4 in Fig. 4

The histogram of predicted threads for executing the Algorithm II in a no stress environment is presented in Fig. 5.

The tool using Dataset 2 confirms the Hypothesis H2, boosting performance in comparison with the most common OpenMP default number of threads which corresponds to the number of cores, i.e 2, in the majority of the runs. It can be concluded that Dataset 4 results in slightly better performance according to the frequency of the Dataset 4 points that lie below Dataset2, which agrees with the minimum average out of these two datasets.

Dataset 3 and Dataset 4 out of the four datasets give the most accurate results according to the average elapsed

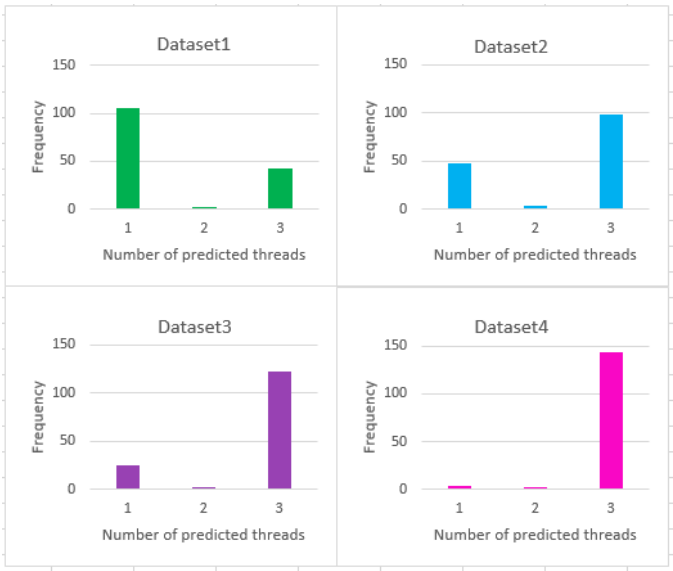


Fig. 5. Histograms of predicted threads for executing Algorithm II in a no stress environment.

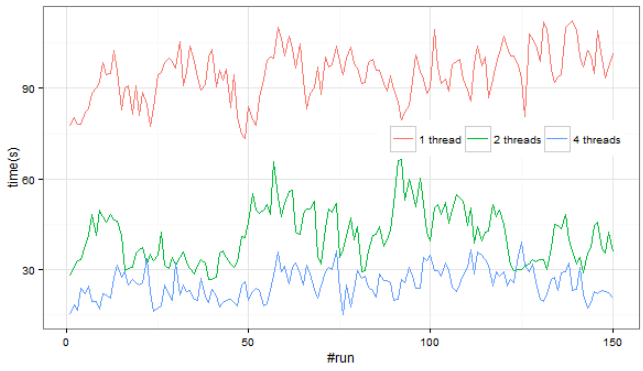


Fig. 6. Performance of executing the Algorithm I in a stress environment with OpenMP default schemes

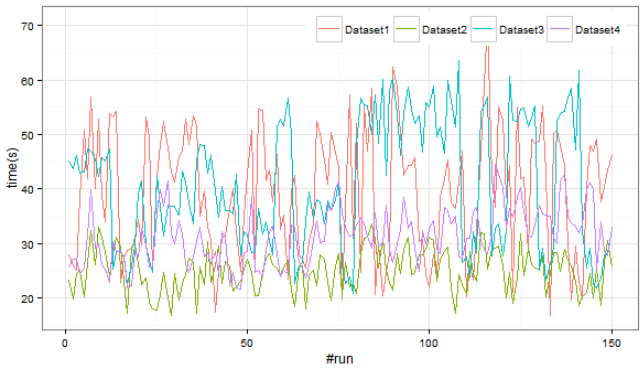


Fig. 7. Performance of executing the Algorithm I in a stress environment with Datasets 1 -4

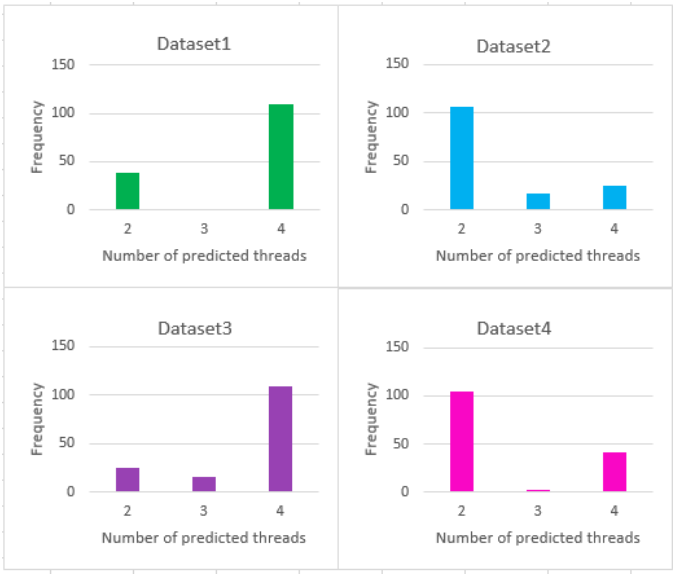


Fig. 8. Histogram of predicted threads for executing the Algorithm I in a stress environment

time. However, the average elapsed time for this algorithm is generally not a good indicator of the whether the predictions optimize the performance since the intervals [mean-sd, mean +sd] of the measurements listed in Table IV all overlap with each other except for 1 thread. Therefore the graphs along with the significant attributes should be analyzed in order to determine which dataset results in the best optimization of the algorithm.

Additionally, no dataset prediction degrades the performance, and results with a number of threads less than the number of physical cores Fig. 5.

C. Algorithm I in a Stress environment

The performances of executing the Algorithm I in stress environment are presented in Fig. 6 for default OpenMP schemes and in Fig. 7 for Datasets 1-4.

Fig. 8 presents the histogram of predicted threads for executing the Algorithm I in stress environment and contains information about the frequency associated with the predictions.

Algorithm I exploits the parallelism at a high degree and is accentuated enough so that in a stress environment using OpenMP default scheme, it is utilizing more cores. Eventually, this yields a better performance, except in very rare cases with some overlap. This overlap is located in the interval $[x - sd, x + sd]$ for a certain value of x as elapsed time. It is a case when the performance of executing the algorithm with a given dataset overlaps with the same interval for execution of an OpenMP scheme.

The execution of the Algorithm I with maximum number of threads in the OpenMP default scheme yields to the best performance, as presented in Table IV). The maximum number of threads is consequently used as a reference value when comparing the tool results with the OpenMP default scheme results.

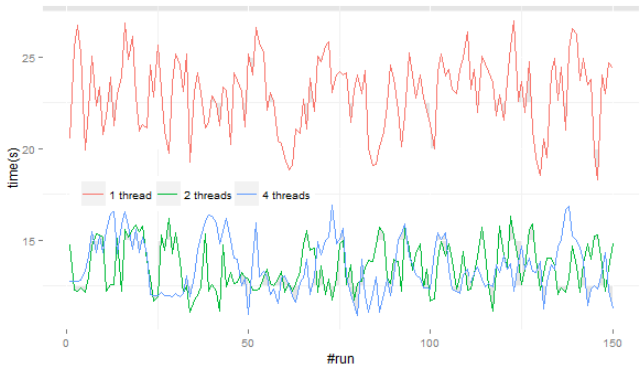


Fig. 9. Performance of executing the Algorithm II in a stress environment with default OpenMP schemes.

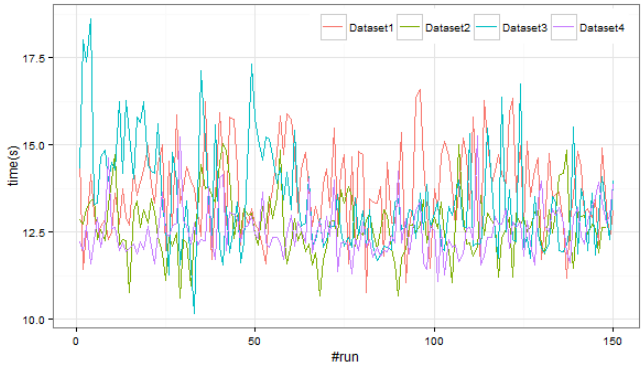


Fig. 10. Performance of executing the Algorithm II in a stress environment with Datasets 1-4.

D. Algorithm II in a Stress environment

Fig. 9 presents the performance of executing Algorithm II in a stress environment with OpenMP schemes and Fig. 10 with Datasets 1-4.

The best performance in the OpenMP schemes results in the configuration with 4 threads judging by the average. The majority of input states benefit more from 2 threads than 4 threads when comparing the overlaps shown in the performance graphs. This happens due to the variation. It is not a negligible percentage, so it should be considered when comparing the tool results with the OpenMP default scheme results. The reference optimal number of threads is therefore not uniquely determined, but is closely related to the runtime.

The results presented in Fig. 9 and Fig. 11 indicate that from the OpenMP default scheme, the optimal number of threads is leaning towards 4 threads. However the same as with Algorithm I in a stress environment, this preference will be reviewed more closely due to the higher variations of the results i.e the unpredictability caused by the the highly dynamic runtime state changes.

V. Discussion

Table IV presents the average values of elapsed times for executing the Algorithm I in a no stress environment (A1NS), Algorithm II in a no stress environment (A2NS), Algorithm I in a stress environment (A1S) and Algorithm II in a stress environment (A2S).

The obtained values of speedups for the experiments are displayed in Table V.

One can conclude that no dataset produces a speedup factor of less than 1 with the number of physical cores equal to the number of threads in (V). The exception is the Dataset 1 in stress environment with a value of 0.986805. Considering the standard deviation of the results, this value can be considered acceptable due to the stress environment. This evidence sup-

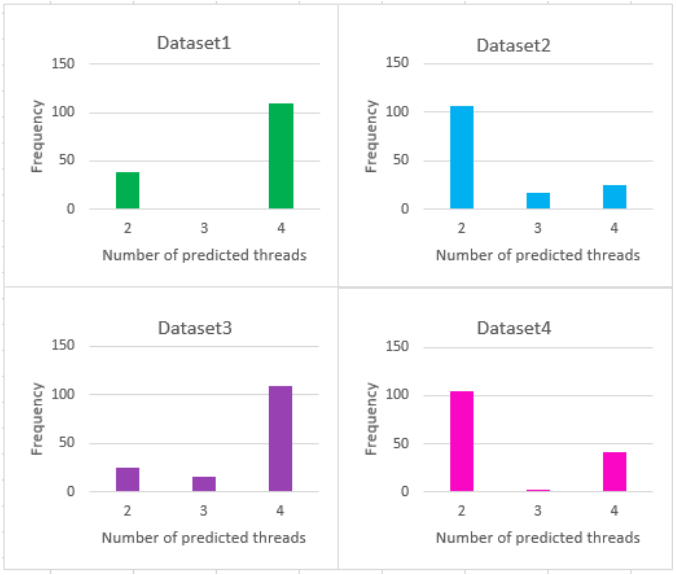


Fig. 11. Histogram of predicted threads for executing the Algorithm I in a stress environment

TABLE IV
AVERAGE ELAPSED TIMES FOR THE EXPERIMENTS

Test run	A1NS	A2NS	A1S	A2S
1 Thread	35.65783	7.752695	94.67928	22.99271
2 Threads	16.79859	5.192298	41.35630	13.52446
4 Threads	08.89994	4.724949	25.33419	13.58029
Dataset 1	11.95526	4.831201	39.43302	13.70530
Dataset 2	08.95332	4.770104	25.10371	12.70467
Dataset 3	12.16092	4.640983	40.83514	13.42115
Dataset 4	09.00934	4.556192	31.29540	12.54097

TABLE V
SPEEDUPS OBTAINED FOR THE EXPERIMENTS

	Test run	Dataset1	Dataset2	Dataset3	Dataset4
A1NS	1 Thread	2.982606	3.982636	2.932165	3.957874
	2 Thread	1.405121	1.876241	1.381358	1.864575
	4 Thread	0.744437	0.994037	0.731847	0.987857
A2NS	1 Thread	1.604713	1.625267	1.670486	1.701573
	2 Threads	1.074743	1.088508	1.118793	1.139614
	4 Threads	0.978007	0.990534	1.018092	1.037039
A1S	1 Thread	2.401015	3.771525	2.318574	3.025342
	2 Threads	1.048773	1.647418	1.012763	1.321481
	4 Threads	0.541023	0.849842	0.522447	0.681704
A2S	1 Thread	1.677651	1.809784	1.713170	1.833408
	2 Threads	0.986805	1.064527	1.007698	1.078422
	4 Threads	0.990874	1.068921	1.011857	1.082874

ports hypothesis H1 for all the datasets.

In summary, the performance evaluation yields the following conclusions. The speedups measured in the stress and no stress environment for both Algorithms I and II suggest that the predictions result in an optimized performance using either Dataset 2 or Dataset 4 for both Algorithms I and II. The indicators are the mean elapsed times combined with the conclusions made by analyzing the performance graphs that compare the results of executing the Datasets with the OpenMP default scheme results.

The maximum number of threads results in best performance for both the no stress and stress environments. Since Algorithm I is scaled vertically more successfully than Algorithm II stress testing does not result in a different optimal number of threads. The tool predictions that yield the best performance using the maximum number of threads are Dataset 2 and Dataset 4 following closely behind with slightly lower speedups. The optimization/performance boost is done by taking advantage of hyperthreading. The successful optimization using Dataset 2 and Dataset 4 confirms Hypothesis H2 for Algorithm I.

The optimal number of threads for Algorithm II in both the no stress and stress environment is not unique as in the previous case with Algorithm I. Therefore it requires more thorough graph analysis of the resulting performances. This analysis shows evidence that the tool succeeds in finding an optimal number of threads that varies from 2 to 4. Despite the high index of variance inevitably caused by the intensive stress testing, close examination of the graph/plots and the statistics of the datasets and the predictions, yields the same conclusion that Dataset 2 and Dataset 4 result again in a better performance than for the other datasets. This evidence confirms hypothesis H2 for Algorithm II.

VI. Related work

Among the research work done in area of adaptive or dynamic runtime OpenMP configuration, one of the most recent works, is based on creating a Multilayer Perception neural network with one hidden layer with back propagation to determine the number of threads for a parallel region based on the external workload [9].

Even though different machine algorithm, benchmarks and profiling tools are used, there are similarities between the goals set and methodologies used between this paper and [9]. The benefits of the approaches to optimization presented in both papers are mainly demonstrated with increase in the workload. A machine learning technique is used to predict the number of threads and is tested in an unknown setting.

Other similar work is done building up on [9] using other approaches other than neural networks in dynamic workload settings, such as reinforced learning and Markov decision processes for unsupervised learning [10]. Random forests are mentioned as a suggestion for another machine learning algorithm that can solve the problem.

Substantial amount of work is done in creating loop scheduling policies using runtime information. Other research work demonstrate promising results by using an adaptive loop scheduling strategy targeting SMT machine [11].

VII. Conclusion

The conducted experiments and performance analysis provide evidence that supports the correctness of hypotheses H1 and H2.

The main achievement is that the analysis tool provides a way to map the dynamic runtime state to the optimal number of threads in a real-time manner being able to respond to the changes in the runtime. The automation of this task becomes greater and greater advantage with increasing the number of processors and hyperthreading scale on a machine.

Therefore the significance of this paper is to explore a way of automating optimization in OpenMP and set up a groundwork for developing a truly generic cross-platform optimization tool that requires minimum configuration to produce good predictions on the optimal number of threads.

The results so far are encouraging, and the tool is able to correct the default scheme and find the optimal number of threads in cases of a highly dynamic runtime when the variation of the parameters is significant enough to provide indication of it.

References

[1] B. Barney. (2015, Aug) OpenMP. [Online]. Available: https://computing.llnl.gov/tutorials/openMP/

[2] M. Curtis-Maury, X. Ding, C. D. Antonopoulos, and D. S. Nikolopoulos, "An evaluation of openmp on current and emerging multithreaded/multicore processors," in *OpenMP Shared Memory Parallel Programming*. Springer, 2008, pp. 133–144.

[3] OpenCV dev team. (2014, Nov) OpenCV documentation, Random Trees. [Online]. Available: http://docs.opencv.org/3.0-beta/modules/ml/doc/random_trees.html

[4] L. Breiman and A. Cutler. (2004) Random Forests. [Online]. Available: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[5] R. Morgan and D. MacEachem. (2010, Dec) Sigar - system information gatherer and reporter. [Online]. Available: https://support.hyperic.com/display/SIGAR/Home

[6] P. V. V. Reddy and L. Rajamani, "Evaluation of different hypervisors performance in the private cloud with sigar framework," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 2, 2014.

[7] B. Barney. (2015, Aug) POSIX Threads Programming. [Online]. Available: https://computing.llnl.gov/tutorials/pthreads/

[8] John Burkardt. (2011, May) C++ Examples of Parallel Programming with OpenMP. [Online]. Available: http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[9] M. K. Emani, Z. Wang, and M. F. O'Boyle, "Smart, adaptive mapping of parallelism in the presence of external workload," in *Code Generation and Optimization (CGO), 2013 IEEE/ACM International Symposium on*. IEEE, 2013, pp. 1–10.

[10] M. K. Emani, "Adaptive parallelism mapping in dynamic environments using machine learning," Ph.D. dissertation, The University of Edinburgh, 2015.

[11] Y. Zhang, M. Burcea, V. Cheng, R. Ho, and M. Voss, "An adaptive OpenMP loop scheduler for hyperthreaded SMPs," in *ISCA PDCS*, 2004, pp. 256–263.

An effective Task Scheduling Strategy in multiple Data centers in Cloud Scientific Workflow

Esma Insaf Djebbar
Department of Computer Science
University of Oran 1, Ahmed Ben Bella
Oran, Algeria
esma.djebbar@gmail.com

Ghalem Belalem
Department of Computer Science
University of Oran 1, Ahmed Ben Bella
Oran, Algeria
ghalem1dz@gmail.com

***Abstract*— Cloud computing is currently the most hyped and popular paradigm in the domain of distributed computing. In this model, data and computation are operated somewhere in a cloud which is some collection of data centers owned and maintained by a third party. Scheduling is the one of the most prominent activities that executes in the cloud computing environment. The goal of cloud task scheduling is to achieve high system throughput and to allocate various computing resources to applications. The Complexity of the scheduling problem increases with the size of the task and becomes highly difficult to solve effectively. In this research, we propose a task scheduling strategy for Cloud scientific workflows based on gang scheduling in multiple Data centers. The experimentation shows the performance of the proposed strategy on response time and average cost of cloudlets.**

Keywords*— *Cloud scientific workflow, task scheduling, gang scheduling, multiple Data centers.

I. INTRODUCTION

Cloud computing is a distributed computing paradigm that mixes aspects of Grid computing, Internet computing, Autonomic computing, Utility computing, and Green computing. Cloud computing is derived from the service-centric perspective that is quickly and widely spreading in the Internet Technology world. From this perspective, all capabilities and resources of a Cloud (usually geographically distributed) are provided to the users as a service, to be accessed through the Internet without any specific knowledge of, expertise with, or control over the underlying technology infrastructure that supports them. Cloud computing offers a user-centric interface, that acts as a unique point of access for users' needs and requirements. Moreover, it provides on-demand service provision, Quality of Service (QoS) guaranteed offer, and autonomous system for managing hardware, software, and data transparency to the users.

Cloud computing has recently received considerable attention, as a promising approach for delivering Information and Communication Technologies (ICT) services as a utility. In the mechanism of providing these services it is necessary to improve the utilization of data center resources which are operating in most dynamic workload environments. Data centers are the essential parts of cloud computing. In a single data center generally hundreds and thousands of virtual servers run at any instance of time, hosting many tasks and at the same time the cloud system keeps receiving the batches of task requests. During this context, one has to notice a few target servers out of many powered on servers, which can fulfil a batch of incoming tasks. So Task scheduling is a valuable issue which greatly influences the performance of cloud service provider. Traditional approaches that are used in optimization are deterministic, fast, and give perfect answers but often tends to get stuck on local optima. The complexity of the task scheduling problem belongs extremely large search space with correspondingly large number of potential solutions and takes much longer time to find the optimal answer. There is no ready made and well outlined methodology to solve the problems under such circumstances. However in cloud, it is tolerable to find near best solution, preferably in a short period of time. This work is a continuation of our work presented in [7].

The rest of the article is organized as follows. Section 2 presents the related work. Section 3 introduces the basic strategy of the proposed approach, gives an example and analyzes the research problem. Section 4 demonstrates the simulation results and the evaluation. Finally, Section 5 addresses our conclusions and future works.

II. RELATED WORKS

There are so many algorithms for scheduling in Cloud computing. The main advantage of scheduling algorithm is to obtain a high performance. The main examples of scheduling algorithms are First Come First Served (FCFS), Round-Robin (RR), Min-Min algorithm, and Max-Min algorithm.

A. FCFS Algorithm

First come First serve Algorithm basis means that task that come first will be execute first.

B. Round-Robin Algorithm (RRA)

In this Scheduling algorithm, time is to be given to resources in a time slice manner.

C. Min-Min Algorithm

Min-Min Algorithm selects the smaller tasks to be executed first.

D. Max-Min Algorithm

Max-Min Algorithm selects the smaller tasks to be executed first.

Scheduling in cloud computing can be categorized into three stages:

- Discovering a resource and filtering them.
- Selecting a target resource (Decision stage).
- Submission of a particular task to a target resource.

Lin and Lu [1] developed an algorithm for scheduling of workflows in service-oriented environments. Unlike algorithms for Grid systems, it is able to utilize dynamic provisioning of resources. However, it misses the capability of considering the cost for utilization of resources required for its utilization in Cloud environments.

In [2] is presented a scheduling algorithm based on cost of workflows for real-time applications. The purpose of the algorithm is to develop a scheduler that minimizes the cost and still meets the time constraints imposed by the user. The workflow is divided into subsets of tasks for establishing a single flow. Tasks that do not form a single flow are separated and each of them runs as an independent subset.

In [3], is presented a strategy of a dynamic workflow scheduling that treats the relationship user/resource. In this approach the resources are not seen individually, but grouped. The scheduler, in this approach, selects the sites and this selection is made by an opportunistic strategy. It aims to spread the tasks of the workflow through Grid sites based on their performance in previous submissions.

III. PROPOSED STRATEGY

In the Cloud computing system, several tasks are running simultaneously, the tasks will have more access to data. To run a task, the data must be aggregated and this requires more data movement. Therefore, if several tasks are used the same data, they must be placed together to minimize the frequency of data movement. The proposed approach includes two important stages. Each of which contains a set of operations to be performed. Figure \ref{fig1} shows a global view of the approach.

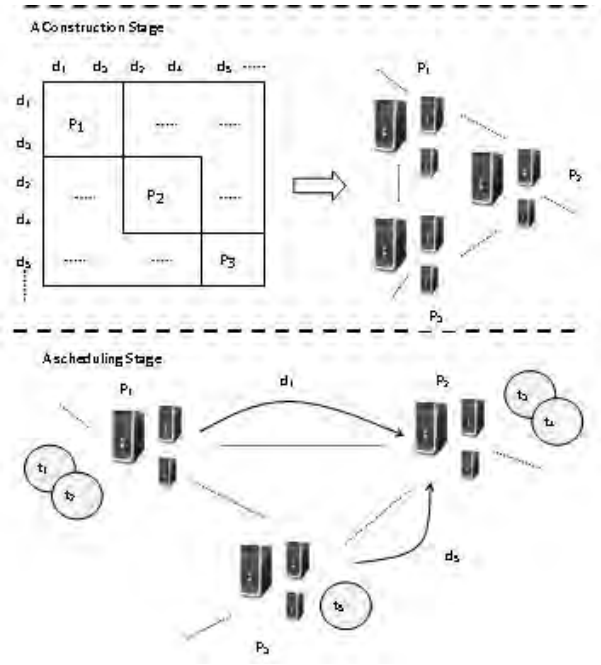


Fig1. A global view of the proposed approach

A. Stage of construction

During the construction phase, we use a matrix model to represent the existing task. We form clusters in the task set by the transformation of matrix, and then we distribute the data set to different data centers as the original partitions to be used for the next step.

First, we compute the task dependencies of all tasks and accumulate a matrix *TM* whose elements TM_{ij}= $dependency_{ij}$. The value is the dependency between the tasks T_i and T_j. It can be calculated by counting data in common between the sets of task that are denoted as T_i and T_j. Specifically, for the diagonal elements of the *TM*, each value means the number of data that will be use by this task. *TM* is a symmetric matrix of dimension *n*n* where *n* is the total number of existing Datasets.

The Bond Energy Algorithm (BEA) is applied to the matrix *TM* in order to group similar values together. Two measures, BEC and BEL are defined for this algorithm. The permutation is done so that these measures (see Formulas1 and 2) are maximized:

$$BE_{C_{ij}} = \sum_{i=1}^{n} TM_{i,j} * TM_{i,j+1} \quad (1)$$

$$BE_{L_{ij}} = \sum_{i=1}^{n} TM_{i,j} * TM_{i+1,j} \quad (2)$$

After applying the algorithm to group the *BEA* similar values in the matrix. The *TM* matrix is partitioned according to the number of data centers in multiple clustered arrays named: *CM1, CM2, CM3, ...*

B. Stage of scheduling

The dependency matrix (i.e*TM*) is dynamically maintained at this phase. When a new tasks are generated or added to the system by user, we calculate their dependencies with all

existing tasks and we add them to the matrix *TM*. We take the dependency matrix (*TM*) as input and generate the clustering dependency matrix (*CM*). In the *CM*, the items with the same values are grouped together.

Before worrying Datasets that will be generated, first we must run existing tasks. Since the movement of Datasets of a data center to another is more expensive than the scheduling of tasks to the Datacenter. A job scheduling algorithm is used (Algorithm of scheduling).

Algorithm 1 Algorithm of scheduling

Require: T: set of tasks, DC: set of datacenters.
Ensure: All tasks sequenced to an appropriate datacenters.
  for each $t_i \in T$ do
    if datasets required by t_i are available then
      Schedule t_i to dc_j to be executed where dc_j has the majority ($\sum_{i=1}^{n} Data/2+1$) of datasets required by t_i
      set status t_i=ready
    else
      status t_i=wait
    end if
  end for
  for each $t_i \in T$ do
    if status t_i=ready then
      Execute t_i
    end if
  end for

In this algorithm, the technique used is based on the placement of Datasets; the ready tasks are scheduled to the Datacenter that contains the majority of the Datasets required. A task is said to be ready if all required Datasets belong to the set of existing Datasets. Once the tasks are completed, new Datasets are generated.

IV. Experimentations and Results

In this section, we describe the experiments that we conducted to evaluate the proposed approach. Experiments were conducted with the CloudSim toolkit [4, 5, 6]. The objective of CloudSim Framework is to provide a generalized and extensible simulation that allows the modeling, simulation and experimentation of new infrastructure cloud and associated application services. We used in our work a simulator CloudSim version 3.0.3.

A. Average Response time

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

In this first simulation, we calculated the average response time with TimeShared and TimeShared Clustering (proposed strategy). For a different number of Cloudlet (20, 40, 60, 80, 100) with a length corresponding to the Cloudlet. In Cloudsim, the TimeShared algorithm can handle multiple requests (Cloudlet) in the same time but cloudlets must share the computing power of the machine, the functioning of the TimeShared algorithm is equivalent to the Round Robin algorithm. The following figures (Figure 2) and (Figure 3) show the result of implementation of the average response time.

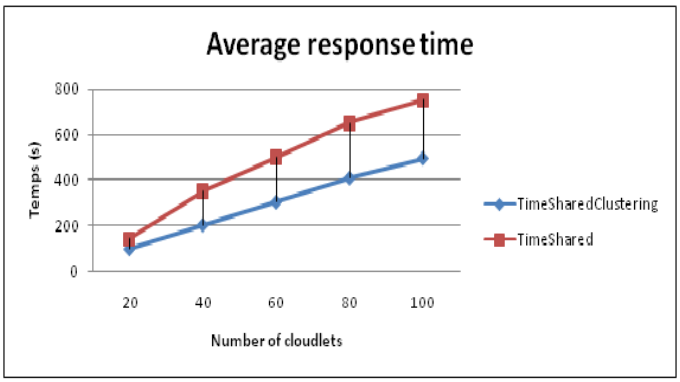


Fig2. Average response time of cloudlets

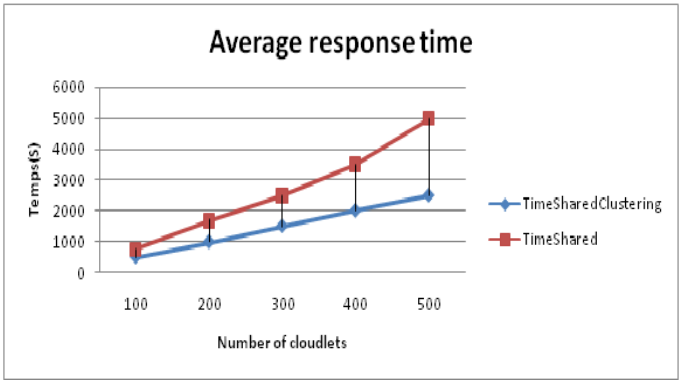


Fig3. Average response time of cloudlets

Based on these results, we note that the average response in Time Shared increases each time that the number of cloudlets rise because many cloudlets are processed at once while running takes time to process all cloudlets therefore the average response time increases each time. In Time Shared Clustering, the average response time is low because the Cloudlet are divided in different Datacenter.

B. Average Cost of cloudlets

In this series of simulation, we calculated the average processing cost of Cloudlet with both algorithms (TimeShared and TimeShared Clustering). In (Figure 3 and Figure 4) are presented, on the bar charts, the main executions performed on that scenario.

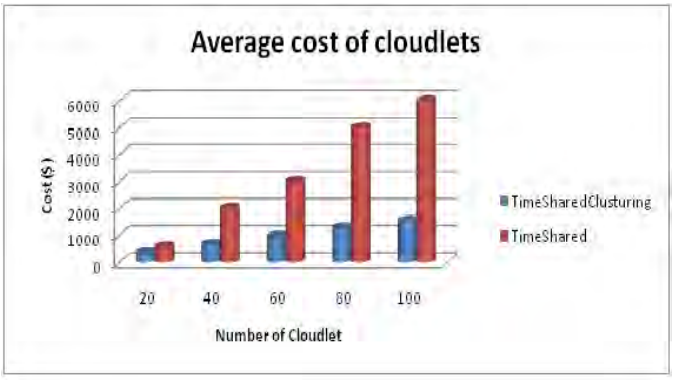


Fig4. Average cost of cloudlets

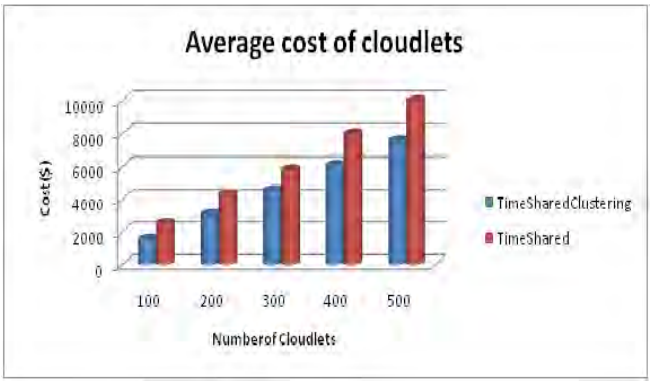


Fig5. Average cost of cloudlets

The objective of this series of simulation is to study the impact of our application on the cost of processing cloudlets. Based on these results, we note that the average cost in TimeShared algorithm is very high compared to TimeShared Clustering because the CPU utilization is less in TimeShared.

V. CONCLUSION AND FUTURE WORKS

Cloud computing implies that the data processing is not performed only on local computers, but on third-party data center. It refers equally to applications delivered as a service over the Internet as infrastructure as a service, i.e. the compute resources and / or storage. These technologies appear to be key solutions for companies and research teams with modest budgets. It allows to provide users with a set of application without having a lot of resources, the user connects to the cloud service provider site uses the applications proposed to it without realizing it accesses machines (virtual or not) different, it can also store personal data on remote servers, all of these services, which are available to users, are provided through the three cloud models, namely SaaS, PaaS and IaaS.

This work deals with the problem of task scheduling and their influence on the performance and cost of execution on the type of cloud IAAS. The initial objectives were to propose a method to distribute costs and improve runtime execution and implement and study the impact of our application on the performance of Cloud Computing. We have established simulations under CloudSim simulator environment because it is the best and the most popular simulator for Clouds. The objective of this simulation was to make experiments of some metric; the wait time, response time and the cost of processing with the two algorithms based on Shared Time (First Came First Served) and SpaceShared which is based on (Round Robin).

This work seeks primarily to improve cloud services including the execution time of applications and the financial cost. For this, we compared the two algorithms TimeShrad and TimeSharedClustering. The results of the experiments are carried out under the CloudSim simulator prove encouraging that meet our expectations. The results of experiments carried out under the CloudSim simulator prove encouraging and have met our expectations.

References

[1] C. Lin and S. Lu, SCPOR: An Elastic Workflow Scheduling Algorithm for Services Computing, in Proc. Int'l Conf. SOCA, pp. 1-8, 2011.

[2] J. Yu, R. Buyya, and C. K. Tham. Cost-based scheduling of scientific workow application on utility grids. In E-SCIENCE '05: Proceedings of the First International Conference on e-Science and Grid Computing, pages 140-147, Washington, DC, USA, 2005.

[3] L. A. V. C. Meyer, D. Scheftner, J.-S. Vckler, M. Mattoso, M. Wilde, and I. T. Foster. An opportunistic algorithm for scheduling workflows on grids. In VECPAR, volume 4395 of Lecture Notes in Computer Science, pages 1-12, 2006.

[4] R.N. Calheiros, R. Ranjan, A. Beloglazov, C.A.F.D. Rose, and R. Buyya, CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms, Softw., Pract. Exper., vol. 41, no. 1, pp. 23-50, Jan. 2011.

[5] Rajkumar Buyya, Rajiv Ranjan and Rodrigo N. Calheiros, Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities.

[6] Rodrigo N. Calheiros, Rajiv Ranjan, Anton Belogla zov, César A. F. De Rose, and Rajkumar Buyya, CloudSim A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms.

[7] Esma Insaf Djebbar, Ghalem Belalem. Optimization of Tasks Scheduling by an Efficacy Data Placement and Replication in Cloud Computing. in International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), pages 22-29, December 2013.

Visualization in the ECG QRS Detection Algorithms

A. Ristovski*, A. Guseva*, M. Gusev**, S. Ristov**
*Innovation Dooel, Skopje, Macedonia
**Ss. Cyril and Methodius University, Skopje, Macedonia
info@innovation.com.mk

***Abstract*—Digital ECG data analysis is a trending concept in the field where applied computer science and medicine coincide. Therefore, in order to meet the requirements that arise, our R&D team has created an environment where developers can test different approaches in data processing. To assist the objective, the platform offers a number of features with the following main target goals: 1) to increase the effectiveness in conducting a proper medical diagnose, 2) to incorporate a unified format of storing the results of the diagnosis conducted and 3) to test various ECG QRS detection algorithms.**

***Index Terms*—*ECG Analysis*; *Modular Software Architecture*; *AI Agent*.**

I. Introduction

The striding process of data digitalization has not left the ECG data immune to it. For an instance, modern professional ECG instruments often feature a digital output in addition to the hard copied records. While hard copied records have the role of feasible hand-outs and provide medical personnel an immediate overview of a patient's condition, the digitalized records have proven to be convenient for storage. Also, there are devices know as Holter monitors, that take continuous readings for a period of one or several days. Holter monitors assist in the diagnosis of conditions that require an extended keep up with the heart's physiology. Their output is inevitably digitalized since it is impossible to keep the lengthly recording on paper. Not long ago, the medical gimmick market became abundant with gadgets that administer the user with simplified ECG readings, again, taking advantage of the most convenient storage form - digitalized data.

No matter the level of exhaustiveness and accuracy of the information contained in the digitalized data, many opportunities rise in the filed of digital data processing. What is of greatest significance, is the opportunity to give a diagnosis on the recordings without the assistance of medical personnel, i.e. to replace human resources with a virtual Artificial Intelligence (AI) agent [1].

The extent of details provided in a conducted diagnosis vary according to the needs. In some cases, only statistical data targeting a specific medical condition or property is collected, and in others, a thorough analysis of the ECG recording is handled - to deliver an accurate cardiac profiling.

Regardless the scope of the diagnosis, visualizing the outcome of the signal processing while testing different algorithms is of substantial benefit. Applying a filter to the signal before it is processed can increase the effectiveness of the algorithm used for the processing. Therefore, for the purpose of having an insight how the filter has effected the initial signal, a graphic representation of the processed signal in contrast to a representation of the original one is a must. That way, by modifying the filters' parameters and immediate observation of the results, the process of improving the effectiveness gains a significant advance.

An ECG signal consists of a sequence of form-specific element set of typical waves that appear in a unambiguous pattern, at least to say unambiguous for most of time. These typical waves are referred to as components of the signal or characteristic points. Although the signal's components can be easily distinguished by the human eye, it is quite the challenge to teach an AI agent how to do so. By visually marking which signal components have been detected and their exact point of detection, one can determine to what extent the detection algorithm has been successful. That way, the left-outs can be consequently analyzed and the algorithm can be modified accordingly, hence, doing corrections in a favorable direction.

Some algorithms may even use pattern recognition in the detection of the signals' components. By visualizing how does a component fit in the pattern it is supposed to fit in, the pattern can be improved in a way so that it better detects the components that have been left out during the detection. In addition, one of the more advanced features when using a pattern recognition is the ability to asses which part of the components deviates from the general form and in what way.

Therefore, using visualization in the development of algorithms which do the ECG component detection is of great interest. The goal is to design an AI agent that will mimic the way a medic would interpret a reading. Although the AI agent itself does not need visualization of any kind to do its job, it is practically inevitable for the development process to feature the use of visualization modules, i.e. provide developers a visual insight in how the approach they are using is affecting the correlation between the input data - the ECG signal and the output data - a legitimate diagnosis. Whereas the paper is focused on the role of visualization in the AI agent design process, it also encloses an architectural design of an ECG diagnosis application that is not only aiming to deliver accurate results, but also, to incorporate functionalities for testing new approaches and improve overall effectiveness.

The paper is organized as follows. The architectural design of the solution is given in Section II. Section III describes the architectural modules in detail and explains how the visualization has been implemented how it assists the process of devising the detection algorithms. An overview of existing related work is handed in Section IV. Section V discusses the conclusions and directions for future work.

II. Architectural Design

The features of a tool designed for delivering an ECG diagnosis are incorporated into the modules of a system with modular software architecture. Such design allows the advantages of distributed development, code re-usability and modular transparency, hence, the integration of new functionalities and the upgrade of existing ones can be done easily.

The process of AI agent assisted diagnosing is very close to how a medic would act upon doing so, with respect to the steps needed to utilize the input and output formats that make a notable exception. The whole routine consist of several stages. These include:

- Unpacking and interpreting the ECG recording, whatever the file format;
- Pinpointing the signal components of the signal:
  - P waves;
  - QRS complexes;
  - T waves;
- Analise the constellation and form of the signal components;
- Set a medical diagnosis;
- Save the annotations;

In addition, the system is capable to evaluate the effectiveness of the applied algorithm on an extensive database of patients. It responds with several key effectiveness indicators, such as *detection rate* (percentage of detected features), *hit rate* (percentage of correctly detected features), *miss rate* (percentage of incorrectly detected features), and also the *extra rate* (percentage of extra determined features, which are not recognized by a human and have to be eliminated). This module is efficient in recognizing those parts of the ECG which caused problems (extra features or misses) and a user can automatically analyze them and improve the algorithm by customizing and adapting its parameters.

The product reviewed in this paper has been designed in accordance with the stages of ECG analysis numbered above and is using a separate software module for each stage. Guidelines for pragmatic processing and several assisting visualization modules are also designed in accordance with the concept of modular architecture. They have been offered at disposal in order to make the whole process simpler and more transparent for developers. Although the architecture makes a notably pragmatic solution for the problem, what is more important is that it excels at increasing the effectiveness of the diagnostic capabilities.

The overall architecture of the system is presented in Fig. 1. The next section gives a description on how each module deals with the problems in respect, and, moreover, focuses on the interest of implementing the modules for visualization.

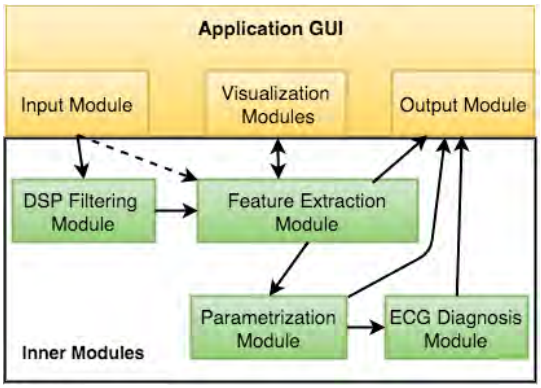


Fig. 1. An overview of the system's architecture

III. Functional Description

One of the fundamental ideas for an application with such architecture is interoperability. That is achieved by allowing data input in different file formats, hence, making it possible for the application to process output data from a variety of ECG sensors. The application's *input module* handles the different data representations containing the actual ECG signal, so that the data is given a unified composition - an inevitable prerequisite for the next stage of the analysis.

There are several standard medical record data formats for storing and retrieving ECG files. Although most of them are XLM based, still there are examples of binary files, such as those used by PhysioNet open source library [2]. The most profound is maybe the USA FDA standard HL7 [3] that specifies annotated ECGs and provides means to systematically evaluate the ECG waveforms and measurement locations. Bond et al. [4] review nine different formats used to store the ECG, such as SCP-ECG, DICOM-ECG, and HL7 aECG. The input module also does the consolidation of additional record information, such as sensor ID, time stamp, location, type of ECG lead being analyzed [5] and alike.

The *ECG signal visualization module* incorporated into the application's GUI is an important module for visualization of the ECG signal. By the means of this feature, the developers have an immediate insight on how does the analog representation of the digital signal look like, as if looking at a hard copied ECG recording.

Therefore, this visualization module is responsible for displaying the consolidated ECG signal. The OpenGL libraries [6] have been used for the purpose of visualizing the data. Since the development environment has incorporated the C# programming language, a wrapper for the OpenGL libraries is needed. The choice of OpenGL wrapper has come down to the OpenTK wrapper [7], superseding the Tao Framework [8].

The application interface displays the additional information of the ECG signal and highlights the analog representation of the ECG signal in a standard ECG graph paper surrounding [9]. A standard record duration of 10 and 30 seconds are the

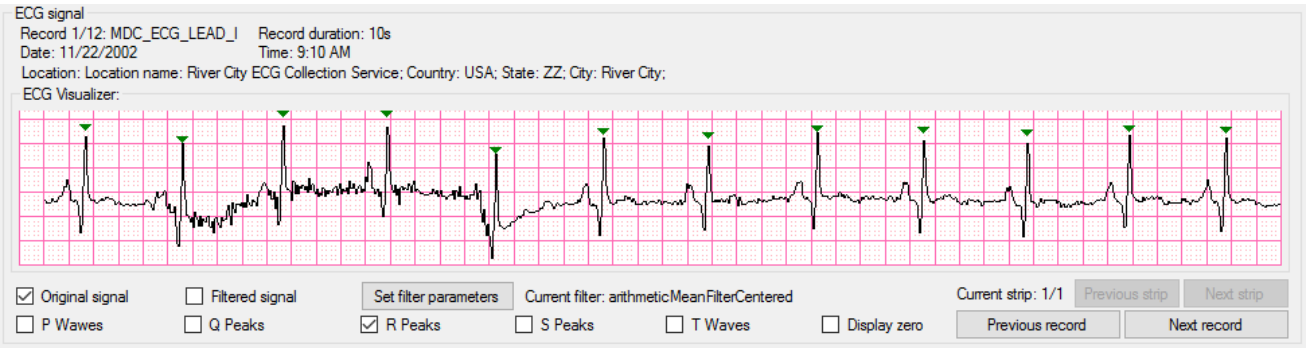


Fig. 2. A snapshot of the signal visualization GUI component

most common length of an ECG record, as the information contained within a 10 seconds reading is sufficient for a standard diagnosis. Moreover, records with these lengths are optimal in terms of appearance since longer records would make the visualized signal appear cluttered.

However, records with longer duration are not nonexistent. Certain diagnosis types require readings that last up to half an hour. Visualization of such records has been also supported, by enabling transition between consecutive 10 second strips, thus, maintaining optimal appearance.

The size of the visualization component in the Graphical User Interface (GUI) is scalable, so that it adapts to the dimensions it is supposed to have, hence, the ECG recording being displayed scales up to provide maximum clarity and visibility. The height of the graph paper squares scales up as well, to conform with the signal's amplitude. Optionally, a zero amplitude line can be displayed as well.

The *DSP module* is another auxiliary application module used for filtering the digital signal from noise. The effectiveness of the upcoming second phase of ECG analysis can be significantly improved by applying a filter to the signal. The filtering can realize various DSP methods, such as, interpolation, moving average, regressive filters, differential filtering, wavelets, discrete Fourier transformations, etc. The module allows use of a combination of any of these filters and customizing their parameters.

The processed signal is at request displayed within the same visualization module, next to the original signal - in order to give the developer an insight of how applying the filter and changing the filters' parameters is affecting the signal. The GUI component is shown on Fig. 2. The way the ECG signal visualization module and the DSP module improve the effectiveness is shown on Fig. 3.

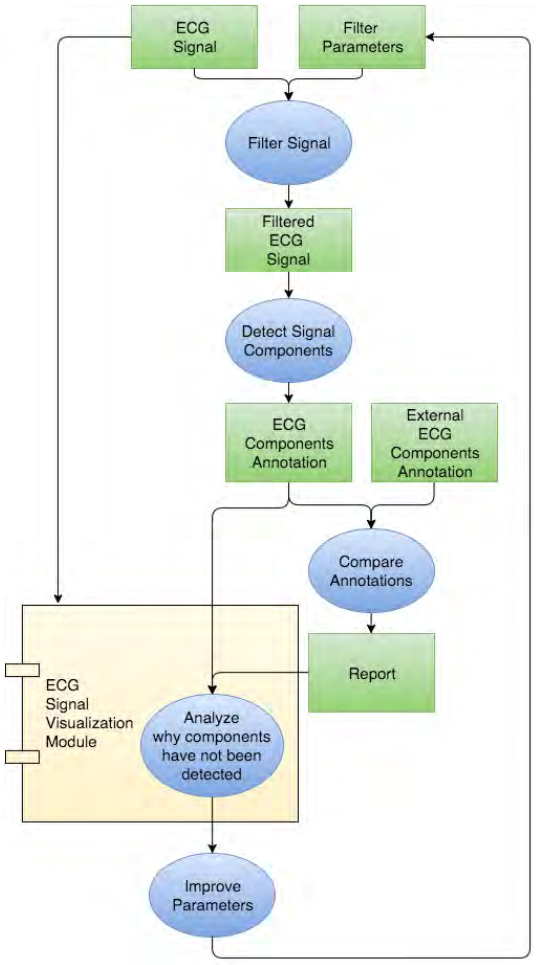


Fig. 3. An iterative process on how the ECG signal visualization improves the overall effectiveness

The second phase involves *the feature extraction module* used for detection and localization of the signal components. The easiest way of doing this is to conduct a QRS complex detection first, as this signal component has the most distinctive form and is the easiest to pick up out of the ECG recording. The other two components, the P waves and T waves, can be optionally detected as well - by conducting a look up for one of each in the interval between two neighbor QRS complexes. This stage can be done in a number of different ways: by analyzing the signal's curve slopes, by using pattern recognition, by looking for a local maximum at predefined intervals, etc. It is up to the developer to make an assessment which approach is of greatest interest. All the detected signal components are then collected in a logical structure, referred to as ECG signal component annotations, where an annotation

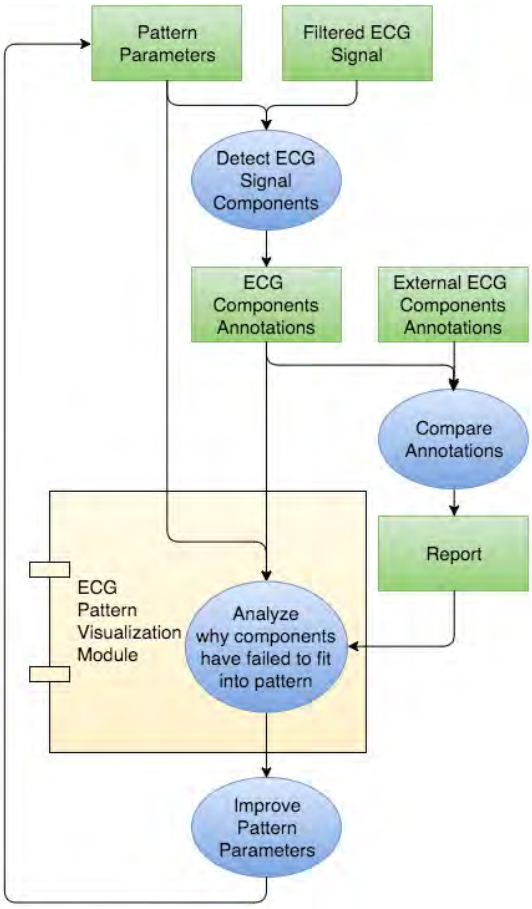


Fig. 4. An iterative process on how the pattern visualization model improves the overall effectiveness

consist of the following couple: component index (time stamp) and component amplitude.

In case of using a pattern recognition, the feature extraction module displays the pattern and the signal component entering the pattern. Because there are multiple components detected, this module enables selection of the signal component displayed on the ECG pattern visualization module. In case of undetected component, a specific part of the signal can be put on display under the pattern so that the pattern can be modified in order to better detect the components. This process is explained on Fig. 4.

In case of using the expert system to diagnose a streaming data from an ECG sensor as a web service, the feature extraction module offers an excellent opportunity to tune and customize various filter and feature extraction parameters of the algorithms to match the specifics of the person.

The ECG signal visualization module can also help with testing the effectiveness of stage two by specifically marking the signal components, i.e. by indicating which components have been successfully detected and which have not. The application's interface offers selection of the signal components that are being displayed. For each type of components, if specified for display, an indicator is drawn right above their positions.

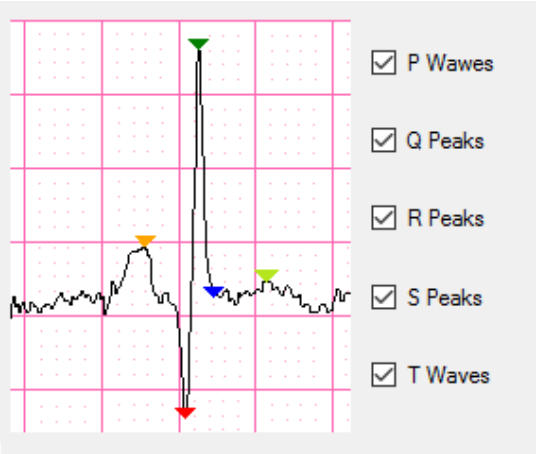


Fig. 5. A snapshot of the ECG component detection visualization

The feature is shown on Figure 5.

After the signal's components have been temporally pinpointed (localized), the third phase is realized by the *parametrization module*. It features two types of analysis: obtaining temporal parameters and obtaining vector polarization parameters. The temporal parameters' analysis starts with the calculation of typical component intervals. These intervals optionally include: PR, QRS, QT, PP and RR interval duration. Out of that data, additional temporal components are derived, such as Beats Per Minute (BPM), Heart Rate Variation (HRV), Beat Fluctuation (BFx), SDNN and other related parameters [10]. Then, the form of the signal components is being investigated into for the vector polarization analysis. Although there is small number of options how the parametrization can be done, since it comes down to a standard medical approach, the way the component extraction is handled offers significant flexibility and is, in fact, very similar to how the second phase - the component detection, is accomplished, as the signal's amplitude is taken into consideration. In case of having only one lead in the ECG recording, very little relevant information can be given for the vector polarization, because the vector polarization is a derived compound from several leads that make up the QRS phasor [5]. Therefore, when processing one lead record, the vector polarization comprises only parameters such as abnormal amplitude levels and component width.

The fourth phase consists of the *diagnosis module* that can closely correlate to how a cardiologist would conduct the ECG analysis. The information collected at the third phase is being used - a combination of a number of parameters, would trigger a medical condition marker. All the triggered markers are then reasoned into a closing composite diagnosis.

Eventually, all the information gathered at the previous phases: the ECG signal and the additional ECG record data (phase one), the ECG component annotations (phase two), the typical ECG parameters (phase three) and the ECG diagnosis (phase four) are consolidated within the *output module* into an output file. Once again, it is up to the developer to determine which information is of interest and needs to be included in

the output. The output data can then be stored into a data base, sent through a network channel or stored locally as an individual file.

Additional functionality of the final phase is the effectiveness check. There are databases that in addition to the ECG recordings offer supplemental component annotations. The annotations generated by the developed system and the pregenerated annotations supplemented to the ECG recordings from the databases can be automatically compared, via another auxiliary module, inbuilt in the output module. An extensive report of such comparison optionally accompanies the output file, and it includes analysis of the effectiveness indicators, such as the detection rate, hit rate, miss rate and extra rate.

The application has a fully functional Graphical User Interface (GUI) that allows the developer to have an actual overview of the diagnosis given, as well as interfaces for handling the application's input and output, a debug window and a status pane, as shown on Figure 6. There are additional interfaces for customizing the parameters used during the component detection process and for customizing the parameters used for the signal filtering.

IV. Related work

There are several tools that help ECG diagnosis, and most of them are open source software packages or offered as web application. ECG has been attracting the researchers a lot and even the mathematical toolkits Mathematica [11] and MatLab [12] offer extensions for ECG interpretation.

PhysioNet [13] have developed a large collection of software for viewing, analyzing, and creating recordings of physiologic signals, and their product Wave is a visual tool that enables an extensible interactive graphical environment for manipulating sets of digitized signals with optional annotations.

An ECG simulator is a device, which can simulate ECG signals recorded previously in the real time mode [14]. The main goal of ECG simulators is to convert the digital ECG signals to analogue counterparts, which have been created in a computer graphic environment [15]. The ECG Simulator enables analysis and study of normal and abnormal ECG waveforms without actually using the ECG machine. One can simulate any given ECG waveform using the ECG Simulator [16]. This simulator can be used both for clinical training of doctors as well as for design, development and testing of automatic ECG machine without having real subject for this purpose.

Most of the simulators on the market are offering services for training of readings of ECG and similar practical clinical skills. Nilsson et al. [17] propose a web-based ECG-interpretation program for undergraduate medical students. Lu et al. [18] designed a tool for modeling and simulation of electrophysiology. Kaur describes a simulator [14] that produces realistic ECG with the possibility of including other biological signals, e.g. blood pressure.

In the literature, software-based ECG simulators have been developed by the help of different programming languages such as Java, C++, C#, and MATLAB. Oefinger et al. proposed a web service Java and CGI-based, which can plot ECG signals in the ECG database called Physionet [19]. Kartnik designed a MATLAB based ECG simulator [16], which can generate normal lead II ECG wave forms.

Another example of a web-based ECG simulator is the "The Six Second ECG Simulator" [20]. The simulator "simECG" was developed using C++ language [21]. Additionally, an ECG simulator of particular interest is WebECG [22] where various ECG signals can be generated and plotted in 3-dimensional view with zooming and moving.

Therefore, most of the related available products are in fact ECG simulators, meant for medical training of medical personnel or medical software developers. However, the system proposed in this paper would assist the medical software developers in a unique way. That is, not for training individuals on how to read and interpret ECG recordings, but to enable developers to design approaches for digital ECG processing and analysis. On the other hand, the proposed solution has functional similarities to other applications meant for diagnosis, such as the products by PhysioNet.

V. Conclusion and Future Work

This developers' environment has been built so that it meets the needs of an ongoing project. The project involves analysis of ECG data recorded by a set of wearable ECG sensors. An AI agent serves the sensors and is capable of detecting any abnormalities in the subject's cardiac condition and also, capable of preventing critical scenarios on time.

So far, the application has substantially contributed to improving the effectiveness of the component detection in a number of ways: by improving the pattern recognition approach and by testing different component detection methods, by contributing in the analysis of how different filters affect the original signal, by following a strict component based architecture which increases the code re-usability in case of transitioning to other platforms.

One of the features that this tool offers is the transitioning from a desktop application to a web service. This is of high interest, because that way out R&D team could continue working on a fully web based platform, hence, make the corrections to the web server that is in charge of the processing of incoming ECG records without much effort.

Howbeit, a greater variety of filters can be implemented, so that the users can choose one or combination of several, all in order to increase the accuracy of the diagnostics. At the same time, most of the added features have an effect that can be seen "on the go" - all thanks to the visualizer module.

To conclude, although the software was designed for a specific cause, it can easily be re-purposed or upgraded with functionalities, all due to its architectural modularity. It is a product that is not of exclusive interest to the team that build it, but assists the work of others that deal with similar challenges.

The main benefit of this tool is the possibility to fine tune the parameters for DSP filtering and feature extraction in the ECG detection and, therefore, customize it towards a specific

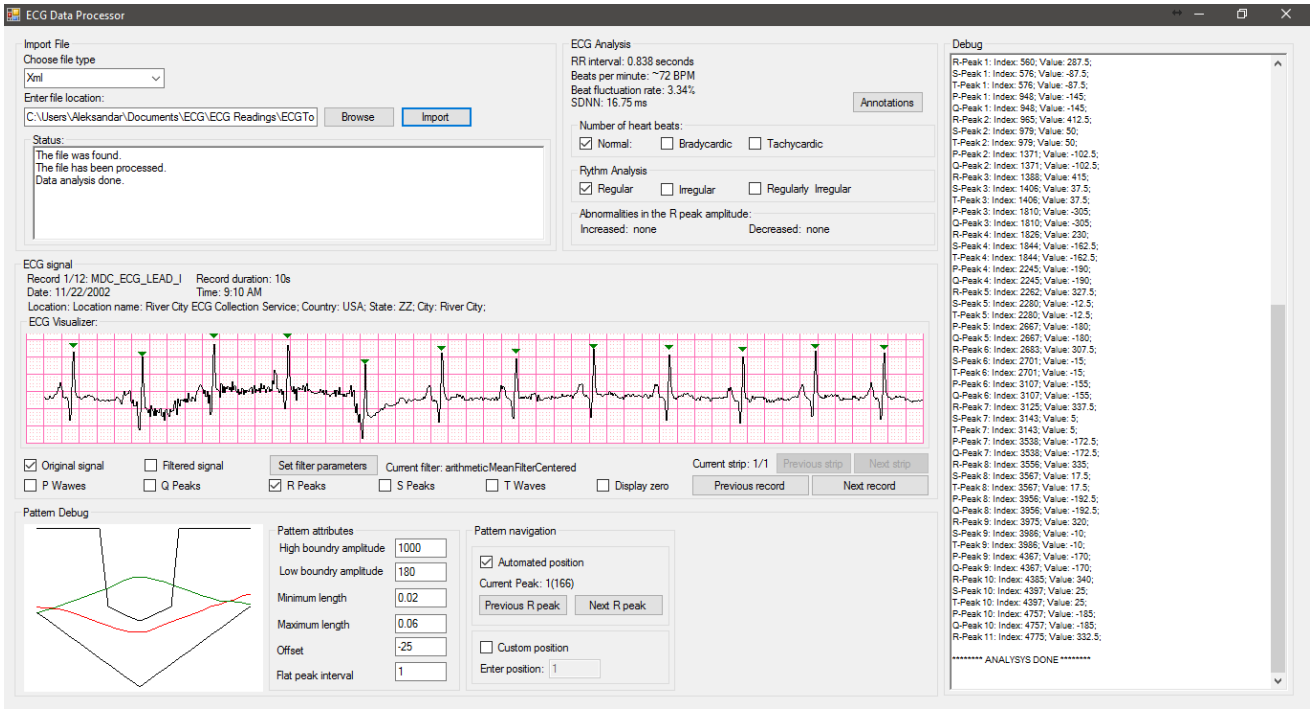


Fig. 6. A snapshot of the visual tool for ECG QRS detection

patient, especially in a case of using a streaming ECG sensor that continuously sends data to a diagnosis system. A machine learning approach can be used for the pattern parametrization in order to boost the effectiveness of the expert system.

As future work, we plan to develop more filters and feature extraction algorithms. Additionally, the system is planed to support simultaneous processing of multiple recordings, hence, it will allow handling of entire databases and record sets. The effectiveness check that covers the annotations generated from a single file, will be extended to cover the annotations of multiple files and subject them to extensive statistical analysis. We believe that such expert system would be of particular interest for students and medical researchers.

REFERENCES

[1] F. Gritzali, "Towards a generalized scheme for qrs detection in ecg waveforms," *Signal processing*, vol. 15, no. 2, pp. 183–192, 1988.

[2] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[3] Health Level Seven International. (2015, June) Hl7 version 3 implementation guide: Annotated ECG (aECG). [Online]. Available: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=102

[4] R. R. Bond, D. D. Finlay, C. D. Nugent, and G. Moore, "A review of ECG storage formats," *International journal of medical informatics*, vol. 80, no. 10, pp. 681–697, 2011.

[5] T. B. Garcia *et al.*, *12-lead ECG: The art of interpretation*. Jones & Bartlett Publishers, 2013.

[6] D. Shreiner, M. Woo, J. Neider, T. Davis *et al.*, *OpenGL (R) programming guide: The official guide to learning OpenGL (R), version 2.1*. Addison-Wesley Professional, 2007.

[7] S. Apostolopoulos, "The Open Toolkit library online documentation," http://www.opentk.com/doc.

[8] D. Hudson, R. Ridge, R. Loach *et al.*, "The Tao Framework," http://sourceforge.net/p/taoframework/wiki/Home/.

[9] E. Einthoven's, "Ecg graph paper."

[10] M. Malik, "Heart rate variability," *Annals of Noninvasive Electrocardiology*, vol. 1, no. 2, pp. 151–181, 1996.

[11] T. Pogorelov. (1998, Sep) Ekg qrs detection algorithm. [Online]. Available: http://library.wolfram.com/infocenter/Demos/4476/

[12] R. Gupta, J. Bera, and M. Mitra, "Development of an embedded system and matlab-based gui for online acquisition and analysis of ecg signal," *Measurement*, vol. 43, no. 9, pp. 1119–1126, 2010.

[13] PhysioNet.org. (2015, July) The WFDB software package, software for viewing, analyzing, and creating recordings of physiologic signals. [Online]. Available: https://www.physionet.org/physiotools/wfdb.shtml

[14] G. Kaur, "Design and development of dual channel ecg simulator and peak detector," *Thapar Institute of Engineering and Technology, Deemed University, Patalia*, 2006.

[15] C. Caner, M. Engin, and E. Z. Engin, "The programmable ecg simulator," *Journal of Medical Systems*, vol. 32, no. 4, pp. 355–359, 2008.

[16] R. Karthik. (2003) Ecg simulation using matlab. [Online]. Available: www.mathworks.com/matlabcentral/fileexchange/10858

[17] M. Nilsson, G. Bolinder, C. Held, B.-L. Johansson, U. Fors, and J. Östergren, "Evaluation of a web-based ecg-interpretation programme for undergraduate medical students," *BMC medical education*, vol. 8, no. 1, p. 25, 2008.

[18] W. Lu, D. Wei, X. Zhu, and W. Chen, "A computer model based on real anatomy for electrophysiology study," *Advances in Engineering Software*, vol. 42, no. 7, pp. 463–476, 2011.

[19] M. Oefinger and R. Mark, "A web-based tool for visualization and collaborative annotation of physiological databases," in *Computers in Cardiology, 2005*. IEEE, 2005, pp. 163–165.

[20] S. L. Canada. (2016) The six second ecg (cardiac rhythm simulator). [Online]. Available: http://skillstat.com/tools/ecg-simulator

[21] M. J. Martins AC, Costa PD. (2016) simecg: Ecg simulator. [Online]. Available: http://simecg.sourceforge.net/

[22] E. Güney, Z. Ekşi, and M. Çakıroğlu, "Webecg: A novel ecg simulator based on matlab web figure," *Advances in Engineering Software*, vol. 45, no. 1, pp. 167–174, 2012.

Analysis and comparison of algorithms in advanced Web clusters solutions

D. Alagić* and K. Arbanas**
* NTH Media, Varaždin, Croatia
** Paying Agency for Agriculture, Fisheries and Rural Development, Zagreb, Croatia
da@nth.ch
krunoslav.arbanas@gmail.com

Abstract – Today's websites (applications) represent an essential part of nearly every business system, therefore it is unacceptable for them to be unavailable due to the ever-increasing competition on the global market. Consequently, such systems are becoming more and more complex to allow for their high availability. To achieve a higher system availability, a greater scalability is used in creating the so-called Web farms or Web clusters. This system requires much more computer power than the traditional solutions. Since such systems are very expensive and complex in nature, the question is how to obtain the best possible results with the least amount of investment. To achieve that, it is necessary to take a look at the component which contains the information on the request or traffic amount, and that is the HTTP/HTTPS load balancer. The system is based on several algorithms, however, there are no comprehensive analyses that indicate which algorithm to use, depending on the Web cluster and the expected amount of traffic. For this reason, this paper provides a detailed comparison of several frequently used algorithms in several different Web cluster scenarios, i.e. loads. Also, examples are given as to when to use a certain algorithm.

I. INTRODUCTION

One of the first cloud computing solutions were simple Web sites, i.e. portals that have become a lot more complex in the past two decades due to globalization of the market. Today, companies and organizations increasingly base their operations on the forms of computing that make the information available anytime and anywhere. That is why no company or organization can afford for their Web sites to be unavailable, since that is detrimental to their reputation, business opportunities and revenue [4][14]. In order to provide constant availability of the system, both architecture and infrastructure of these solutions have become a lot more complex and, consequently, more expensive. Therefore, it is almost impossible to find the traditional solution for the Web sites; instead, everything is done with the help of Web clusters or Web farms [17]. Such cluster solutions present a kind of parallel and distributed computer system consisting of computers which are interconnected (in this case a Web server), which represent one computing resource whose performance is far greater than that of regular computers. This allows for a horizontal way of spreading computing resources to more instances of Web servers that are usually in the form of virtual machines (multiple usage of the existing resources) and a lot cheaper than the vertical growth, which represents a strong and independent purchase of powerful and expensive servers [16][20].

Such solutions offer many advantages such as a scalable growth, meaning that it is possible to add more Web servers to handle HTTP / HTTPS requests depending on the increase of the traffic. In addition, such a system allows for greater availability because the requests - with the help of the HTTP / HTTPS load balancer (hereinafter referred to as the LB) - are being distributed on several Web servers (multi-node architecture). Due to that, the chance of a single point of failure is decreased. Since the system consists of several Web servers, such solutions are more favorable when dealing with necessary system changes and upgrades, because it is possible to work on the servers one by one without causing any downtime unlike the traditional concept [8][19].

In order for that to be possible, the architecture of such systems is notably more complex, i.e., it consists of multiple components, thus entailing a greater risk that one of them will undergo an "operational dropout" or a malfunction that would result in the collapse of the entire system. In addition to a larger number of components (servers, storage, etc.), the system is more complex because it uses a more sophisticated equipment with various distributed systems, as well as a large number of protocols for all that to function without a flaw. Besides, such complex systems are quite expensive because they generate large OPEX and CAPEX costs, wherein their usage is not always brought to the maximum. An example of OPEX costs is the maintenance of the infrastructure (servers, network, etc.), as well as other items, such as other monthly expenses: Internet connection, electricity, air conditioning, renting of the space, etc. An example of CAPEX costs is an investment in new servers or network equipment necessary for the operation of Web farms [5].

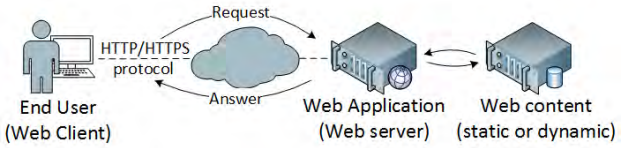


Figure 1. Schematic representation of the communication between end user and the Web server (the traditional model)

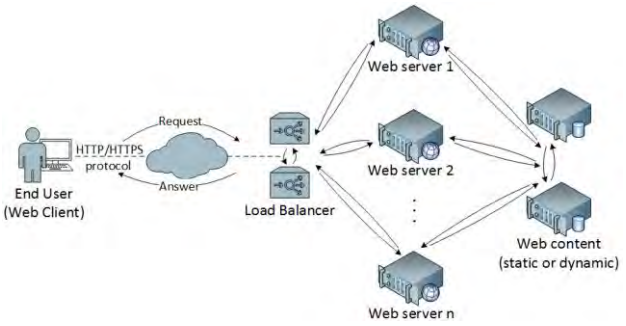


Figure 2. Schematic representation of the communication between the end user and the Web servers in the Web farm

The main problem here is that the equipment amortization rate is very high, therefore the equipment is trying to be put to a multiple use [10][21].

Due to all these costs, the already existing resources need to be used in the best possible way to minimize the costs, all the while making sure to not hinder the results. To do so, we will take a look at one of the main components of a Web farm, which is the LB system based on a number of algorithms that have been used for many years. Considering the fact that the aforementioned algorithms are widely accepted in the LB systems and many others, the question is which algorithm to use and in which case for the optimal utilization of the system.

In order to find out which algorithm will provide the best results in certain cases, it is necessary to fully understand their operating principles. Depending on the type of distribution of the tasks, i.e. the requirements, the LB algorithms are divided into static (SLB) and dynamic (DLB) ones [6][13][14].

In SLB algorithms, the requirements are assigned to the executors (i.e. nodes) from the start, depending on their characteristics. During the operation, it is not possible to change the tasks or increase the number of nodes.

TABLE I. COMPARISON BETWEEN DIFFERENT ALGORITHMS FOR THE DYNAMIC LOAD DISTRIBUTION [1]

Algorithm	*The possibility of verifying the state*	*Performance*
Rand	No	Excellent
Prand	Partially	Excellent
Acwn	Yes	Good
Pacwn	Yes	Good
Cyclic	Partially	A little better than the rand
Probalistic	Partially	Good
Threshold	Partially	Better
Least	Partially	Better
Reception	Partially	Not so good
Central	Yes	Excellent
Global	Yes	Good
Offer	Yes	Bad
Radio	Yes	Good
Sed	Yes	Good
Nq	Yes	Good

Another disadvantage is that the SLB algorithms do not collect any feedback on the nodes' status [1][9]. Due to these limitations, we will consider further only the DLB algorithms since they are quite flexible, and can be modified and improved easily. The following table shows the comparison between the existing DLB algorithms.

This paper will not compare and test all those algorithms, but only those which are commonly used and are generally available (open source).

As previous research showed (see the table above), there is not a large number of DLB algorithms used in different situations, i.e. there is no universal DLB algorithm to be applied to all computing distributed systems [1]. Although almost all DLB algorithms have the ability to verify the nodes status, i.e. the flow of information and their load, none of them has the complete information on the status of compute power. Their verification applies only to the status and availability at the application level, but not at the level of the operating system. Thus, it is presumed that a LB system could be more effective if it had a complete overview of the system, i.e. an overview of its resources and the actual load. However, that is a topic for another research.

II. RESEARCH PROBLEM

Previous research in cloud computing is related to the load balancer systems in Web clusters [2][3][11] [16][17][18] [19], but also to the algorithms used in such arrangements [1][7][15][17][20]. There are many researches that deal with the topic, but there are two problems:

1) **Testing repeatability** - in the tested environment, the measurements are described poorly or not at all, while the test samples are quite specific. In other words, it is not possible to track those results in order to further expand the existing research.
2) **Partial analysis** - those load balancer algorithms that were compared are incomplete, i.e. a small number of algorithms was compared, and in many cases the observed variables are not clearly stated or defined.

For this reason, this paper will provide a complete analysis of the LB system as well as the comparison between its algorithms. During this research, the test environment will be described in detail and the applied technologies will be open source to enable for a higher solution availability which enables the obtained data to be used for future research.

III. DATA SOURCES

The relevance of the data is provided by inclusion of the data from a number of IT companies which provide cloud computing services, or more accurately, Web hosting services. These solutions are based on the Web clusters that were previously explained.

All research and testing was carried out in one of the data centers with a full access to the equipment and the necessary financial data, i.e. the indicators essential for this research.

In the given example, it is evident that the company has a number Web clusters, categorized according to the type of business or content. There are several reasons as to why this type of architecture was opted for; one of them is to achieve a high system availability, since a certain group of Web sites is constantly exposed to security attacks causing a hindrance and potential unavailability of the system. The chart also makes it possible to conclude that the resource utilization is not linear and that it depends on the type of the content. In other words, informative and business Web sites are visited more during the day, while adult sites and sites containing video games are frequented by night.

Given that there is a large part of computing resources that are hibernating (that are not used optimally in the course of 24 hours), the question is how to increase the level of utilization of the computing resources. If we present all this data on one chart, i.e. on a single Web cluster, it can be noted that an overload is present on the entire cluster. In other words, not even a larger system could handle the entire load, wherein there would be time periods when it would have very little traffic or load. Thus, whether it be a number of separate Web clusters or one major cluster, the computing resources would not be used properly, and if we go back to the beginning of this topic, we can remember how these resources generate large monthly and capital costs.

Thus, the goal is to maximize the existing resources without increasing the number of Web servers, as it is done with commercial tools for the virtualization and configuration of Web clusters, such as VMware. Such tools perform a continuous monitoring of all Web servers and their resources, causing a deploy of virtual machines or Web servers in critical situations (e.g. in the case of a sudden increase of traffic and the subsequent system overload). This guarantees for the stability and availability of the system, but it also generates additional costs due to the increased number of servers. One way to make a better use of computer resources is to choose the best LB algorithm depending on the situation and the amount of traffic that is present or expected.

This may not necessarily be true, but in order to remove any doubt, this paper provides the analysis and the comparison of a number of most commonly used algorithms according to several different cases varying in the request amount. The results will be gathered and compared from a number of computer resources monitoring tools, but also from the LB system in which it is possible to monitor the load and the number of visits in real time. The purpose is to determine whether considerable savings are possible in regard to the computing resources, or if the results are so similar that it does not matter neither which algorithm is applied nor in which case.

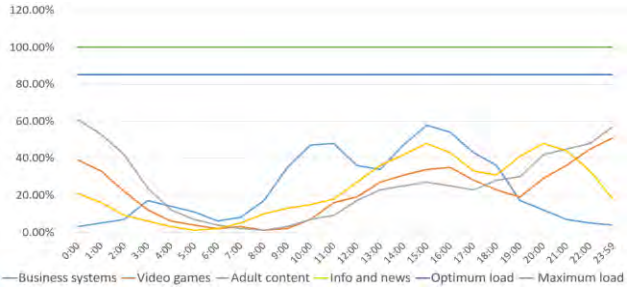


Figure 3. The system load by Web site category for 24 hours

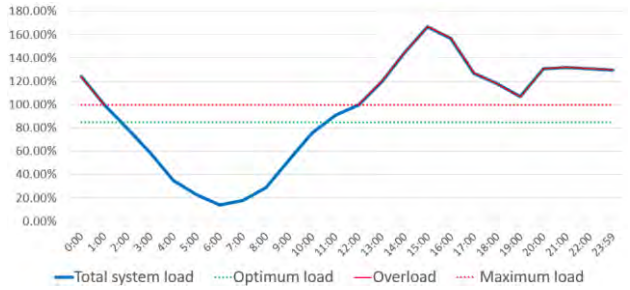


Figure 4. The total system load in all Web sites categories of for 24 hours

IV. TEST ENVIRONMENT AND CASES

The entire test environment is based on open source tools, services and operating systems allowing for reproducibility and greater availability of for further analysis and research. Given that the test environment's main objective is to test the load of Web nodes, the complete redundancy of all service was not employed: e.g. only one LB was configured, but not his passive replica. Also, in order to simplify the reproducibility of the test, classic local hard drives were used rather than a central storage, since the main focus is on the computing resources of the Web server and not on the data processing performance of the storage. However, those local disks contain a database that is configured in master - master replication, i.e. an active synchronization of data between two databases.

The environment consists of two Web clusters with three Web servers, one of which is Apache and the other nginx. Both Web services are included in order to fully test both of them, depending on the type of content (static/dynamic). As for the Website, or the content, a generic Web site for CMS (Content Management System) system was used, called WordPress (available at the following URL: https://wordpress.org/latest.zip). For the testing purposes, one Web site was installed on each Web cluster. The content was not altered so that the test could be repeatable, as well as used for further research.

As far as the infrastructure is concerned, the solution consists of three physical servers with the following virtual machines:

- Physical server host1 with virtual machines: dt-lb1 (LB), dt-db1 (first database) and dt-db2 (second database).
- Physical server host2 with virtual machines: dt-web1, dt-web2, dt-web3 (nginx Web cluster).

- Physical server host3 with virtual machines: dt-web4, dt-web5, dt-web6 (Apache Web cluster).

Server specifications:

- host1: 2x CPU, both quad-core, RAM 32GB, local hard drive 140GB.
- host2: 2x CPU, both quad-core, RAM 32GB, local hard drive 140GB.
- host3: 2x CPU, both quad-core, RAM 32GB, local hard drive 140GB.
- dt-lb1: CPU, quad-core, RAM 4GB, local hard drive 10GB.
- dt-web1: CPU, quad-core, RAM 4GB, local hard drive 10GB.
- dt-web2: CPU, quad-core, RAM 4GB, local hard drive 10GB.
- dt-web3: CPU, quad-core, RAM 4GB, local hard drive 10GB.
- dt-web4: CPU, quad-core, RAM 4GB, local hard drive 10GB.
- dt-web5: CPU, quad-core, RAM 4GB, local hard drive 10GB.
- dt-web6: CPU, quad-core, RAM 4GB, local hard drive 10GB.
- dt-db1: CPU, quad-core, RAM 4GB, local hard drive 20GB.
- dt-db2: CPU, quad-core, RAM 4GB, local hard drive 20GB.

Test environment specifications:

- Physical server: HP ProLiant DL360 G7
- OS: CentOS 6.7
- Virtualization platform: Foreman Version 1.9.1 and Libvirt (virsh) 0.10.2
- Database: MySQL mysql Ver 14.14 Distrib 5.6.12
- LB: HA-Proxy version 1.5.4 2014/09/02
- Scripting language: PHP 5.6.13
- Web servers: Apache/2.2.15 and nginx/1.8.0
- CMS (Web site): WordPress Version 4.3

Tools used for the assessment and comparison of the results:

- ApacheBench - enables traffic generation for a specific Web site
- Observium 0.15.12.7231 - monitoring of computer resources and service load
- HAProxy version 1.5.4 - monitoring of traffic, i.e. the load of Web serves in real time

Three different cases were employed for test purposes:

- Case 1 - Linear load of 100 consecutive requirements
- Case 2 - Load of 1000 requirements, 200 are done simultaneously
- Case 3 - Load of 10000 requirements, 500 are done simultaneously in the course of 10 seconds

Tested algorithms [12]:

- ***Roundrobin*** - according to this algorithm, each server is used in turns according to the weights. It works best when the processing time on the server is equally distributed. It is dynamic in nature, meaning that the server weight can be adjusted. It is limited to 4095 active servers per backend. In some particular (and very rare) cases, it can take several hundred requests for the server to be re-integrated after being down for a short time.
- ***Static-rr*** - just as with the previous algorithm, each server is used in turns according to the weights. However, this algorithm is static, meaning that if the server's weight is changed during the process, it will have no effect whatsoever. Next, this algorithm is not limited regarding the number of servers, and when the server goes up, it is always immediately reintroduced into the farm. It also uses slightly less CPU to run (approximately 1% less CPU).
- ***Leastconn*** - with this algorithm, the server with the lowest number of connections will receive the connection. Round-robin is performed within server groups working with the same load to ensure that all servers will be used. It is recommended to use this algorithm with very long sessions, such as LDAP, SQL, TSE, etc. Yet, it is not very well suited for protocols using short sessions, such as HTTP. The algorithm is dynamic, meaning that server weights may be adjusted during the process.
- ***First*** - here, the first server with available connection slots will receive the connection. Once a server reaches its maxconn value, the next server will be used. The purpose of this algorithm is to always use the smallest number of servers so that extra servers can be powered off during non-intensive hours. The server weight is not important, and the algorithm is more efficient in long session such as RDP or IMAP than HTTP, though it can be used there as well. To efficiently use this algorithm, it is recommended to check server usage regularly and turn off unused servers, as well as to regularly check backend queue to turn servers on when the queue inflates. Alternatively, using "http-check send-state" may provide information on the load.
- ***Source*** - The source IP address is hashed and divided by the total weight of servers to determine which server will receive the request, ensuring that the same client IP address will always reach the same server if the number of servers remains the same. If the hash result changes, many clients will be directed to a different server. The algorithm is generally used in TCP mode where no cookie may be inserted, but can also be used to provide a best-effort stickiness to clients refusing session cookies. It is static by default, meaning that changing a server's weight in the process will have no effect, but this can be altered by using "hash-type".
- ***Uri*** - This algorithm hashes either the left part of the URI or the whole URI and divides the hash value by the total weight of servers, determining which server will receive the request. This is used with proxy caches and anti-virus proxies in order to maximize the cache hit rate. Note that this algorithm may only be

used in an HTTP backend. It is static by default, meaning that changing a server's weight in the process will have no effect, but this can be changed using "hash-type". The algorithm supports two optional parameters - "len" and "depth", both followed by a positive integer number, which can be helpful when you need to balance the servers based on the beginning of the URI only. With the "len" parameter, the algorithm only considers that many characters at the beginning of the URI to compute the hash. The "depth" parameter indicates the maximum directory depth to be used to compute the hash.

- ***rdp-cookie*** - The RDP cookie <name> (or "mstshash" if omitted) will be looked up and hashed for each incoming TCP request. This is useful as a degraded persistence mode, always sending the same user (or session ID) to the same server. If the cookie is not found, the normal *roundrobin* algorithm is used instead. Note that the frontend must ensure that an RDP cookie is already present in the request buffer. For this you must use 'tcp-request content accept' rule combined with 'req_rdp_cookie_cnt' ACL. This algorithm is static by default, meaning that changing a server's weight in the process will have no effect, but this can be changed using "hash-type".
- ***url_param*** - The URL parameter specified in argument will be looked up in the query string of each HTTP GET request. If the modifier "check_ post" is used, an HTTP POST request entity will be searched for the parameter argument, when it is not found in a query string after a question mark. The message body will be analyzed only after the advertised amount of data has been received or if the request buffer is full. In the rare case of using chunked encoding, only the first chunk is searched. If the parameter is followed by the "=" sign and a value, the value is hashed and divided by the total weight of the running servers, designating which server will receive the request. This is used to track user identifiers in requests and ensure that a same user ID will always be sent to the same server if the number of servers remains the same. If no value or parameter is found, a round robin algorithm is applied. Note that this algorithm may only be used in an HTTP backend. It is static by default, meaning that changing a the weight in the process will have no effect, but this can be changed using "hash-type".
- *hdr* - The HTTP header <name> will be looked up in each HTTP request. Just as with the equivalent ACL 'hdr()' function, the name in parenthesis is not case sensitive. If the header is absent or does not contain any value, the *roundrobin* algorithm will be used instead. An optional 'use_domain_only' parameter is available for reducing the hash algorithm to the main domain part with some specific headers such as 'Host'. This algorithm is static by default, which means that changing a server's weight on the fly will have no effect, but this can be changed using "hash-type".

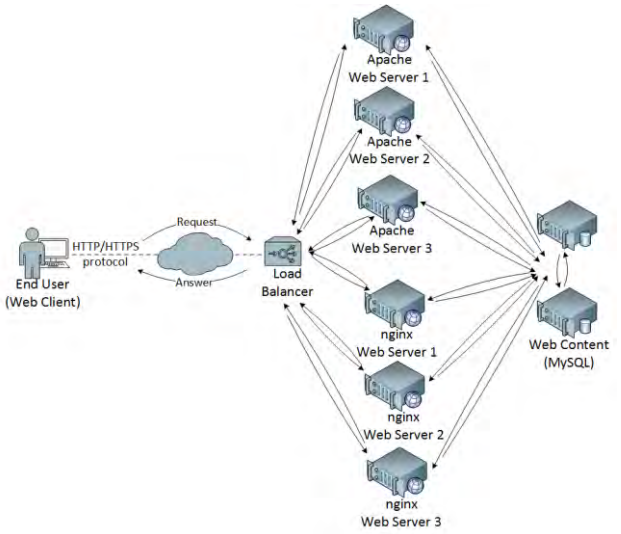


Figure 5. Test environment architecture

The algorithms ***url_param*** and ***hdr*** were left out since they are very specific and rarely used in clusters with a larger number of domains.

V. COMPARISON AND RESULTS

The testing was done Apache and nginx Web clusters for all the above mentioned cases. During testing, two main variables were monitored: the time necessary for the processing of all requirements (measured in seconds) and the system load, measured by the amount of computational work performed by a computer system.

TABLE II. APACHE WEB CLUSTER – CASE 1

Algorithm	*Time of execution*	*Average system load*
leastconn	22.448	0.5566
roundrobin	17.117	1.10333
static-rr	16.098	0.71000
first	21.604	0.07000
source	3.128	0.02666
uri	2.870	000001
rdp-cookie	19.358	0.07666

TABLE III. APACHE WEB CLUSTER – CASE 2

Algorithm	*Time of execution*	*Average system load*
leastconn	44.078	29.63667
roundrobin	34.655	23.50667
static-rr	34.802	22.73667
first	103.678	45.97333
source	13.239	1.55666
uri	8.979	2.59
rdp-cookie	33.612	21.41333

TABLE IV. APACHE WEB CLUSTER – CASE 3

Algorithm	*Number of successfully executed requests*	*Average system load*
leastconn	581	9.65
roundrobin	402	5.26
static-rr	252	23.65666
first	22	5.40334
source	1009	2.87666
uri	1139	3.11
rdp-cookie	10.037	4.24333

TABLE V. NGINX WEB CLUSTER – CASE 1

Algorithm	*Time of execution*	*Average system load*
leastconn	17.599	0.00667
roundrobin	16.301	0.07000
static-rr	16.336	0.10333
first	23.504	0.02
source	0.087	0.000001
uri	0.084	0.000001
rdp-cookie	19.844	0.00666

TABLE VI. NGINX WEB CLUSTER – CASE 2

Algorithm	*Time of execution*	*Average system load*
leastconn	6.771	0.38333
roundrobin	35.342	20.8
static-rr	33.843	20.221666
first	68.544	27.9666666
source	0.309	0.000002
uri	0.188	0.000001
rdp-cookie	35.993	22.95333

TABLE VII. NGINX WEB CLUSTER – CASE 3

Algorithm	*Number of successfully executed requests*	*Average system load*
leastconn	25272	0.61333
roundrobin	404	6.23
static-rr	351	3.53667
first	69	4.63667
source	50000	0.000001
uri	50000	0.000001
rdp-cookie	403	6.33

VI. CONCLUSION

The analysis shows that some algorithms provide much better results than others. Namely, when looking at the ratio of spent resources and the execution time for the first case of Apache cluster, the best results were obtained by the *source* and *uri* algorithms. These algorithms also provided the best results in the second case, while the worst second case result was given by the algorithm *first* (two to three times worse results). In the case of a higher load, i.e. in the third case for the Apache cluster, the best results were again obtained by algorithms *source* and *uri*. With the help of these algorithms, the largest number of successfully processed requests featuring a lesser system load was obtained in comparison with the remaining algorithms.

A similar situation was found in the nginx cluster, where the best results in all the cases were again obtained with the help of *source* and *uri* algorithms, while the algorithm *first* again gave the worst result. Other algorithms provided more or less the same results, except for the algorithm *leastconn*, which provided solid results in the third case, compared to algorithms other than *source* and *uri*.

The purpose of the testing was to identify the algorithms which provide the best results in various cases. This research singles out algorithms *source* and *uri*, since they are multiple times more successful that the rest, which is a surprising fact since algorithms such as *lastconn* and *roundrobin* are most frequently in use. It was noted not only that algorithms can provide excellent resuts, but also that they can enable multiple savings of computer resources which in turn positively reflect on the CAPEX and OPEX costs mentioned at the beginning.

REFERENCES

[1] Ali, M.F.: The Study On Load Balancing Strategies In Distributed Computing System. Int. J. Comput. Sci. Eng. Surv. 3, 2, 19–30 (2012).

[2] Aversa, L., Bestavros, a.: Load balancing a cluster of web servers: using distributed packetrewriting. Conf. Proc. 2000 IEEE Int. Performance, Comput. Commun. Conf. (Cat. No.00CH37086). (2000).

[3] Chen, W. et al.: Design and implementation of server cluster dynamic load balancing in virtualization environment based on OpenFlow. Proc. Ninth Int. Conf. Futur. Internet Technol. 691–697 (2014).

[4] Chung-Cheng Li, Kuochen Wang: An SLA-aware load balancing scheme for cloud datacenters. Int. Conf. Inf. Netw. 2014. 58–63 (2014).

[5] Gruber, C.G.: CAPEX and OPEX in Aggregation and Core Networks. 2009 Conf. Opt. Fiber Commun. - incudes post deadline Pap. 9–11 (2009).

[6] Iyer, S.: Load balancing and parallelism for the internet. July, (2008).

[7] Ji, Z., He, B.: A dynamic load balancing method for parallel rendering and physical simulation system based sort-first architecture. Proc. 2011 Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2011. 3, 1792–1796 (2011).

[8] Jung, J. et al.: Self-Adapting Load Balancing for DNS. 564–571 (2001).

[9] Khiyaita, a. et al.: Load balancing cloud computing: State of art. Netw. Secur. Syst. (JNS2), 2012 Natl. Days. 106 – 109 (2012).

[10] Knoll, T.M.: A combined CAPEX and OPEX cost model for LTE networks. (2014).

[11] Lin, Z. et al.: A content-based dynamic load-balancing algorithm for heterogeneous web server cluster. Comput. Sci. Inf. Syst. 7, 1, 153–162 (2010).

[12] Luke, S. et al.: Essentials of Metaheuristics A Set of Undergraduate Lecture Notes by BibTEX : (2011).

[13] Mcheick, H. et al.: Load Balancing Mathematical Model. 2011 Dev. E-systems Eng. 581–586 (2011).

[14] Pham, V. et al.: Gateway load balancing in future tactical networks. Proc. - IEEE Mil. Commun. Conf. MILCOM. 1844–1850 (2010).

[15] Popa, L. et al.: A Cost Comparison of Data Center Network Architectures. Proc. 6th Int. Conf. Co-NEXT '10. (2010).

[16] Rajavel, R.: De-Centralized Load Balancing for the Computational Grid Environment. Comput. Intell. 419–424 (2010).

[17] Teodoro, G. et al.: Load balancing on stateful clustered Web servers. Proceedings. 15th Symp. Comput. Archit. High Perform. Comput. (2003).

[18] Ungureanu, V. et al.: Effective load balancing for cluster-based servers employing job preemption. Perform. Eval. 65, 8, 606–622 (2008).

[19] Werstein, P. et al.: Load Balancing in a Cluster Computer.Pdf. (2006).

[20] Yong Meng Teo, Ayani, R.: Comparison of Load Balancing Strategies on Cluster-based Web Servers. Simulation. 77, 5-6, 185–195 (2001).

[21] Opex-based data centre services: Co-location, managed services and private cloud business support. 7,.

Metamodeling as an Approach for Better Computer Resources Allocation in Web Clusters

D. Alagić* and D. Maček **
* NTH Media, Varaždin, Croatia
UniCredit S.p.A. Zweigniederlassung, Wien, Austria
da@nth.ch
davor.macek@foi.hr

Abstract - Constant changes are inherent to information technology, the proof of which can be found in many challenges, such as the business sector's cloud computing. Because of the recent economic and financial crisis companies are forced to provide as best results as possible with the least amount of costs. The main issue of cloud computing are the computer resources (i.e. compute power). Due to their less than optimal usage, computer resources generate high-level costs. The term cloud computing encompasses a wide range of technologies, therefore this paper will focus only on Web clusters. It will describe the issue of improper use of resources in these systems and its main cause, as well as provide suggestions for further optimization. Additionally, the paper will present a new concept of HTTP / HTTPS traffic analysis which should enable a more efficient usage of computing resources. The concept has emerged from metamodeling of two methods. The first one is method for distribution of HTTP / HTTPS requests, and the other one is the Analytic Hierarchy Process (AHP) method which is used for classification and prioritization of requirements so that the entire Web cluster could operate as best as possible with the least amount of resources.

I. INTRODUCTION

Note: The views expressed in this article are those of the authors and do not necessarily reflect the views of the NTH Media, UniCredit S.p.A., or Zagrebačka banka d.d.

In the last decade, the concept of cloud computing has been used more frequently and we could say that it has become an important part of any modern business system as such. Cloud computing is a type of computing that relies on sharing the computing resources with another computer, starting with the applications and including a variety of related services [4][9][10].

In order for the cloud computing to work, an IT (Information Technology) infrastructure is necessary, i.e. the Data Center. It is a well-known fact that such centers are not cheap, considering their direct, indirect and general costs. For a data center to survive in today's market, continuous investments in the growth and improvement of the system are required. In other words, capital expenditures (CAPEX) as well as operating costs (OPEX) are present [8][21].

The following chart represents the total cost of the ownership of a single data center.

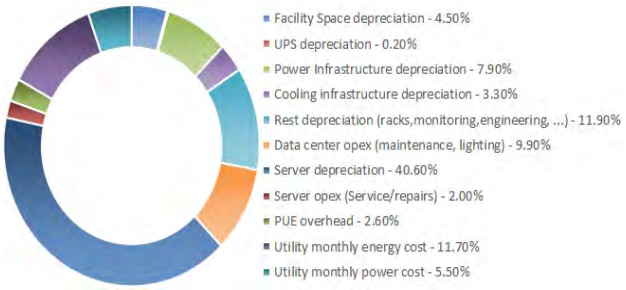


Figure 1. Total Cost of Ownership for DC [16]

An example of CAPEX cost in a single data center represents an investment in new servers or network equipment, while OPEX cost refer to the cost of maintaining the system, such as: Internet connection, electricity, air conditioning, space, maintenance and replacement of hardware, rental space, etc. [11][16].

Due to these costs, the cloud computing providers are trying to engage in multiple usage of the existing resources. In other words, they strive to achieve the automation and optimization of the system. Previous results have shown that servers represent the biggest costs because they result in high capital costs, and at the same time have a high amortization rate. They also generate high maintenance costs [16]. Therefore, the question is how to apply multiple usage to such infrastructure, i.e. to servers in order to minimize the mentioned costs. Due to this problem, the virtualization of computing resources appeared thirty years ago with which it has become possible to distribute the resources of a single physical device (in this case, the server) between several, smaller virtual environments, i.e. between a number logical or application process [15][19].

However, virtualization is now present in all data centers and it is not "sufficient" as such [18]. In other words, it is necessary to search for new ways and technologies for multiple utilization and optimization of computing resources next to virtualization, since that would result in a smaller number of servers, and ultimately reduce the consumption of electricity and other operating expenses. Thus, by employing a single variable - computer power - multiple savings are possible. To achieve it, it is necessary to define the limits, i.e. to define what kind of technology and services will be considered, since cloud computing includes a number of them and

also, they are rather specialized in terms of allocation and utilization of resources.

Cloud computing services are categorized as follows: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS) [3][5][9]. This paper mainly focuses on PaaS and SaaS services based on the HTTP / HTTPS protocol. An example of such services are Web sites where the user may or may not have a complete access to and control over the application and its contents or the data. Sets of such combined solutions, i.e. Web sites are called Web farms, whose architecture and design depend on the technology and type of content (static or dynamic). For a better understanding of the main cause of computer power consumption, we have to start from scratch, i.e. from the amount of information that must be processed in the allocated time.

II. RESEARCH PROBLEM

In the past decade, there has been a rapid increase in the use of cloud computing. To enable the existence of such infrastructure, a large number of data centers was made [6][10]. Such infrastructure is quite expensive considering that its primary goal is to allow continuous operation of all cloud computing services. In order to allow for it, the entire system must be redundant (power installations, UPS, air conditioning, network infrastructure, etc.), wherein it has to have a spare solution or system (a power generator, a BGP connection, other geographic collocation for business continuity purposes, etc.). All these items generate extremely high CAPEX and OPEX costs which represent the main motivation for improvement so as to allow savings at all levels of the infrastructure and services [16].

Furthermore, this paper pinpoints the problems as well as open issues representing the main motivation for this research:

- ***High costs of data centers*** - High operating costs, such as maintenance and electricity consumption, as well as the renting of the DC area which is usually charged by the square meter. Large capital costs at every growth systems and at the same high rate of depreciation on this type of goods. In other words, the question is: how to allow for a multiple usage of resources so as to decrease the amount of investment.
- ***System complexity and availability*** - Nowadays, all important business systems (but also all others) are located in the data center. The vast majority of these systems must not have any interruptions in their work, meaning that the information systems and infrastructure of a data center must be available at all times. In order to achieve that, information systems are becoming increasingly complex, i.e. they have an increasing number of components to be installed, configured and maintained (e.g. redundancy, backup, etc.). Thus, the question is how to simplify such systems without jeopardizing their functionality and availability. There is a number of commercial solutions, but are complex and expensive, and they also require a lot more resources and investments in infrastructure for the availability of the information systems.
- ***Computer power utilization*** - As it was already mentioned, the virtualization of resources has existed for the past few decades. However, the systems are becoming larger and more complex, so that the current solutions are no longer sufficient if we want to achieve a greater utilization of computer resources. Modern solutions for a multiple utilization of resources are mostly specialized, i.e. their application scope is limited, and thus the savings are not large-scale. In other words, the solution should be scalable and adaptable to multiple systems.

Due to all these problems, the question is how to achieve a more efficient utilization of computing resources in order to reduce the costs while simplifying the system. Today, there are several models of Web clusters. However, all of them share the system for the allocation of HTTP / HTTPS requests (hereinafter referred to as the LB system). Since the LB system contains information on the real amount of traffic which ultimately affects the computing resources, it can be assumed that it is possible to achieve the goals of this research if we improve the operation of the LB system. To achieve that, this paper analyzes two methods whose combination, or metamodel, should allow for more efficient results.

III. METHODS AND THEIR MODELS

This paper presents two methods: the first method represents the basis for the operation of the LB system, while the second method serves for the prioritization of the type of Web traffic in the cloud infrastructure.

1) Load Balancing method

The method which represents the basis of the LB system is always the same, only the operating algorithms change. The following algorithms are well-known:

- Round robin – This method for the allocation of HTTP/HTTPS traffic is based upon one of the simples and most frequently used algorithms. The entire procedure is done by allocating a time interval to each process in which it has to be done. Upon the completion of the allocated time interval, the system will stop the request and go on to the next process on the list. If the process is completed before the time runs out, the system will process another request. Having used its allocated time interval, the process goes on to the end of the list. It is important to note that the algorithm ensures that a time interval is assigned to each and every request. The main issue here is the determination of the optimal length of time. If shorter, the system assigns a larger number of requests, thus consuming more processor time while transferring requests from one process to the other. If the time interval is longer than necessary, the response for the execution of

requests is also longer, resulting in a decreased number of processed requests [12].

- Weighted Least Connections – As with the previous method, it is likewise necessary to have a knowledge of the system and the expected number of requests, since that is the basis for the determination of the maximum number of requests (connections) which can be processes by a single nod. The process is done in cycles, wherein the request first go to the node with the largest number of free connections, i.e. the largest amount of available resources [12].

The metamodel of the method for the allocation of HTTPS/HTTPS request is shown through the example of a company dealing with Web hosting. The following table shows all key entities and their attributes.

TABLE I. METAMODEL METHOD OF THE LB SYSTEM

Meta model	***Object:*** *Entity* Attribute				
Data model	**Entity:** "department" **Attribute:** "department_id" "name" "team"	**Entity:** "resource" **Attribute:** "resource_id" "web_cluster_id"	**Entity:** "result" **Attribute:** "result_id"	**Entity:** "person" **Attribute:** "person_id" "first_name" "last_name" "type_of_job"	**Entity:** "method" **Attribute:** "method_id" "lb_algorithm_id"
Appearance data	**Department:** "213" **Department name:** "ICT" **Team:** "IT"	**Resource:** "server specification" **Web cluster:** "web01-corporate"	**Result:** "url"	**Person:** "422432" **First name:** "Tom" **Last name:** "Clark" **Type of job:** "Technician"	**Method:** "Load Balancing" **Algorithm:** "Round Robin"
Real world appearance	ICT	10 CPU cores and 50 GB of Memory	http://www.company.com/contact	Tom Clark	Round Robin

Data model	**Entity:** "process" **Attribute:** "process_id"	**Entity:** knowledge **Attribute:** knowledge_id department_id	**Entity:** it_service **Attribute:** service_id	**Entity:** external_sources **Attribute:** end_user_id http_method http_request
Appearance data	**Process:** "HTTP GET request"	**Knowledge:** "131" **Department:** "213"	**Service:** "562"	**End user:** "33.176.216.1" **HTTP method:** "POST" **HTTP request:** "http://www.company.com"
Real world appearance	http://www.company.com/contact_form.asp?name1=value1&name2=value2	Web Clustering	Hosting Web Site	33.176.216.1

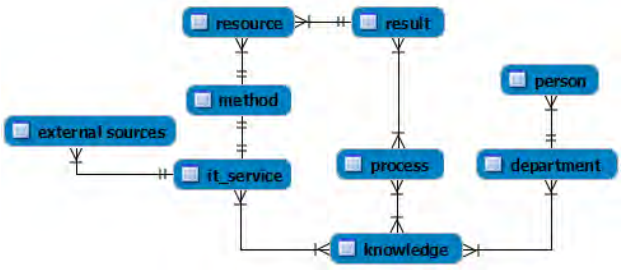


Figure 2. Business model - Load Balancing method

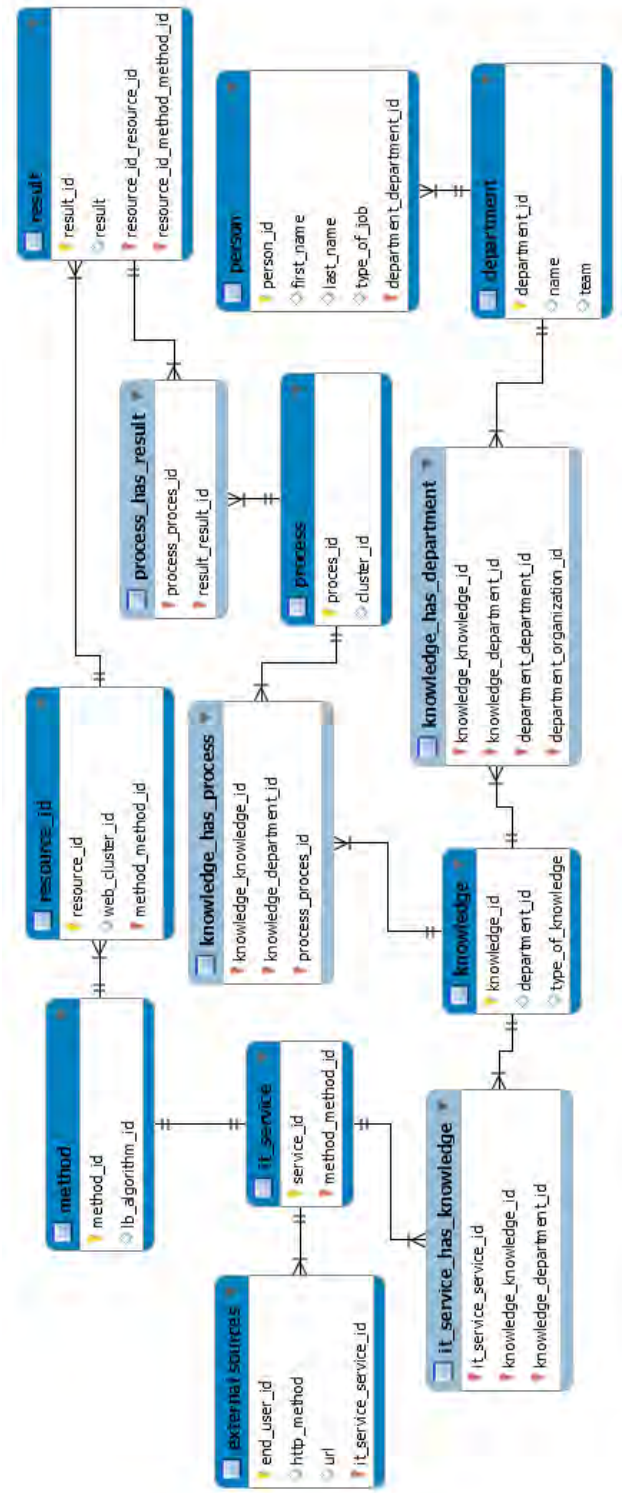


Figure 3. ER model - Load Balancing method

2) Method for the prioritization of the type of Web traffic in the cloud infrastructure

It is clear that load-balancing algorithms for the utilization of some parts of the cloud infrastructure are already developed and well-known. But there is a problem regarding the efficient prioritization of the usage of Web farm resources according to the type of Web traffic. For example, due to high CAPEX and OPEX costs, it would be quite unacceptable to allocate a Web

farm with plenty of resources (CPU, RAM) to a less critical type of HTTP or HTTPS traffic. This would present an imminent financial risk due to inadequate managing of the available resources. Operational risk would likewise be present, as well as the reputational one, if a critical type of traffic is left without enough resources. Therefore, in the context of the cloud computing and Web traffic, the aim is to choose the optimal load-balancing system depending on the type of Web traffic. This leads us to the multi-criteria decision making problem regarding the conditions of uncertainty and risk. In order to address this complex problem, it is necessary to use one of the multiple-criteria decision making (MCDM) or decision analyses (MCDA) techniques available today.

The reason for choosing a multi-criteria decision-making technique rather than a technique such as the SWOT analysis (Strengths, Weaknesses, Opportunities, and Threats) lies in its applicability. The SWOT analysis indeed provides a sufficient amount of quite descriptive data in regard to each element of the analysis. Yet, it does not include a defined criteria by which some individual system or solutions could be evaluated and/or prioritized. In this specific case, that would mean to choose an appropriate Web cluster system according to the type of Web traffic. Also, the final result of the evaluation done with the help of the MCDM technique provided some numeric/quantitative ratios among alternatives, i.e. the Web cluster nodes. This is certainly measurable (and absolutely necessary), unlike the descriptive information provided by the SWOT analysis, which does not facilitate the selection of the optimal cluster node. There are a few MCDM techniques available for complex decision making issues, but the AHP and TOPSIS techniques are among the most frequent techniques used today [2]. The reason for their extensive usage is their proven mathematical foundation.

The Analytic Hierarchy Process (AHP) is a structured technique for organizing, analyzing and making complex decisions, based on mathematics and psychology. The AHP is a multi-criteria decision-making (MCDM) approach, introduced by mathematician Thomas L. Saaty. The AHP is a decision making support tool which can be used to solve complex decision problems. It uses a multi-level hierarchical structure of objectives, criteria, sub-criteria and alternatives. Important data are derived by using a set of pairwise comparisons. These comparisons are further used to obtain the weights of importance of the decision criteria and the relative performance measures of the alternatives in terms of each individual decision criterion [7].

The Analytic Hierarchy Process is among the most widely exploited decision making techniques when the decision is based on several tangible and intangible criteria and sub-criteria. It is recognized as one of the leading theories in multicriterion decision making field. The application of the AHP technique has increased significantly in the recent years, especially in the field of IT due to its mathematical proven basis. The reason for choosing the AHP and not some other MCDM technique lies in the fact that there were already many proven researches on applying the AHP to the load-balancing issues [1][20][22]. It is true that the TOPSIS MCDM technique also has a certain number of proven applications in multi-criteria decision making problems related with load-balancing issues, but that number is still lesser when compared to the AHP technique [13][17]. Also, according to the research, the AHP has more applications in banking industry [2].

So, the AHP is an effective MCDM tool that meets the objective of decision making by ranking the choices according to their merits. Now, upon clarifying which MCDM method will be used, it is necessary to define the AHP criteria for the prioritization of types of Web traffic to choose the optimal load-balancing system. To make a set of necessary criteria for the types of Web traffic, it is also necessary to know which kind of business environment it entails. In this case, we concentrated on the banking business environment and their Web traffic. The selected criteria are influenced by business requirements and levels of criticality of individual network services that support certain business banking models. Thus, for the bank Web traffic we defined the following evaluation and prioritization criteria, in addition to the selection of the optimal cluster node: public Web, e-banking, m-banking, e-commerce and IP POS traffic. The organization uses the AHP (Analytic Hierarchy Process) technique as a method for the prioritization of Web traffic (domain) in the cloud environment. The actual goal is to prioritize the cloud Web traffic based on its type, and to select an adequate (or optimal) alternative (web cluster node) according to the defined criteria – a kind of Web traffic tagging.

The following table shows the metamodel (Table II.) of the AHP method for the prioritization of the cloud Web traffic, according to which the business model was designed (Figure 4.). From the defined business model shown in Figure 4, the ERA (Entity-Relationship-Attribute) model was derived with all of its tables, attributes and necessary relationships (Figure 5.).

TABLE II. METAMODEL OF THE AHP METHOD FOR WEB TRAFFIC PRIORITIZATION

Meta model	***Object:*** *Entity* *Attribute*					
Data model	**Entity:** "organization" **Attribute:** "organization_id" "name" "address"	**Entity:** "method" **Attribute:** "method_id" "name"	**Entity:** "domain" **Attribute:** "domain_id" "description"	**Entity:** "goal" **Atribut:** "goal_id" "description"	**Entity:** "criteria" **Atribut:** "criteria_id" "name"	**Entity:** "alternative" **Atribut:** "alternative_id" "name"
Appearance data	**Organization:** "213" **Name:** "CloudOcean" **Address:** "30 Algonquin, New York, NY, USA"	**Method:** "7" **Name:** "Analytic Hierarchy Process (AHP)"	**Domain:** "Web Site" **Description:** "url"	**Goal:** "58" **Description:** "Web Traffic Prioritization"	**Criteria:** "14" **Name:** "e-banking"	**Alternative:** "22" **Name:** "Web Cluster web01-corporate"
Real world appearance	CloudOcean	Analytic Hierarchy Process (AHP)	http://www.company.com	Web Traffic Prioritization	e-banking	Web Cluster web01-corporate



Figure 4. Business model for the AHP method for Web traffic prioritization

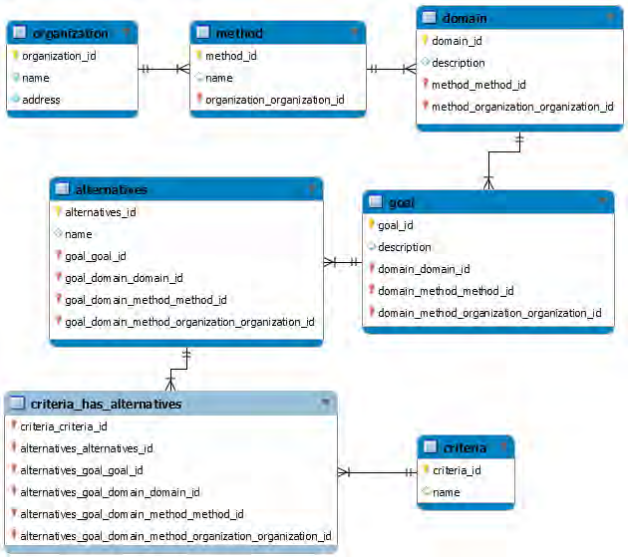


Figure 5. ER model for the AHP method for Web traffic prioritization

IV. COMMON METAMODEL METHODOLOGY

The Load Balancing (LB) method and the AHP method for Web traffic prioritization have only two common points – the *method* entities and Web traffic type (*IT_service* in LB_method and its corresponding table *Domain* in the AHP_method). However, these two common points are actually crucial for the existence of this new hybrid metamodel. By merging the LB method and the AHP method, we succeeded in developing a new hybrid metamodel capable of prioritizing Web traffic according to its type. The metamodel can also help load balancers to direct the traffic to the appropriate Web cluster node, ultimately resulting in a better resource allocation and reduction of costs which represent the ultimate goals of this research.

V. EXPECTED CONTRIBUTIONS

The purpose of this research is to propose a concept model combining two methods, the implementation of which would entail multiple social and business justifications for the following reasons:

- ***Wide availability*** - today there are many commercial solutions that are very expensive and demanding at the same time, i.e. it takes a lot of compute power for their proper operation. The purpose of this paper is to use the open source technology and solutions so that everyone can benefit from them, not only ICT businesses, but also the scientific community for research purposes. The new algorithm will also be publicly available to everyone.
- ***Improved utilization of resources*** - one of the main objectives is: "How to achieve multiple usage of the existing compute power"? As noted above, the virtualization of computing resources is present almost everywhere however, for it requires the so-called bare bone (physical resources) whose price range and amortization are quite high the new algorithm should orchestrate Web servers and their free resources on its own and allocate them where necessary. This is a drastic change compared to the commercial solutions which solve the lack of resources by configuring and delivering new instance of Web server, which ultimately consumes even

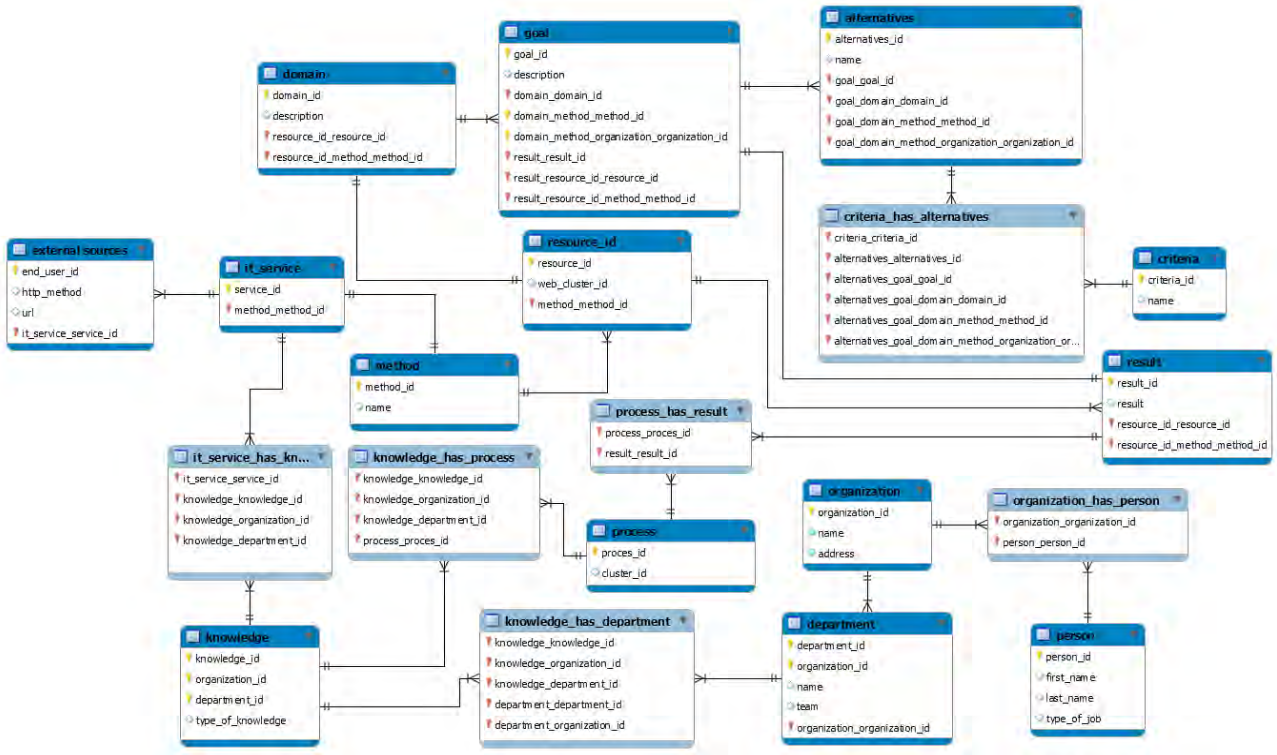


Figure 6. ER model for the common metamodel

more compute power.

- ***Increased system availability*** - since compute power can be better utilized, this means that higher scalability and system availability are also possible.
- ***Wide use*** - the goal of this solution represents its broader application and the fact that the new algorithm can be applied to all types of Web services (Nginx and Apache), but also for both types of content, that is the static and the dynamic.
- ***Increased financial profitability*** (lower OPEX) - the main requirement for any cloud computing is greater availability. In order to achieve this, a larger and more complex infrastructure is used that is expensive to maintain. The new algorithm should allow multiple utilization of computing resources that provides even greater scalability with a lesser server density, i.e. the necessary infrastructure. Since the computing resources can be better utilized, we can assume that the new algorithm should reduce the number of physical servers which would result in savings because it would allow lower maintenance costs (less equipment and network infrastructure), a smaller number of equipment (cost of maintaining and replacing hardware), lower monthly expenses of the data center (electricity and air conditioning) and the like.
- ***Environmental protection*** - due to the already mentioned high costs of a single data center, very few of them are in compliance with the environmental regulations. In other words, the electricity they use for their infrastructure does not come from renewable energy sources, therefore the aim is to reduce their electricity consumption. The new algorithm should reduce the number of servers due to the possibility of reallocating computing resources which would ultimately result in lower energy consumption.

VI. CONLUSION

In the last decade, there has been a rapid increase regarding the usage of the cloud computing. In order for a such infrastructure to exist, a large number of data centers has been made [6][10]. Such infrastructure is quite expensive since its primary goal is to enable a continuous operation of all cloud computing services. To achieve that, the entire system must be redundant (power installations, UPS, air conditioning, network infrastructure, etc.), with a secured backup system (a power generator, a BGP connection, other geographic collocation, etc.). All these items generate extremely high CAPEX and OPEX costs which is the main motivation for improvement so as to allow for savings at all levels of the infrastructure and services [16].

Consequently, this paper proposes a new model created through a combination of two methods, which could lead to achieving the set goals along with further upgrading and improvement. For future research, we propose the development of some scripts (e.g. Python scripts) in the first phase. In the second phase, we propose the development of the entire software solution in a certain programming language to support this new metamodel. Also, we propose to investigate as to whether is possible to use some other multi-criteria decision making tool like TOPSIS, instead of the AHP. In case of success, that would be additional forte for this hybrid metamodel because of its modularity and scalability.

REFERENCES

[1] Aghazarian, V.: RQSG-I : An Optimized Real time Scheduling Algorithm for Tasks Allocation in Grid Environments. 205–210 (2011).

[2] Aruldoss, M.: A Survey on Multi Criteria Decision Making Methods and Its Applications. Am. J. Inf. Syst. 1, 1, 31–43 (2013).

[3] Celesti, A. et al.: An approach to enable cloud service providers to arrange IaaS, PaaS, and SaaS using external virtualization infrastructures. Proc. - 2011 IEEE World Congr. Serv. Serv. 2011. 607–611 (2011).

[4] Chung-Cheng Li, Kuochen Wang: An SLA-aware load balancing scheme for cloud datacenters. Int. Conf. Inf. Netw. 2014. 58–63 (2014).

[5] Detal, G. et al.: Multipath in the middle (box). Work. Hot Top. middleboxes Netw. Funct. virtualization. 1–6 (2013).

[6] Dixon, D., Basiliere, P.: Cloud Computing for the Sciences. Anal. June, 1–6 (2009).

[7] Engineering, I. et al.: Using the Analytic Hierarchy Process for Decision Making in Engineering Applications : Some Challenges. Int. J. Ind. Eng. Theory, Appl. Pract. 2, 1, 35–44 (1995).

[8] Gruber, C.G.: CAPEX and OPEX in Aggregation and Core Networks. 2009 Conf. Opt. Fiber Commun. - incudes post deadline Pap. 9–11 (2009).

[9] Islam, S.S. et al.: Cloud computing for future generation of computing technology. 2012 IEEE Int. Conf. Cyber Technol. Autom. Control. Intell. Syst. 129–134 (2012).

[10] Jadeja, Y., Modi, K.: Cloud computing - Concepts, architecture and challenges. 2012 Int. Conf. Comput. Electron. Electr. Technol. ICCEET 2012. 877–880 (2012).

[11] Knoll, T.M.: A combined CAPEX and OPEX cost model for LTE networks. (2014).

[12] Luke, S. et al.: Essentials of Metaheuristics A Set of Undergraduate Lecture Notes by BibTEX : (2011).

[13] Ma, F. et al.: Distributed load balancing allocation of virtual machine in cloud data center. Softw. Eng. Serv. Sci. (ICSESS), 2012 IEEE 3rd Int. Conf. 20 – 23 (2012).

[14] Popa, L. et al.: A Cost Comparison of Data Center Network Architectures. Proc. 6th Int. Conf. Co-NEXT '10. (2010).

[15] Rubenstein, B., Faist, M.: Data Center Cold Aisle Set Point Optimization through Total Operating Cost Modeling.

[16] Sampson, J., Tullsen, D.M.: Battery Provisioning and Associated Costs for Data Center Power Capping. (2012).

[17] Sayari, Z., Harounabadi, A.: Evaluation and Making a Tradeoff between Load Balancing and Reliability in Grid Services using Formal Models. 127, 9, 5–10 (2015).

[18] Singh, A. et al.: Server-storage virtualization: Integration and load balancing in data centers. 2008 SC - Int. Conf. High Perform. Comput. Networking, Storage Anal. SC 2008. November, (2008).

[19] Soundararajan, V., Herndon, B.: Benchmarking a Virtualization Platform. 99–109 (2014).

[20] Wang, M. et al.: Analytic hierarchy process in load balancing for the multicast and unicast mixed services in LTE. 2012 IEEE Wirel. Commun. Netw. Conf. 2735–2740 (2012).

[21] Wiboonrat, M.: Life Cycle Cost Analysis of Data Center Project. (2014).

[22] Zhang, J.-D. et al.: Load Balancing Based on Group Analytic Hierarchy Process. 2013 Ninth Int. Conf. Comput. Intell. Secur. 758–762 (2013).

Showers Prediction by WRF Model above Complex Terrain

T. Davitashvili *, N. Kutaladze **, R. Kvatadze ***, G. Mikuchadze **, Z. Modebadze *, and I. Samkharadze*

* Iv. Javakhishvili Tbilisi State University/ I.Vekua Institute of Applied mathematics, Tbilisi, Georgia

** Georgian Hydro-meteorology Department, Tbilisi, Georgia

*** Georgian Research and Educational Networks Association, Tbilisi, Georgia

**** Iv. Javakhishvili Tbilisi State University/ Exact and Natural Sciences Faculty, Tbilisi, Georgia

tedavitashvili@gmail.com

Abstract - In the present article we have configured the nested grid WRF v.3.6 model for the Caucasus region, taking into consideration geographical-landscape character, topography heights, land use, soil type, temperature in deep layers, vegetation monthly distribution, albedo and others. Computations were performed using Grid system GE-01-GRENA with working nodes (16 cores+, 32GB RAM on each) located at the Georgian Research and Educational Networking Association (GRENA) which had been included in the European Grid infrastructure. Therefore it was a good opportunity for running model on larger number of CPUs and storing large amount of data on the Grid storage elements. Two particulate cases of unexpected heavy showers which took place on 13th of June 2015 in Tbilisi and 20th of August 2015 in Kakheti (eastern Georgia) were studied. Simulations were performed by two set of domains with horizontal grid-point resolutions of 6.6 km and 2.2 km. The ability of the WRF model in prediction precipitations with different microphysics and convective scheme components taking into consideration complex terrain of the Georgian territory was tested. Some results of the numerical calculations performed by WRF model are presented.

I. INTRODUCTION

Regional climate formation above the territory of complex terrains is conditioned due to joint actions of large-scale synoptic and local atmospheric processes where the last one is basically stipulated by complex topography structure of the terrain. The territory of Caucasus and especially territory of Georgia are good examples for that. Indeed compound topographic sections (85% of the total land area of Georgia are mountains) play an impotent role for spatial-temporal distribution of meteorological fields. As known the global weather prediction models can well characterize the large scale atmospheric systems, but not enough the mesoscale processes which mainly associated with regional complex terrain and land cover. For modeling these smaller scale atmospheric processes and its characterizing features it is necessary to take into consideration the main features of the local terrain, its heterogeneous surfaces and influence of large scale atmosphere processes on the local scale processes. The Weather Research and Forecasting (WRF) models are widely used by many operational services for short and medium range weather forecasting [1]. The WRF model version 3.6 represents a suitable mean for studding regional and mesoscale atmospheric processes such are: Regional Climate, Extreme Precipitations, Hails, influence of orography on mesoscale atmosphere processes, sensitivity of WRF to physics options etc. [1]. As a matter of fact the Advanced Research Weather Research and Forecasting Model (WRF-ARW) is a convenient research tool, as it offers multiple physics options and they can be combined in many ways [2]-[9]. Indeed in WRF-ARW model the main categories of physics parameterizations (microphysics, cumulus parameterizations, surface physics, planetary boundary layer physics and atmospheric radiation physics) mutually connected via the model state variables (potential temperature, moisture, wind, etc.) by their tendencies and via the surface fluxes [1]-[5]. Taking into account this broad availability of parameterizations it is not easy to define the right combination that better describes a meteorological phenomenon dominated above the investigated region. Many works have been dedicated to the problem of identification the best combination of parameterizations in WRF model that better represents the atmospheric conditions above the investigated region [2]-[26]. Three main combination of the microphysics parameterization schemes (WRF Single-Moment 3-class (WSM3) scheme[11], Eta Ferrier scheme [12], Purdue Lin scheme [13]) and 3 cumulus schemes (Kain-Fritsch [7], Betts-Miller-Janjic [14]-[15], Grell-Devenyi ensemble [16]) were chosen for identification which combination in WRF model was better simulated the atmospheric lightning conditions in the Brazil southeastern [5]. Analysis of numerical calculations in [5] had shown that sets of WSM3 with Kain-Fritsch schemes and Eta Ferrier with Betts-Miller-Janjic schemes, were better represented temperature and wind at surface, low and medium levels in the atmosphere, than Purdue Lin with Grell-Devenyi ensemble schemes. Although the sets of the WSM3 with Kain-Fritsch schemes and Eta Ferrier with Betts-Miller-Janjic schemes, have been highlighted again among all the combinations and the set WSM3 with

Kain-Fritsch schemes was the one which presented the best results for the meteorological variables [5]. Using WRF model six different convective parameterization schemes (CP's) were investigated for studying the impact CP's on the quality of rainfall forecast over Tanzania for two rainy seasons [2]. The results of numerical experiments had showed that for extreme rainfall prediction the Betts-Miller-Janjic (BMJ) and Ensemble Grell-Devenyi (GD) Schemes gave better results than any other CP's for different regions and seasons in Tanzania. Also WRF model to some extent performs better in the cases of extreme rainfall [2]. Eight microphysics schemes (Lin, WSM5, Eta, WSM6, Goddard, Thompson, WDM5, WDM6) and three DA techniques were examined in the WRF-ARW for reproduction observed strong convection and 3 heavy precipitations over the US SGP for events of 27_31 May 2001 [3]. To reproduce vertical structure and evolution of the warm-season convective events (27_31 May 2001) high temporal resolution data and cloud measurements millimetre cloud radar (MMCR) were used. From all simulation results an important deduction was that quite accurate reproductions of lower tropospheric moisture, temperature and wind profiles were necessary for the successful application, implementation of cloud-resolving regional models to simulate deep convection [3]. In [4] results of 24 h. predictions by WRF model had shown that the Kain-Fritsch BMJ and Grell schemes were not in satisfactory quality and the simplified Arakawa cumulus parameterization scheme with Lin et al. micro-physics scheme, RUC Land surface model, Asymmetric Convective Model (ACM2), planetary boundary layer physics gave a better result than others for produce reasonable predictions in the Kelani River basin in Sri Lanka [4]. The sensitivity of quantitative precipitation forecasts to various modifications of the KF scheme and determination at which grid spacing values the KF scheme may no longer be needed on simulated precipitation was studied in [17]. By the way the Kain-Fritsch scheme [7],[18],[19] is frequently used to improve forecasts for convective parameterization at grid spacing below 20 km, likely because it has been shown that KF scheme perform better convective parameterization than other CPSs such as the Betts-Miller-Janjic and Grell-Devenyi schemes [8], [9], [10]. Also the KF scheme outperformed others for the 4 km simulation that used no convective scheme, so KF scheme can be used to improve forecasts even at such high resolutions [8],[17],[20]. Atmosphere processes and parameterization of physics for the Caucasus territory have been tested by the WRF model using the following schemes: WSM 3-class simple ice scheme, RRTM scheme, Dudhia scheme, unified Noah and-surface model, Yonsei University scheme, Kain-Fritsch (new Eta) scheme and Noah land-surface model scheme [25], [26]. In this study, WRF is using for prediction heavy showers and hails for different set of physical options over the regions characterized with the complex topography. Mesoscale Convective Systems (MCS) have been studied using real data and WRF simulations based on grid spacing in the range from 2.2 km to 19.8 km with an emphasis on 2.2 km. The ability of the WRF model in prediction precipitations with different microphysics and convective scheme components taking into consideration complex terrain of the Georgian territory have been tested.

EU Commission Project H2020 VI-SEEM №675121

II. DATA AND METHODOLOGY

A. Observational data

Hydro-meteorological observations (HMO) and research in Georgia began in 1887 and in 1974 it was performed in 33 HMO stations allocated in 11 towns of Georgia. Since 1992 in the system of HMO arouse some problems due to political and economical processes developed on the territory of Georgia. Namely there was decreased a number of HMO, functioning stations were out of modern devices. At present air quality monitoring performs by National Agency of Environment and under his jurisdiction are 7 observation stations distributed in the 5 cities of Georgia: Tbilisi, Rustavi (eastern Georgia) Kutaisi, Zestafoni and Batumi (western Georgia). Each city has only 1 or 2 observation stations and only exception is capital city of Georgia Tbilisi were for the last ten-year period the observation were carried out in 8 posts, located in different districts of Tbilisi. It is obvious that these numbers of stations are not enough for assessment of hydro-atmosphere statement over the territory of Georgia. In fact we have hydrometeorological information only for separated areas were the stations are located. We have obtained and analyzed only scant information on air temperature, air humidity precipitation amount, wind (speed, direction), of 13-14 June and 20-21 August 2015 in Tbilisi. All the above data were obtained from Hydro-meteorology Department of Georgia and from the meteorological post of Tbilisi State University. Also we have analyzed radar's information on clouds structure located in Kakheti region and these data have been used for assessment WRF model results.

B. Observed convective events during 13-14 June 2015

Weather on the night of 13 to 14 June 2015 in Tbilisi was terrible with showers, thunderstorms and lights. According to the Department of Hydrometeorology of Georgia the maximal temperature on 13th of June reached 29^0 C. There were transfer of heat from the south by wave and it stipulated high temperature showers with thunderstorms and lights in Tbilisi. A shocking accident took place on the night of 13 to 14 June 2015 in Tbilisi. Namely late on 13 June 2015 during 1.5-2 hours there was heavy shower and following of the heavy rainfall, a landslide was released above the village of Akhaldaba, about 20 km southwest of Tbilisi. The collapsed 1 million m^3 of land, mud, rocks and trees moved down from the Akhaldaba mountain into Tbilisi and dammed up the Vere river (the Vere river is flows from the Akhaldaba mountain through the territory of Tbilisi's zoo and further discharge into the Kura river by a tunnel under square of Heroes). A big wave (constructed by mass of slush, rocks and trees) run across the Vere canyon and washed everything away until the square of Heroes. The resulting flood inflicted severe damage on the Tbilisi Zoo, Heroes'

Square and nearby streets and houses (see Fig.1). Unfortunately this process has been resulted in at least 20 deaths, including three zoo workers and leaving half of the Tbilisi Zoo's animal inhabitants either dead or on the loose.



Figure 1. After flood on 14 June 2015 inTbilisi

C. *Observed convective events 20-21 August 2015*

It was dominated western atmospheric processes above the territory of Georgia from 19 to 21 August 2015. There were developed inner massive processes above the territory of Tbilisi and it was hailed in the evening of 19th August 2015. In the evening of 20th August 2015 it was raining with thunderstorms and lights in Tbilisi. The maximal temperature on 20th of August reached 36^0 C and on 21st August it reached 31^0 C. Also on 20th of August 2015 a heavy rainfall was observed above the Kakheti region (Kakheti is famous wine-making region in eastern Georgia) of Georgia. Downpours with hail cause destruction to some regions of Kakheti and resort suburbs of Tbilisi Kojori and Kiketi, where ground floors of many houses were flooded in the evening of 20th of August 2015. Namely, caused by the violent weather the rain with hail lasted for half an hour and in some settlements of the Gurjaani, Lagodekhi and Kvareli districts broke roofs and even walls of houses. As a result, 50 percent of crops have been destroyed and many fruit trees in orchards were damaged (Source: http://eng.kavkaz-uzel.ru/articles/21952/). The latest largest hail storm in Kakheti was recorded in July 2012 when 100 percent of all crops were destroyed in several villages. Furthermore, hundreds of houses were left without roofs and many cattle died. Hail has always been a major issue for people in Kakheti. Every year a large portion of the agricultural sector, particularly grapes, were damaged by hail, leaving farmers with no crops. It became necessary to put into practice some implementation for solution to the hail problem. More than 80 anti-hail firing points and equipments (radars) have been installed in Kakheti for reduction the damage to crops.

According to radar's allocated in the Kakheti region we have obtained the following information concerning on cloud characteristics progression in the atmospheric column over the investigated region. Namely at 19:00 o'clock of 20th August from south-west of radar system there was outbreak of cloud systems having atmospheric front appearance (looks), which was moving towards to north-east direction, which from 19:42 began weakening. At 19:20 a new clod systems were formed and began moving from north-west to the town Akhmeta and 19:49 it achieved Akhmeta and the atmospheric column over the region has 15km height with maximal reflection 60dB. At 20:18 the cloud system with height 15km and maximal reflection 60dB from the territory of Akhmeta continued moving towards south-east direction and at 21:13 it reached territory of Kvareli with height 16km and maximal reflection 60dB. The cloud system continued migration and at 23:07 it shifted to the Lagodekhi territory the height and reflection of the cloud system began depletion. At 01:59 were formed a new clod system and it began moving from north-west to the north-east direction and at 02:19 it has 10km height, maximal reflection 50dB at 02:42 it achieved Akhmeta territory continued moving toward to the north-east direction and leaved investigated region.

D. *WRF model simulation design*

The Advanced Research WRF ARW model version 3.6 was used to simulate the warm-season heavy precipitation cases of the 13-14 June 2015 and 19-21 August 2015. Below some details of the numerical schemes are present. In our study we have used one-way nested domains centered on the territory of Georgia. Namely simulations were performed using a set of 2 domains with horizontal grid-point resolutions of 6.6km and 2.2 km, both defined as those currently being used for operational forecasts. The coarser domain has a grid of 94x102 points which covers the South Caucasus region, while the nested inner domain has a grid size of 70x70 points mainly territory of Georgia. Both use the 54 vertical levels including 8 levels below 2 km. A time step of 10 seconds was used for the nested domain. The WRF model contains a number of different physics options such as micro physics, cumulus parameterization physics, radiation physics, surface layer physics, land surface physics, and planetary boundary layer physics. Microphysics firmly determine water vapor, cloud, and precipitation processes [1]. There are number of microphysics such as Kessler scheme, Purdue Lin scheme, WRF Single-Moment3-class (WSM3) scheme, WSM5 scheme, WSM6 scheme, Eta Grid-scale cloud and precipitation scheme, Thompson et al. scheme, Goddard cumulus ensemble model scheme and Morrison et al Moment scheme. In our study we have chosen WSM6, Thompson, Purdue Lin, Morrison 2 Moment and Goddard schemes. Cumulus parameterization schemes are responsible for the sub-grid-scale effects of convective and/or shallow clouds and theoretically valid only for coarser grid sizes [1]. The types of cumulus parameterization schemes are Kain-Fritsch scheme, Betts – Miller - Janjic scheme, Grell – Devenyi ensemble scheme, Grell-3d ensemble scheme and Simplied Arakawa scheme. We have chosen Kain-Fritsch, Betts – Miller - Janjic and Grell – Devenyi ensemble schemes for our experiments.The planetary boundary layer (PBL) is responsible for vertical sub-grid-scale fluxes due to eddytransports in the whole atmospheric column [1]. Parameterization of the PBL directly influences on vertical wind shear, as well as precipitation evolution [21],[22]. In [23] summarized the main characteristics that explain the differences among WRF PBL schemes

and also there was investigated how the PBL evolves within the ARW using 4-km grid spacing. There are number of PBL schemes such as Yonsei University scheme, Mellor-Yamada-Janjic scheme, MRF scheme, Asymmetric Convective Model, Quasi-Normal Scale Elimination and Mellor-Yamada Nakanishi and Niino schemes are the varieties. According to [23] we have mainly chosen Yonsei University scheme. The land-surface models use atmospheric information from the surface layer scheme, radiative forcing from the radiation scheme, and precipitation forcing from the microphysics and convective schemes, together with internal information on the land's state variables and landsurface properties, to provide heat and moisture fluxes over land points and sea-ice points[2]. Five-layer thermal diffusion model, Noah Land Surface Model, RUC Land Surface Model and Pleim-Xiu Land Surface Model are the species of land-surface models. We have chosen Noah Land Surface Model. After considering various combinations of microphysics, Cumulus parameterization schemes, Land surface-physics and planetary boundary layer physics its combination for our experiments are given in the Table 1.

TABLE 1. Five set of the WRF parameterizations used in this study.

WRF Physics	Set1	Set 2	Set 3	Set 4	Set 5
Micro physics	WSM 6	Thom pson	Purdu e Lin	Morrison 2-Moment	Goddard
Cumulus Paramet erization	Kain-Fritsc h	Betts-Miller Janjic	Kain-Fritsc h	Grell-Devenyi ensemble	Kain-Fritsch
Surface Layer	MM5 Simil.	MM5 Simil.	MM5 Simil	(PX) Similarity	MM5 Similarit
Planet. Boundar y Layer	YSU PBL	YSU PBL	YSU PBL	ACM2 PBL	YSU PBL
Land-Surface	Noah LSM	Noah LSM	Noah LSM	Noah LSM	Noah LSM
Atmosph eric Radiat.	RRT M/Du dhia	RRT M/Du dhia	RRT M/Du dhia	RRTM/D udhia	RRTM/ Dudhia

III. RESULTS AND DISCUSSION

Simulation of 2 atmospheric accidental events occurring during June and August 2015 over the territory of Georgia were performed by WRF-ARW model on the basis of the five different combinations of different physics options such as micro physics, cumulus parameterization physics, radiation physics, surface layer physics, land surface physics, and planetary boundary layer physics represented in the Table 1. Results of numerical calculation showed that not one of the combinations listed in the Table 1 were able to model true atmospheric event which took place on the 13th of June 2015. Namely results of numerical calculations showed that 24h predictions by these schemes were not in satisfactory quality as they were not able to account of the small-scale processes that lead to the development of deep convection. For example on the Fig.2 and Fig.3 are presented predicted fields of the relative humidity on the 850 hPa for 13 June (21UTC) and 14 June (00UTC) 2015, respectively, which were simulated by WRF Physics Options set1 (it gave a better result than others). The calculated amounts of water vapor presented on the Fig.2 and Fig.3 (nested domain with 6.6 km resolution) at the accidental moments when atmospheric event were in full swing are not in satisfactory agreement with real situation which took place in Tbilisi and surroundings on 13 June 2015.

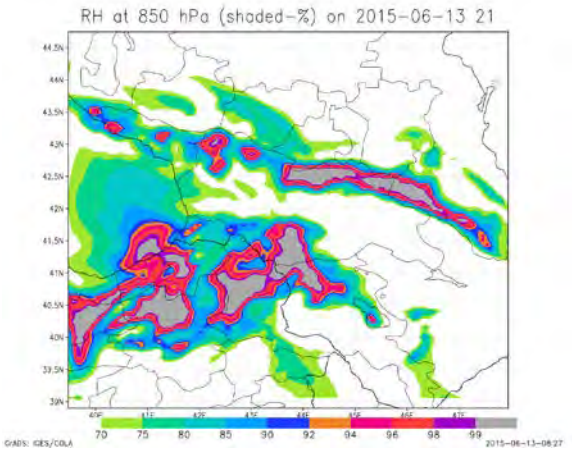


Figure.2 Map of the relative humidity at the 850 hPa for 13 June 2015 (21UTC) simulated for the nested domain with 6.6 km resolution.

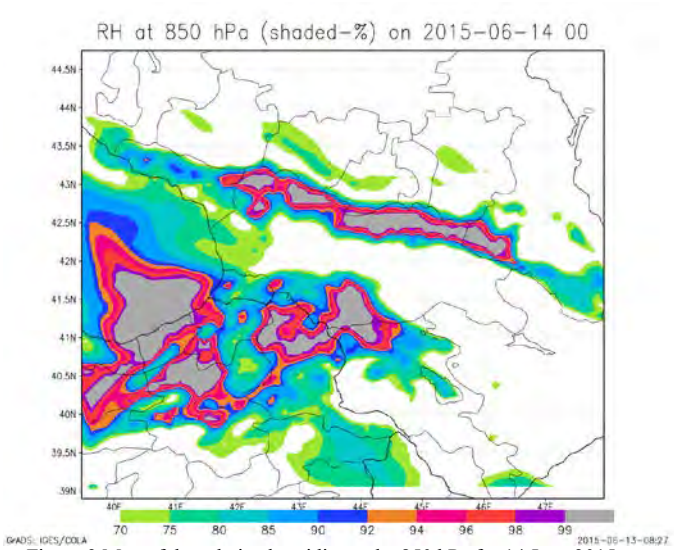


Figure.3 Map of the relative humidity at the 850 hPa for 14 June 2015 (00UTC) simulated for the nested domain with 6.6 km resolution.

On the figures 4 and 5 are presented forecasted precipitation fields on the 850 hPa height for 13 June 2015 (21UTC) and for 14 June 2015 (00UTC) respectively. Both of the figures demonstrate 24 h WRF-ARW forecast failure and especially in the investigated region where the both nested models predicted almost dray conditions (insignificant precipitations). Namely comparison Fig.4 with Fig.5 shows that considerably increased amount of accumulated precipitations at the coastal area of the Black Sea nearby of the Poti city, but there is diffuse spectrum of accumulated precipitations on

the Fig.5 in comparison with Fig.4 in the investigated area. Unfortunately for this case study, all of the precipitation simulated in the region of interest (Tbilisi and suborbs) was not convective in nature, only small amount of precipitation was produced by the model. As it is known the CPSs are producing precipitation, not the microphysics [8], so this indicates that the set of choices of CPSs were not producing precipitation for this mesoscale case study. An important deduction from all simulation results was that quite accurate reproductions of lower tropospheric temperature and wind profiles but these were not necessary for the successful simulation mesoscale deep convection which took place on 13 June 2015 in Tbilisi and suburbs. In our opinion, it is necessary to strengthen initial and boundary conditions through data assimilation, and to improve the physical linkages between the radiation physics, surface layer physics, land surface physics.

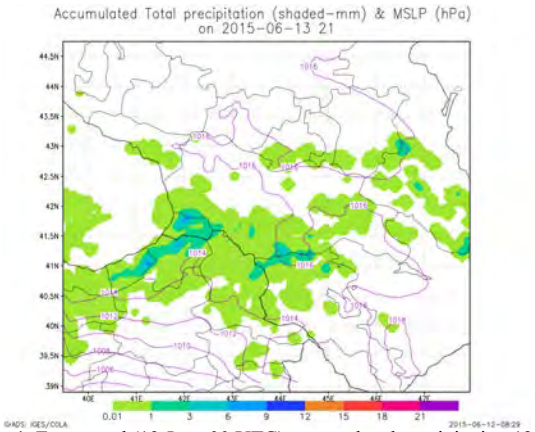


Fig.4. Forecasted (13 June 00 UTC) accumulated precipitation 12 h sum simulated for the nested domain with 6.6 km resolution.

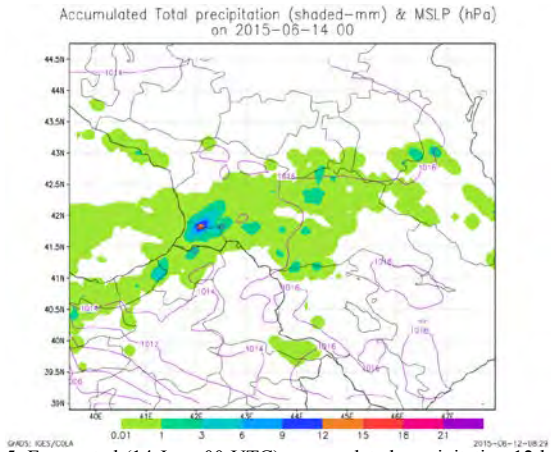


Fig.5. Forecasted (14 June 00 UTC) accumulated precipitation 12 h sum for nested domain with 1-way nesting method and 6.6 km resolution.

Also the WRF–ARW model, version 3.6, was used for simulation shower which took place on 20 August with 5 different combinations of physical schemes (see Table.1). Numerical calculations have shown that in all cases, orographic forcing plays an important role in the localization and intensification of precipitation in and nearby of complex terrain. In the pre-convective environment, caused by differential heating of valleys and mountain slopes, pronounced small-scale secondary circulations develop during the day. They lead to the development of convergence lines, e.g., over mountain crests, and mesoscale vortices influencing atmospheric boundary layer evolution as well as the transport of humidity and other tracers. Numerical calculations have shown that combination of the Purdue Lin scheme with Kain-Fritsch scheme and MM5 Similarity Surface Layer (Set 3) and Goddard scheme with Kain-Fritsch scheme and MM5 Similarity Surface Layer scheme (Set 5) gave the better results than others. Selected convective cases for Set3 and Sey5are shown in Fig.6 and Fig.7 respectively. Numerical calculations have shown that there are indeed 'natural' scales of activity for the convective parameterization within WRF.

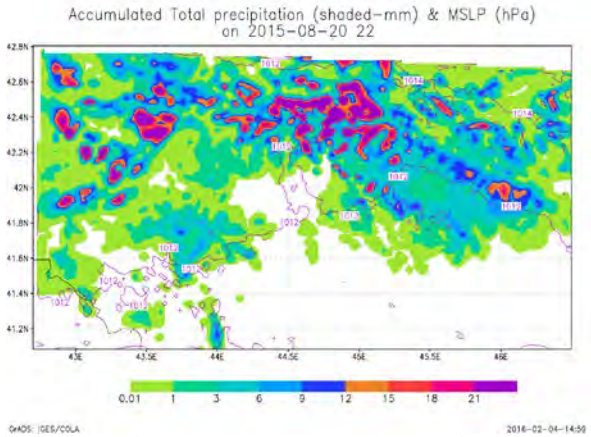


Fig.6. Forecasted (Set-3, 20 August 21 UTC) accumulated precipitation 12 h sum for nested domain 2.2 km resolution.

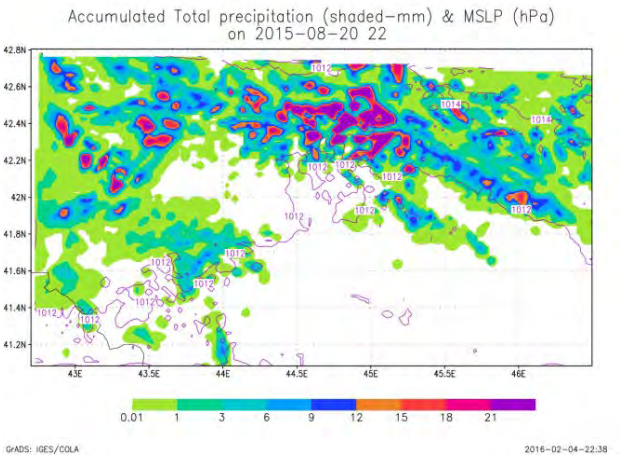


Fig.7. Forecasted (Set-5, 20 August 21 UTC) accumulated precipitation 12 h sum for nested domain 2.2 km resolution.

Comparison Fig.6 with Fig.7 shows that the main features of accumulated precipitations are predicted almost similarly, but accurate study of the dynamics and its comparison with the data of observations have shown that Set 3 was able to model that true atmospheric event which took place on the 20 -21 August 2015. In summary it can be said, that above mentioned model can be

successfully used for local weather extremes prediction for western type synoptic processes such was 19-21 August atmospheric circulation above the Georgian territory.

IV. CONLUSION

In this study some comparisons between WRF forecasts was done in order to check the consistency and quality of WRF model for the day with the heavy precipitations occur on the territory of Georgia. This first analysis allowed verifying that in general the set of combinations of Purdue Lin scheme with Kain-Fritsch scheme and MM5 Similarity Surface Layer (Set 3) and Goddard scheme with Kain-Fritsch scheme and MM5 Similarity Surface Layer scheme (Set 5) gave the better results than others for western atmosphere processes dominated above the territory of Georgia. Also for evolution and improvement of model skill for different time and spatial scale the verification and assimilation methods should be used for further tuning and fitting of model to local conditions.

ACKNOWLEDGMENT

The authors are supported by the EU Commission Project H2020 "VRE for regional Interdisciplinary communities in Southeast Europe and the Eastern Mediterranean" VI-SEEM №675121

REFERENCES

[1] W. C.Skamarock, J. B. Klemp, J.Dudhia, D. O. Gill, D. M. Barker, W. Wang, J.G. Powers, "A description of the advanced research WRF Version 2. *NCAR Tech. Notes*. Natl. Cent. for Atmos. Res., Boulder, Colorado. 2005

[2] A. L. Kondowe, "Impact of Convective Parameterization Schemes on the Quality of Rainfall Forecast over Tanzania Using WRF-Model", *Natural Science,* 2014, Vol.6, pp.691-699

[3] Z. T. Segele, Lance M. Leslie and Peter J. Lamb, "Weather Research and Forecasting Model simulations of extended warm-season heavy precipitation episode over the US Southern Great Plains: data assimilation and microphysics sensitivity experiments", Tellus A 2013, V. 65, pp.1-26.

[4] G.T. De Silva, S. Herath, S.B.Weerakoon and Rathnayake U.R., "Application of WRF with different cumulus parameterization schemes for precipitation forecasting in a tropical river basin", Proceedings of the 13th Asian Congress of Fluid Mechanics 17-21 December 2010, Dhaka, Bangladesh, pp.513-516

[5] G. S. Zepka and Jr. O. Pinto**,** "A Method to Identify the Better WRF Parameterizations Set to Describe Lightning Occurrence", 3rd meteorological Lightning Conference, 21-22 April, 2010, Orlando, Florida, USA pp.1-10

[6] Y. Yair, B. Lynn, C. Price, V. Kotroni, K. Lagouvardos,; E. Morin, A. Mugnai and Llasat M. d. C., "Predicting the potential for lightning activity in Mediterranean storms based on the Weather Research and Forecasting (WRF) model dynamic and microphysical fields". *J. Geophys. Res.*, **115**, D04205, doi: 10.1029/2008JD010868.

[7] J. S. Kain, "The Kain–Fritsch Convective Parameterization" An Update. *J. Appl. Meteor.*, 2004, **43**, 170–181.

[8] E. K. Gilliland, , and C. M. Rowe, "A comparison of cumulus parameterization schemesin the WRF model" . Preprints, *21st Conf. on Hydrology*, San Antonio, TX, Amer.Meteor. Soc., 2007, P2.16.

[9] W. Wang and N. L. Seaman, "A comparison study of convective parameterization schemes in a mesoscale model". *Mon. Wea. Rev.*, 1997, **125**, 252-278.

[10] L.-M. Ma and Z.-M. Tan, 2009: Improving the behavior of the cumulus parameterization for tropical cyclone prediction: convection trigger. *Atmos. Res.*, **92**, 190-211.

[11] S.Y. Hong, J. Dudhia, and S.H. Chen, "A Revised Approach to Ice Microphysical Processes for the Bulk Parameterization of Clouds and Precipitation". *Mon. Wea. Rev.*, 2004, **132**, 103–120.

[12] B.S. Ferrier, Y. Jin, Y. Lin, T Black, E. Rogers, G. DiMego, "Implementation of a new grid-scale cloud and precipitation scheme in the NCEP Eta model". Preprints 15th Conf. on Numerical Weather Prediction,San Antonio, TX, *Amer. Meteor. Soc*., 2002, 280–283.

[13] Y. L. Lin, R. D.Farley, H. D. Orville,"Bulk Parameterization of the Snow Field in a Cloud Model". *J. Appl. Meteor.*, 1983, **22**, 1065–1092.

[14] Z.I.Janjić, "The Step-Mountain Eta Coordinate Model: Further Developments of the Convection, Viscous Sublayer, and Turbulence Closure Schemes. *Mon. Wea. Rev.*, 1994, **122**, 927–945.

[15] Z.I. Janjić, "Comments on "Development and Evaluation of a Convection Scheme for Use in Climate Models". *J. Atmos. Sci.*, 2000, **57**, 3686.

[16] G. A. Grell and Dévényi, D."A generalized approach to parameterizing convection combining ensemble and data assimilation techniques". *Geophys. Res. Lett.*,2002, **29** (14), 1693, doi: 10.1029/2002GL015311.

[17] J. D. Duda, "WRF simulations of mesoscale convective systems at convection-allowing resolutions *Graduate Theses and Dissertations.* , 2011. Paper 10272.

[18] J. S. Kain and J. M. Fritsch, "The role of the convective "trigger function" in numerical forecasts of mesoscale convective systems"**.** *Meteor. and Atmos. Phys.*,1992, **49**, 93–106.

[19] J. S Kain,., and J. M. Fritsch, "Convective parameterization for mesoscale models: The Kain–Fritsch scheme". *The Representation of Cumulus Convection in Numerical Models, Meteor.Monogr.*, No. 24, Amer. Meteor. Soc.,1993, 165–170.

[20] S.Yavinchan, , R. H. B. Exell, and D. Sukawat, 2011: Convective parameterization in a model for the prediction of heavy rain in southern Thailand. *J. Meteor. Soc. Japan*, 89A,pp. 201–224.

[21] A. S.Monin, A. M. Obukhov," Basic laws of turbulent mixing in the surface layer of the atmosphere "Tr. Geofiz. Inst., Akad. Nauk SSSR, 24, 1954, pp.1963–1967. (in Russian).

[22] S.Y. Hong, and J. O. J. Lim, "The WRF single-moment 6-class microphysics scheme (WSM6)". J. Korean Meteor. Soc., vol. 42, 2006, pp.129–151.

[23] A.E.Cohen, S.M.Cavallo, M.C.Coniglo, H.E. Brooks, "AReview of Planetary Boundary Layer Parameterization Schemes and Their Sensitivity in Simulating Southeastern U.S. Cold Season Severe Weather Environments", vol.30, 2015, pp.591-612

[24] I., W. A. Jankov, M.Gallus, Segal, B. Shaw, and S. E. Koch, 2005: The impact of different WRF Model physical parameterizations and their interactions on warm season MCS rainfall. Wea. Forecasting, 20, 1048–1060, doi:10.1175/WAF888.1.

[25] T. Davitashvili, R. Kvatadze, N. Kutaladze, **"**Weather Prediction Over Caucasus Region Using WRF-ARW Model, MIPRO, 2011, *Proceedings of the 34th International Convection*, 2011, Print ISBN: 978-1-4577-0996-8, Opatija, Croatia, pp.326-330

[26] T. Davitashvili, G. Kobiashvili, R. Kvatadze, N. Kutaladze, G. Mikuchadze, "WRF-ARW Application for Georgia" Report of SEE-GRID-SCI User Forum, 2009, Istanbul, Turkey, pp.7-10

Methods and Tools to Increase Fault Tolerance of High-Performance Computing Systems

I.A. Sidorov
Matrosov Institute for System Dynamics and Control Theory of SB RAS, Irkutsk, Russia
ivan.sidorov@icc.ru

Abstract - This work focuses on the development of models, methods, and tools to increase a fault tolerance of high-performance computing systems. The described models and methods are based on automatic diagnostics of the basic software and hardware components of these systems, the use of automatic localization, correction of faults, and the use of automatic HPC-system reconfiguration mechanisms. The originality and novelty of the offered approach consist of creating the multi-agent system with universal software agents, capable of collecting node state data for analysis and thereby enabling the agent to make the make necessary decisions directly.

I. INTRODUCTION

The solution of resource-intensive scientific problems using high-performance computing (HPC) systems is complicated owing to a number of difficulties faced by the applied specialist. Examples of such difficulties include requirements to optimize parallel programs (PP) for the architecture of the HPC-system; complexity in the decomposition of the initial data; necessity to select a specific queue type from the list of the distributed resource manager; requirements with respect to the limitations of administrative policies of the HPC-system (priorities for tasks, limitations in the number of tasks in the queue, limitations in the maximum execution time, etc.), and other conditions required for the preparation and execution of the PP. One of the main problems of large-scale computational experiments that require a large number of simultaneously involved computing resources is the task of providing fault tolerance to the computing process execution. Because PP execution failure directly affects the economic performance, fee for the computing resources, fee for the disruption of experiment deadlines, and fee for labor resources spent searching for problems and failures, etc.

The reasons for the PP execution failures can be classified into the following categories: mistakes made by the developer in the PP code; mistakes made by the user when preparing the initial data, and failures in the hardware and software components of the HPC-system.

The first two categories of failures in some cases can be identified and eliminated during the debugging process on a small number of nodes. The subsequent launch of large-scale computational experiments on a large number of nodes often leads to failures due to the technical problems of computational nodes.

The failure of hardware or software components on one node (processor, memory, hard drives, network devices, system services, etc.) will lead to the total failure of the PP on all involved nodes. Checkpoint mechanism can improve the reliability of the calculations, however, when the program executes on a large number of computing nodes, effective load significantly reduces. This occurs because the PP spends most of the CPU time for organization of a checkpoint and restoration after a failure [1, 2]. In this regard, it is the actual task of creating and applying effective methods and tools that will increase the fault tolerance of HPC-systems, and in turn, will reduce the number of failures of the PP executing on these systems [3]. One approach to solve this problem is the use of monitoring and diagnostics systems.

Known monitoring systems (Ganglia [4], Nagios, Zabbix) focused primarily on the collection of data and providing aggregated information to the operator about status of the HPC-system components. More attention is given to the tools monitoring the PP execution to collect comprehensive data about effectiveness of the use of computing resources (Lapta [5], mpiP, IPM). There are a number of commercial tools for the control of a HPC-system engineering infrastructure (ClustrX Watch, EMC ViRP SRM, HP OpenView [6]) and open-source tools for control of a HPC-system computing infrastructure (Iaso [7] and Octorun [8]).

However, the level of automation at all stages of monitoring, diagnostics, and troubleshooting is unacceptably low and does not support the required level of HPC-systems reliability. Moreover, the big problem for the existing approaches in creating monitoring and diagnostics systems for large-scale HPC-systems (and in the future for exascale systems) is their obvious hierarchical architecture. In accordance with this architecture, the clients of the monitoring system send data to the central (or intermediate) nodes, where the data is collected and processed, and if necessary, control actions are generated. The hierarchical architecture imposes significant limitations on the scalability of the monitoring system, leading to an overload on the network components, data storage systems, and as result, late reaction to the important and critical events.

This paper proposes a model, method, and tools to increase the fault tolerance of HPC-systems with automatic monitoring and intelligent diagnostics of software and hardware components. Originality and novelty of the offered approach consists in creating the multi-agent system with universal software agents, capable of collecting node state data for analysis and decision-making directly on the agent side.

II. Model and Methods of the Computing Node Diagnostics

The model of computing node diagnostics can be represented by the following structure:

$$S = \langle O, Z, T, C, F, R, P, Q, L, I \rangle,$$

where:

- O – is the node components set;
- Z – is the measured characteristics set for each node component;
- T – is the measured characteristics types set;
- C – is the predicates set;
- F – is the control-diagnostics operations set;
- R – is the control-diagnostics operations types set;
- P – is the production rules set;
- Q – is the queue of the control-diagnostics operations that are ready for execution;
- L – is the log of the diagnostics process, and
- I – represents the intervals for starting diagnostics process on the node.

Elements of the set O are different components of a HPC-system node, failure of which may result in failure of the PP, because of the instances of PP being executed on this node. The elements of O can be hardware components (active cooling system, processor, memory banks, hard disk drives, network devices, etc.) and software components (operating system services and processes, operating system network interfaces, connections to network storages, etc.).

Characteristics of Z are divided into the following categories:

- workload computing characteristics of node components (workload of CPU, CPU cores, memory, network and hard-drives input/output workload , etc.);
- characteristics of physical state of node components (temperature of CPU and motherboard, uninterruptible power supply state, hard drives state, etc.);
- characteristics of executed program state (priority, CPU-usage time, memory usage, hard-drive, and network storages utilization, etc).

Elements of T include following value types of measured characteristics: logical (Boolean); integer; float; percent (unsigned integer in interval [0…100]); chars; string (up to 256 chars); text (up to 2^{20} chars).

Elements of F include operations designed to perform control actions to collect data of the hardware and software component states and diagnostic actions to localize and troubleshoot the identified faults. Control-diagnostic operations implement mappings: $F: O \times Z \to Z$ or $F: O \to Z$.

Elements of R include the following control-diagnostic operations types:

- r_1 – data collecting operations (obtain data about the current state of active cooling systems, the temperature of the motherboard and processors, the hard-drive SMART info, the amount of free memory on storage devices, the status of network devices, the status of operating system services, the status of connections to the network drives, etc.);
- r_2 – test operations (fast or full memory tests, testing of the hard drive devices for badblocks, testing of the network components, etc.);
- r_3 – fault localization operations (identification number of the failed RAM bank; identification of processes using a lot of memory or generating a huge workload for hard-disk drives or network components; identification of the physical location of the failed hard drive; identification of failed cooler number on the motherboard etc.);
- r_4 – troubleshooting operations (cleaning temporary files from storage devices, restarting services of operating system services, remounting network drives, etc.);
- r_5 – troubleshooting verification operations (check the free disc space, check the free amount of RAM, check stopped processes of the operating system, check temperature of the CPU, etc.);
- r_6 – critical situations operations (node shutdown, disconnecting node from the HPC-system resource manager, killing all calculating processes to decrease CPU-temperature, sending a notification to the system administrator, etc.).

The type of operation defines the priority of its execution in the operations queue Q: r_1 – priority 1, r_2 – priority 2, r_3 – priority 3, r_4 – priority 4, r_5 – priority 5, r_6 – priority 0 (maximal).

It should be noted that the operations of types r_1, r_2, r_3, and r_5 are classified as data obtain operations, which collect information about the state of the node components without making any changes. Operations r_4 and r_6 are designed to perform automatic control actions aimed at troubleshooting problems in software and hardware components. Control actions can change the mode of the node components in order to provide fault-tolerant functioning of this node in the HPC-system. In case of failure to resolve a critical situation triggered, control actions should automatically withdraw the defective node from the pool of available resources of the HPC-system.

Analysis of the control-diagnostic characteristics from the set Z is performed by predicates from the set C, written as follows:

$$c_i : z_j < logical\ \ operator > \ \ const,$$

where $c_i \in C$, $z_j \in Z$, $i = \overline{1, n_C}$, $j = \overline{1, n_Z}$.

On the left side is specified characteristic of the set Z, is the value to be analyzed. Among the logical operators that can be used are $<$, $\leq$, $>$, $\geq$, $==$, and $!=$. The value type of the $const$ on the right side and the characteristic value z_k should be of the same type. It allows the construction of complex expressions using operators AND, OR, XOR, and NOT.

The set P contains the production rules of the form $IF\ c_k\ THEN\ f_i$. This structure is interpreted as follows: if the predicate c_k is true, it is necessary to perform the operation f_i. Production is ready when the values for all characteristics included in the predicate are defined. Production is completed if the predicate is true and the operation on the right side has been added to the queue.

The process of diagnostics of the computing node includes the following steps. With some interval from I, the monitoring agents that are working on the nodes run the diagnostic procedure. The initial data of the diagnostics is a vector containing values of the logical characteristics from Z, which depend on the current interval from I, which define modes of the diagnostic process (for ex. use or not troubleshooting operations) and the quality of the diagnostic process (quick testing, full testing or advanced testing).

There is a certain order of interpretation of productions and the addition of operations in the queue. It was stated earlier that there are 6 defined operations types in the model. In the first step, ready productions are chosen where the right side has the control-diagnostic operations of types r_1, r_2 or r_3 (data obtain operations). Interpretation of selected productions was conducted, which included the calculation of predicates and in case their value was true, the operations were added in the queue Q.

Processing operations in the queue Q are performed in the following mode. One of the requirements of the diagnostics tools is to reduce the total execution time of operations on the computing node. To achieve this, a strategy was selected for the execution of the maximum number of operations in unit time. According to this strategy, all control-diagnostic operations added to the queue Q executed immediately in a separate thread in parallel mode with other operations. At the same time, the only condition to start operation in the queue is checking to ensure that the same component from O does not perform other operations. This condition is checked in order to avoid execution of mutually contradictory operations.

After completion of an operation, the queue Q expands the set of defined characteristics values, and as a result, expands the list of ready productions. The procedure for interpretation of ready and not previously completed productions starts again.

It should be noted that the second and the subsequent steps of the productions interpretation, in addition to the productions with specified operations of types r_1, r_2 and r_3, also selected productions of type r_6 at the right side (actions in critical situations). In the case the predicate is true for such production, this operation is placed at the first position in the queue. This approach ensures immediate response to critical situations.

After all the operations of types r_1, r_2, and r_3 are completed, the next stage is executed. At this stage, the selected productions on the right are the operations of r_4 type (troubleshooting operation). After completion of all r_4 type operations, the next stage is executed with selected productions of r_5 type operations at the right side (troubleshooting verification operations). At the last stage, when type r_6 operations are performed and when all other operations are completed, the diagnostics process is considered complete.

It should be noted that all elements of sets O, Z, C, F, P, I for the structure S are provided with informative descriptions. During the interpretation of the predicates and operations to the diagnostic log file L sequentially added the actions taken. According to this principle it is performed the construction of the diagnostics log file.

III. Example of Node Diagnostics

As an example, consider a simple diagnostics process of computing node RAM. Let the elements of the sets for the structure *S* be as follows:

O:

- o_1 – RAM of the computing node;
- o_2 – user tasks that executed on the node.

Z:

- z_0 – initial condition of diagnostic (type is string);
- z_1 – total RAM size (type is integer);
- z_2 – used RAM size (type is percent);
- z_3 – identification number of failed RAM bank (type is integer);
- z_4 – status of sending notification to the administrator of HPC-system (type is logical);
- z_5 – status of sending notification to the user of task o_2 (type is logical);
- z_6 – status of command for stopping task execution o_2 (type logical);
- z_7 – status of finishing task o_2 (type logical).

C:

- c_0: z_0 == *"quick"* - performs quick diagnostics;
- c_1: $z_1 != 2^{33}$ - total size of RAM should be 8 Gb;
- c_2: $z_3 > 0$ - number of failed RAM bank was successfully indentified;
- c_3: $z_2 > 99$ - node RAM used more than 99 %.

F:

f_1: $\{o_1\} \rightarrow \{z_1,z_2\}$ - operation to obtain node RAM information (type is r_1);

f_2: $\{o_1\} \rightarrow \{z_3\}$ - operation to localize a failed RAM bank (type is r_3);

f_3: $\{o_1, z_3\} \rightarrow \{z_4\}$ - operation to notify HPC-system administrator about failed RAM bank (type is r_6);

f_4: $\{o_2, z_2\} \rightarrow \{z_5\}$ - operation to notify the user of the task o_2 about exceeding the limit of RAM (type is r_6);

f_5: $\{o_2\} \rightarrow \{z_6\}$ - stop task execution (type is r_4);

f_6: $\{o_1, o_2\} \rightarrow \{z_7\}$ - check the status of stopping program execution o_2 (type is r_5).

P:

p_0: *IF* c_0 *THEN* f_1 - perform quick node diagnostics;

p_1: *IF* c_1 *THEN* f_2 - RAM size was changed, need to localize a faulty RAM bank;

p_2: *IF* c_2 *THEN* f_3 - notify the administrator about identified failed RAM bank;

p_3: *IF* c_3 *THEN* f_4 - notify user of task o_2 about exceeding the limit of the RAM;

p_4: *IF* c_3 *THEN* f_5 - stop task o_2, because the limit of the RAM was exceeded.

The diagnostics process for such a structure may work as follows. Let z_0=*"quick"*. Operations queue after interpretation of the ready productions will include the following operation: $Q=\{ f_1 \}$. When the operation f_1 completes, the values of parameters z_1 и z_2 will be defined. Suppose that a node has two 4GB RAM banks, and one bank has failed, then $z_1 = 2^{32}$ (in such cases, the operating system often continues to work, excluding the address space of the failed bank). Also, assume that $z_2 > 99$ (critical memory usage reached with the possibility of a swap usage). In the next step, the predicates c_1 and c_3 are true. Operations of localization $Q=\{ f_4, f_2 \}$ were added to the queue in accordance with the order of productions processing. In the next step, the predicates c_2 and c_3 are true and the following operations, $Q=\{ f_3 , f_5 \}$, will be added to the queue. And in the last step, the operation $Q=\{ f_6 \}$, for checking process stopping, will be added to the queue.

It should be noted that this example is demonstrative of the above models and methods and does not contain a complete list of operations and productions for a comprehensive diagnostic of computing node RAM.

IV. Realization of Diagnostics Tools

Software implementation of the diagnostics tools is realized as a part of the meta-monitoring toolkit [9, 10] designed for heterogeneous large-scale HPC-systems, which includes hundreds of thousands of calculating nodes. This toolkit is based on the use of service-oriented technology, multi-agent technology, methods of creation of expert systems, methods of decentralized data processing, distributed data storage, and decentralized decision-making.

The above model and methods were implemented as part of the autonomous software agents of the meta-monitoring system that functions on the compute nodes in the background mode. Agents perform the functions of node state data collection, collected data analyzing, generating and executing the control actions, and communications with other agents. Ability to analyze data and to make the necessary decisions on the node side is the key feature of the implemented approach. As discussed above, in most monitoring systems, the collected data are always sent to the central node where they are processed and stored. This implementation significantly reduces the number of operations to analyze data on the node side; however, regular sending of data to the master node creates an additional load on the network protocols stack, which also requires CPU time (formation, analysis and control of network packets). In our approach, the analysis of the data at the node side allows sending data to the central node only if necessary. Comparative analysis of the Ganglia monitoring client (gmond) and agent of the developing system showed that for the same frequency of the data gathering (every 30 sec.), our agent with the functions of local data analysis and local decision-making consumes 37% less CPU time.

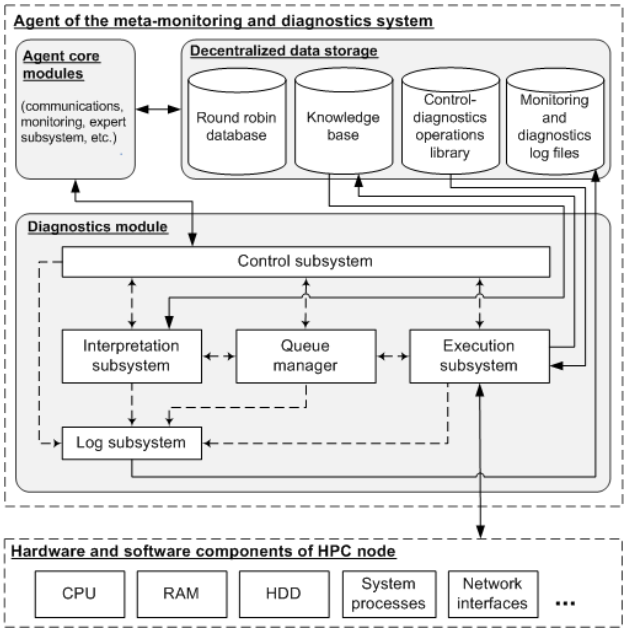


Figure 1. The architecture of the base subsystems of the diagnostics tools

Diagnostics tools include the following base subsystems (Fig. 1):

- decentralized data storage, which includes knowledge base (contains specifications for characteristics, productions, predicates, list of previously detected problems and faults etc.), control-diagnostics operations library (contains modules realized using the specialized language), round-robin database to store periodically obtained monitoring data, and log files;
- control subsystem, which performs functions of coordinating node diagnostics and includes mechanisms to start the diagnostics process in certain time intervals, or in moments when computing node is idle;

- subsystem for interpretation of productions and predicates;
- the control-diagnostics operations queue manager;
- execution subsystem for control-diagnostics operations, and
- the logging subsystem.

Program realization of the agent and all base subsystems was implemented using the C++ programming language. Collection of information about node state is implemented using SIGAR [11] library. Implementation of control-diagnostics operations is performed using a specialized language developed by the author. This language is a subset of ECMA Script [12], which was extended to support calls of external commands, output stream processing and a number of other mechanisms, which allow obtaining data about the state of non-standard software and hardware devices. Interpreter for productions and predicates are based on the ECMA Script interpreter too. All the specifications of the diagnostics tools are described in the JSON format.

To store and have access to collected data, was implemented a solution based on round-robin database principles. The developed solution showed a higher performance for our tasks in comparison with the universal tools (RRDtools and MRTG [13]). For storage of productions, predicates, list of previously detected problems and other information, the light-weight embeddable relational database SQLite [14] is used.

V. Experemental Results

The developed methods and tools of diagnostics were successfully approved in the Supercomputer Center of ISDCT SB RAS [15]. Configuration of heterogeneous HPC-system includes the following resources:

- 110 nodes with AMD Opteron 6276 processors (total number of cores 3520);
- 20 nodes with Intel Xeon 5345 processors (total number of cores 160);
- 2 nodes with GPU Nvidia Tesla C1060 accelerators (total number of cores 1920).

During diagnostics tools testing, several software and hardware resources in critical condition and in an error state were detected:

- [node011.mat.icc.ru]: RAM bank #4 failed;
- [node-12.bf.icc.ru]: CPU temp is more 87°C;
- [node-04.bf.icc.ru]: SMART health status bad;
- [node103.mat.icc.ru]: /store used 100%;
- [node087.mat.icc.ru]: iface ib0 errors detected;
- [tesla01.icc.ru]: CPU warnings in system log file.

The automatic testing of the nodes components, fault localization and troubleshooting, isolation of unreliable nodes, and HPC-system reconfiguration ensured the prevention of many computational processes failures.

VI. Conlusion

The described approach allows to control, diagnose, and troubleshoot software and hardware components of HPC-system nodes in a finite number of steps. It ensures minimization of run-time diagnostics and troubleshooting processes by using the mechanisms of parallel operations execution, and the approach also supports an increase in the fault tolerance of nodes thereby increasing the fault tolerance of PP executed on these nodes. All these features make it possible to enhance the fault tolerance of the HPC-system.

Acknowledgment

The study was supported by Russian Foundation of Basic Research, projects no. 15-29-07955-ofi_m and no. 16-07-00931. The author is grateful to A. G. Feoktistov for discussions and valuable comments.

References

[1] F. Cappello, A. Geist, W. Gropp, S. Kale, B. Kramer, M. Snir, “Toward exascale resilience: 2014 update,” Supercomputing Frontiers and Innovations, vol. 1, no. 1, pp. 5-28, April 2014.

[2] P. Kogge, J. Shalf “Exascale Computing Trends: Adjusting to the "New Normal" for Computer Architecture,” Computing in Science & Engineering, vol. 15, no. 6, pp. 16-26, November 2013.

[3] B. Mohr “Scalable parallel performance measurement and analysis tools – state-of-the-art and future challenges,” Supercomputing frontiers and innovations, vol. 1, no. 2, pp. 108-123. July 2014.

[4] M. Massie, B. Li, V. Nicholes, V. Vuksan, Monitoring with Ganglia Tracking Dynamic Host and Application Metrics at Scale, O’Reilly Media, November 2012.

[5] A. Adinets, P. Bryzgalov, V. Voevodin, S. Zhumatii, D. Nikitenko, K. Stefanov “Job Digest: an approach to dynamic analysis of job characteristics on supercomputers,” Numerical methods and programming: Advanced Computing. vol. 13, no. 4, pp. 160–166, 2012 (in Russian).

[6] “HP OpenView,” http://www.openview.hp.com/solutions/ams/ams_bb.pdf [online, accessed: 31-Jan-2016].

[7] K. Lu, X. Wang, G. Li, R. Wang “Iaso: an autonomous fault-tolerant management system for supercomputers,” Frontiers of Computer Science, vol 6, no. 3, pp. 378-390, May 2014.

[8] A. Antonov, V. Voevodin, A. Daugel-Dauge, S. Zhumatii, D. Nokonenko “Ensuring effective operational control and battery life of the MSU Supercomputer,” Bulletin of South Ural State University, vol. 4, no. 2, pp. 33-43, 2015 (in Russian).

[9] I. V. Bychkov, G. A. Oparin, A. P. Novopashin, I. A. Sidorov “Agent-Based Approach to Monitoring and Control of Distributed Computing Environment,” Lecture Notes in Computer Science, vol. 9251, pp. 253-257, September 2015.

[10] I. A. Sidorov, A. P. Novopashin, G. A. Oparin, V. V. Skorov “Methods and tools of meta-monitoring of distributed computing envinronments,” Bulletin of South Ural State University, vol. 3, no. 2, pp. 30-42, 2014 (in Russian).

[11] “System Information Gatherer and Reporter API,” https://github.com/hyperic/sigar, [online, accessed: 31-Jan-2016].

[12] “Standard ECMA-262,” http://www.ecma-international.org/publications/standards/Ecma-262.htm [online, accessed: 31-Jan-2016].

[13] H. Allen, P. Regnauld “Network management workshop: MRTG/RRDTool,” Apnic 29, Kuala Lumpur, March 2010.

[14] G. Allen, M. Owens, The Definitive Guide to SQLite, Apress, November 2010.

[15] “Irkutsk Supercomputer Center of SB RAS,” http://hpc.icc.ru/ [online, accessed: 31-Jan-2016].

Logical-Probabilistic Analysis of Distributed Computing Reliability

A.G. Feoktistov and I.A. Sidorov
Matrosov Institute for System Dynamics and Control Theory of SB RAS, Irkutsk, Russia
agf@icc.ru

Abstract - The aim of the study is to develop tools of increasing a problem solving reliability in a heterogeneous distributed computing environment by applying a diagnostics of computing resources components and using an analysis of problem solving schemes. A scheme (a plan) is an abstract program for a problem solving. A special attention is paid to the calculation of a problem solving scheme reliability on the basis of a logical-probabilistic method. This method is based on transiting from Boolean functions for a reliability description of a problem solving scheme to probability functions for determining indicators of such reliability. Improving a problem solving scheme reliability is carried out by a resource reservation. The resource reservation applied in a problem solving scheme provides obtaining of a reliability indicator that approximates maximally the predetermined criterion of reliability, taking into account limitations on a number of reserve resources. The example of the problem solving scheme and calculating its reliability is represented.

I. INTRODUCTION

Today, special attention of specialists in the field of cloud infrastructures and Grid-systems is given to developing fundamentals of distributed computing for solving large-scale scientific problems in various subject areas. The development of new methods and algorithms for a computing management in a heterogeneous distributed computing environment (DCE), where nodes have a complex hybrid structure, is actively carried out.

The heterogeneity of a DCE, the presence of hybrid components in it structure, the wide range of fundamental and applied problems and the necessity of scalable computing actualize the researches related to improving a reliability of computational processes.

In theory of computer systems reliability there exist two traditional approaches to provide a reliability of computational processes [1]. The first approach aims to make an attempt to restore a computational process after some fault of software or hardware components of a computer system. The second approach aims to provide an operability (fault tolerance) of a computer system while fault of software or hardware components.

The first approach is typically realized by using checkpoint mechanisms. Unfortunately, these mechanisms don't allow to correctly restore a computational process after a fault for some classes of parallel programs. In addition, a total time of checkpointing and restarting of a computational process can be comparable to a time of solving a problem in a heterogeneous DCE [2].

The second approach is based on using of a redundancy of hardware, processes and data. Within this approach there exist problems with switching from a faulty component to a reserved component [3, 4].

A promising approach to improve a computer system reliability [5, 6] is proactive detecting and troubleshooting of current and potential hardware faults and using this knowledge for planning computational processes and allocating computing resources for an execution of these processes.

However, the implementation of this approach for a heterogeneous DCE raises some important problems. Among them: the lack of universal models and methods of determining reliability indicators for a heterogeneous DCE; the lack of methods and tools for applying reliability indicators in the traditional resource managers; the insufficient automatization of troubleshooting in traditional monitoring systems for DCE.

In this paper some aspects of an approach to ensure a reliability of solving problems in a heterogeneous DCE are considered. This approach is developed to the DCE, which is characterized by following features:

- Computer clusters act as nodes of the DCE. These clusters can consist of hardware and software components (computational elements) of various architecture and configuration for different parallel programming technologies. Clusters are organized based on both allocated and non-allocated computational elements and, hence, significantly differ by the reliability degree of their computing resources.
- Different levels of the DCE have various categories of users working at them. Clusters are used by both local users of these clusters and global users of the DCE.
- To solve a problem, a user should form a job for the DCE. The job is a specification of solving problem that include information on the required computing resources, executable applied programs, input/output data and other required information.
- The specialized multi-agent system (MAS) for a distributed computing management [7] is used for formulating problems, planning problem solving schemes and forming jobs of global users. Formed jobs are decomposed by the MAS into a set of subjobs for clusters. These subjobs and jobs of

local users are managed by traditional local resource managers at the cluster levels. Agents of the MAS observe for subjobs execution processes.

- Clusters do not have enough free resources to simultaneously process all jobs in its queues.

The features of our approach are listed below.

A high competition of users for common resources of the DCE leads to the necessity to take into account all characteristics of these resources in a process of their allocation in order to achieve the required quality of job execution [8]. Usually, a time or cost of job execution, indicators of resources load balancing, coefficients of resources efficiency or performance are used as main criteria of the efficiency for a distributed computing management. Today, besides the listed criteria, the reliability of job execution in the DCE is the practical importance criterion. Increasing the distributed computing reliability allows guaranteeing the fulfillment of other criteria of job execution in a greater degree.

The calculation of a problem solving scheme reliability is carried out on the base of a logical-probabilistic method [9]. This method provides transiting from Boolean functions for a reliability description of a system under study to probability functions for determining indicators of such reliability.

A reservation is one of the simplest and effective methods to improve computer system reliability [10]. In this case, the reservation is to use additional (reserve) nodes. These additional nodes are expected to assume the faulty nodes functions in the execution process of a problem solving scheme. Additional nodes are loaded reserve. These reserve nodes are also used as main nodes. Using loaded reserve due to the fact that other forms of a reservation [11] such as an unloaded reserve, a hot reserve and a multi-version job realization lead to significant overhead and are more suitable for real time systems. The reservation is carried out for single elements (nodes). A large number of nodes may be required for such reservation. Thence, in this paper the problem of forming a set of reserve nodes providing of a reliability indicator that approximates maximally the predetermined criterion of reliability, taking into account limitations on a number of reserve resources.

Data about the nodes reliability are provided by the author's meta-monitoring system considered below.

The successful results of using the logical-probabilistic method for studying distributed systems are known [12, 13]. A practical comparison of similar methods is discussed in [14].

II. META-MONITORING SYSTEM

In the considered DCE the meta-monitoring system is used to test the DCE nodes, to collect data about current state of the DCE nodes, to detect nodes faults, to diagnose and to partially repair these faults. In this system the processes of collecting data, monitoring nodes, detecting and diagnosing faults are based on using multi-agent technologies and the unique methods for decentralized processing and distributed storage of data. Unlike the known, the meta-monitoring system provides:

- an ability to obtain data from the most popular high-performance monitoring systems (Ganglia, Nagios, Zabbix etc.);
- a wide range of traditional and original functions for obtaining and collecting data about hardware and software components of the DCE nodes;
- the high-level toolkit for the implementation of the original functions for obtaining and collecting data as modules in various programming languages;
- the specialized tools for collecting and analyzing data of the engineering infrastructure of computer clusters and data centers;
- the auxiliary tools for a unification and aggregation of data received from different sources;
- the new intelligent tools for an automated expert analysis of data and generation of control actions for changing nodes states;
- the new service tools for periodic testing nodes, detecting, diagnosing and partial repairing nodes faults;
- the special application programming interface based on open standards for providing access to the monitoring data for external software systems.

An accumulation of data about faults is carried out in both automatic and manual modes. The registration of faults in an automatic mode is performed by means the meta-monitoring tools listed above. Similar registration of faults in a manual mode allows the DCE administrator to describe hardware and software faults that were not recognized in an automatic mode.

The basic information about a fault is the following information:

- the object (computing node, storage system, facility of an engineering infrastructure for a computer cluster, etc.) containing a faulty component;
- the faulty component (hard drive, bank of RAM, network adapter, etc.);
- the degree of a component fault impact: without consequences for the object, a fault of a computational process in the object, a critical fault and, as result, non-operability of the object;
- the time of a fault detection;
- the time of troubleshooting;
- the result of troubleshooting: a complete correction of the fault, a partial correction of the fault and further using the object in the DCE, excluding the object from the DCE configuration and reconfiguring the DCE.

An estimation of reliability indicators of the DCE objects is based on accumulated data about the object component faults.

The meta-monitoring system, mentioned above, is described in details in [15].

III. LOGIC-PROBABALISTICS ANALYSIS

A. Conceptual Model

Denote by the symbols F, Z, N and A the sets of program modules, modules parameters, computing nodes and agents correspondently. The relations between these sets are denoted by the symbols R_{in}, R_{out}, R_{af} and R_{an} such that $R_{in} \subset Z \times F$, $R_{out} \subset Z \times F$, $R_{af} \subset A \times F$ and $R_{an} \subset A \times N$. Then, a conceptual model of the DCE can be described by the structure

$$M = < F, Z, N, A, R_{in}, R_{out}, R_{af}, R_{an} > .$$

The relations R_{in} и R_{out} define respectively the sets of input and output parameters for modules and, thus, determine the information-logical connections between modules. The relation R_{af} determines connections between agents and modules that can be performed by agents. A type of the relations R_{in}, R_{out} and R_{af} is the "many-to-many". The relation R_{an} determines agents' nodes. A type of the relation R_{an} is the "one-to-many". Nodes of various clusters have different degrees of the reliability. Computer clusters are represented by agents in the management system of the DCE.

The user's request for computational services of the DCE is defined as a problem in the nonprocedural form for the management system: "calculate the values for parameters of the subset $Z_{out} \subset Z$ knowing the values for parameters of the subset $Z_{in} \subset Z$". In general, a set of schemes for solving this problem can exist in a model of the DCE. Each of these schemes determines what modules and in what order should be executed. Denote by S the set of schemes.

It is required to form the set S, to choose a single scheme $s \in S$, to determine the agents, which will to execute the modules of the scheme s, and to allocate the nodes for an execution of these modules. Indicators of the scheme performance must satisfy to such criteria of the problem solving as the time or the cost. An additional criterion is the scheme reliability – the probability $p^*(t)$ of the scheme performance in the time moment t. All these criteria are set by a user. A satisfaction degree of the scheme performance indicators to the user's criteria defines a level for the quality of service in the DCE. In this formulation, the problem of computation planning and resource allocation is the NP (Non-deterministic Polynomial)-hard [16].

In order to support further generality reasoning, two fictitious modules f_1 and f_2 are entered into the set F. The initial module f_1 has the empty set of input parameters and the subset $Z_{in} \subset Z$ as the set of output parameters. This module defines initial data of the problem. The target module f_2 has the subset $Z_{out} \subset Z$ as the set of input parameters and the empty set of output parameters. This module defines target parameters of the problem.

The set S can be represented by the bipartite directed acyclic graph $G = < V, U >$. The set V includes the subsets V_Z and V_F of vertices corresponding to parameters from the set Z and modules from the set F.

Denote by n_f the number modules included in the scheme s. Assume that, information-logical connections between modules of the scheme s are described by the Boolean matrix W. Dimensions of this matrix are $n_f \times n_f$. The matrix element $w_{i,k} = 1$ means that the module f_i depends on the module f_k.

B. Multi-agent System

A hierarchical structure of the MAS can include two or more layers functioning agents. Agents can play different roles and perform different functions at their levels of the hierarchical structure of the MAS. The agents roles can be constant and temporary, occurring at discrete time moments due the need to organize collective interaction. The hierarchy levels of agents differ by the amount of their knowledge. Agent of the higher level hierarchy has a large amount of knowledge in comparison with agents of the lower hierarchy level.

A subsystem of the MAS for distributed computing management includes the agents of computation planning and resource allocating. These agents are designed for creating problem solving schemes and allocating resources for schemes performance. Any agents may be united in a virtual community of agents (VCA). In such VCA the agents interact between themselves on the base of a competition or cooperation. In detail, the multi-agent algorithm of computation planning and resource allocation is considered in [17]. This algorithm provides the obtaining scheme s satisfied to the time or cost criteria, which defined by a user. The data from meta-monitoring system are used in MAS.

Let us now consider new aspects of the multi-agent management related with the scheme s reliability.

C. Logical-Probabalistics Model

Denote by n_a the number of agents in some VCA. These agents are expected to participate in the scheme s performance. Denote by x the set of the Boolean variables $x_{i,j,k}$ representing the events of execution ($x_{i,j,k} = 1$) or not-execution ($x_{i,j,k} = 0$) for the module f_i in the k-th node allocated by the agent a_j, $i = \overline{1, n_f}$, $j = \overline{1, n_a}$. The index k of the variable $x_{i,j,k}$ defines correspondently the main ($k = 1$) and reserve ($k \geq 2$)

nodes of the agent a_j. These nodes are allocated for an execution of the module f_i. Each agent a_j can allocate the number c_j of the reserve nodes in the scheme s performance. Assume that, all nodes of the same agent are homogeneous. A possibility to run modules in the main or reserve nodes causes various scenarios of the scheme s performance.

Let the Boolean function $y_i(x)$ determines conditions of the module f_i execution in the scheme s performance and the function $p_{i,j,k}(t)$ shows a probability of the module f_i in k-th node of the agent a_j. The recurrent formulas (1)-(3) describe a logical circuit of the execution reliability of a problem solving process:

$$y_1(x) \equiv 1, \tag{1}$$

$$y_i(x) = \begin{cases} h(x), & if\ i = 2, \\ x_{i,j_i,1} h(x), & if\ i \geq 3, \end{cases} \tag{2}$$

$$h(x) = \bigwedge_{\forall k: w_{i,k}=1} y_k(x), \tag{3}$$

where $i \in \overline{1,n_f}$, $k = \overline{1,n_f}$, $j_i \in \overline{1,n_a}$. Initially, the agent a_{j_i} allocates the main ($k = 1$) node for the module f_i.

The Boolean function $y_2(x)$ corresponding to the target module f_2 determines the scheme s reliability indicator. After all substitutions in (1)-(3), the function $y_2(x)$ takes the form

$$y_2(x) = \bigwedge_{i=3}^{n_f} x_{i,j_i,1}, \tag{4}$$

where $j_i \in \overline{1,n_a}$.

A transition [18] from the Boolean function $y_2(x)$ to the probability function $P(t)$ is implemented through the rules in Table I. The probability function $P(t)$ has the following form:

$$P(t) = \prod_{i=3}^{n_f} p_{i,j_i,1}(t).$$

The function $P(t)$ calculates the probability for the single scenario of the scheme s performance. If $P(t) < p^*(t)$, then function $y_2(x)$ transformation by means of the probability indicators improvement for the function structure elements is carried out. The function $y_2(x)$ determined in (4) is represented in the disjunctive normal form. Therefore, this function is monotonic. Such property provides the preservation or improvement of the scheme s probability when the probability indicators for the function structure elements are improved.

TABLE I. RULES OF TRANSITING FROM BOOLEAN FUNCTIONS TO PROBABILITY FUNCTIONS

The Boolean function element	The probability function element
$x_{i,j,k}$	$p_{i,j,k}(t)$
$\bar{x}_{i,j,k}$	$1 - p_{i,j,k}(t)$
$\wedge$	$\cdot$
$\vee$	$+$

Denote by J the set of indexes for such agents that $c_j = 0$, $J = \{j : c_j = 0, j \in \overline{1,n_a}\}$. The algorithm for the transformation process of the function $y_2(x)$ corresponding to the scheme s without nodes reservation into the function $y_2'(x)$ corresponding to the scheme s with nodes reservation includes following stages.

- If the following condition is fulfilled

$$\sum_{j=1}^{n_a} c_j = 0,$$

then the transformation process is finished. Else, the determination of the index for module with minimum probability of this module execution is carried out:

$$k = \underset{i=\overline{3,n_f}, j_i \notin J}{\arg\min} \left(1 - \prod_{l=1}^{e} \left(1 - \bar{p}_{i,j_i,l} \right) \right),$$

where the number of nodes allocated by the agent a_{j_i} for the module f_i is denoted by n_{i,j_i} and $e = n_{i,j_i}$.

- The number c_{j_k} is reduced on one unit.
- The reservation of the additional node by the agent a_{j_k} for the module f_k is carried out by replacing the element $x_{k,j_k,e}$ of the transformable formula for the function $y_2(x)$ by the element $x_{k,j_k,e} \vee x_{k,j_k,e+1}$, where the number of nodes allocated by the agent a_{j_k} for the module f_k is denoted by n_{k,j_k} and $e = n_{k,j_k}$.
- The number n_{k,j_k} is increased on one unit.
- The function $y'_2(x)$ obtained in the transformation process of the function $y_2(x)$ is represented in the following form:

$$y'_2(x) = \bigvee_{l=1}^{n} K_l, \tag{5}$$

$$K_l = \bigwedge_{i=3}^{n_f} x_{i,j_i,e},$$

$$n = \prod_{i=3}^{n_f} \sum_{j=1}^{n_a} n_{i,j},$$

where $e = k_{j_i}$, $k_{j_i} \in \overline{1, n_{i,j_i}}$. Each elementary conjunction in (5) represents the one of the possible scenarios of the scheme s performance. All elementary conjunctions are numbered from 1 to n in according to their rank ascending. The orthogonalization of the function $y'_2(x)$ is carried out on the base of the algorithm offered in [18]. Such orthogonalization of the function $y'_2(x)$ provides an incompatibility of the possible scenarios of the scheme s performance. The orthogonal function $\tilde{y}'_2(x)$ has the following form:

$$\tilde{y}'_2(x) = K_1 \vee \overline{K}_1 K_2 \vee \ldots \vee \overline{K}_1 \overline{K}_2 .. \overline{K}_{n-1} K_n .$$

- A simplification of the function $\tilde{y}'_2(x)$ by means of the removing of the conjunctions that are identically zero and the redundant conjunctions is carried out.
- A transition from the Boolean function $\tilde{y}'_2(x)$ to the probability function $P'(t)$ is implemented through the rules in Table I. The probability function $P'(t)$ has the following form:

$$P'(t) = \sum_{i \in I} p'_i(t),$$

where the probability of the i-th scenario for the scheme s performance is denoted by $p'_i(t)$.

- The scheme s reliability is calculated with the using of the function $P'(t)$.
- If $P'(t) < p^*(t)$, then the transition to the first stage is carried out. Else, the transformation process is finished.

The nodes reservation process considered above provides scheme reliability indicator that is as close as possible to the given criterion of scheme reliability taking into account the limitations on the number of nodes allocated by each agent. These limitations ensure convergence of the nodes reservation process.

IV. Scheme Reliability Calculation Example

Let the set S (Fig. 1) includes the two schemes $s_1 : f_1, f_3 - f_4, f_6, f_2$ and $s_4 : f_1, f_3 - f_5, f_6, f_2$ for the problem "calculate the parameter z_4 value knowing the parameter z_1 value".

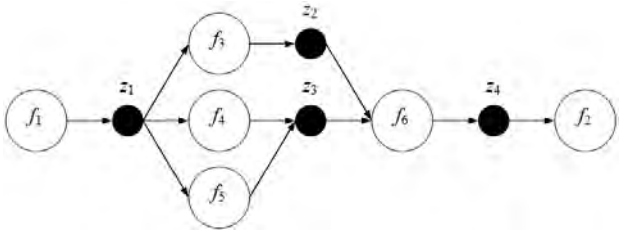


Figure 1. Bipartite directed acyclic graph for the set S

Such constructions as "f_i, f_j" and "$f_i - f_j$" mean correspondingly that the modules f_i and f_j are executed sequentially only or may be executed in parallel. The scheme reliability indicator is defined as following: $p^*(t) = 0{,}995$. All Boolean functions $y_i(x)$ for the problem solving schemes are shown in Table II.

TABLE II. Boolean Functions $y_i(x)$ for Problem Solving Schemes

Scheme	
s_1	s_2
$y_1(x) \equiv 1$,	$y_1(x) \equiv 1$,
$y_3(x) = y_1(x) x_{3,j_3,1}$,	$y_3(x) = y_1(x) x_{3,j_3,1}$,
$y_4(x) = y_1(x) x_{4,j_4,1}$,	$y_5(x) = y_1(x) x_{5,j_5,1}$,
$y_6(x) = y_3(x) y_4(x) x_{6,j_6,1}$,	$y_6(x) = y_3(x) y_5(x) x_{6,j_6,1}$,
$y_2(x) = y_6(x) =$	$y_2(x) = y_6(x) =$
$= x_{3,j_3,1} x_{4,j_4,1} x_{6,j_6,1}$.	$= x_{3,j_3,1} x_{5,j_5,1} x_{6,j_6,1}$.

Let the set $A = \{a_1, a_2, a_3, a_4, a_5\}$ includes 5 agents and the nodes reliability of these agents are correspondingly following: 0,999, 0,999, 0,9999, 0,9999 and 0,99. The modules distribution by the agents are shown in Table III. The performance probabilities for the schemes s_1 and s_2 without nodes reservation are correspondently equal 0,9988 and 0,9898. Thus, the reservation of the additional node in scheme s_2 is necessary.

TABLE III. Modules Distribution by Agents

Scheme	Module	Agent				
		1	*2*	*3*	*4*	*5*
s_1	f_3	–	–	–	+	–
	f_4	–	+	–	–	–
	f_6	–	–	+	–	–
s_2	f_3	–	–	–	+	–
	f_5	–	–	–	–	+
	f_6	–	–	+	–	–

After the modules distribution by agents, the function $y_2(x)$ corresponding scheme s_2 has the following form:

$$y_2(x) = x_{3,4,1}x_{5,5,1}x_{6,3,1}.$$

The reservation is implemented by the agent a_5, because the resources of this agent provide the minimum probability for execution of the module f_5 in comparison with probabilities for execution of the modules f_3 and f_6. After reservation, the functions $y'_2(x)$ и $\tilde{y}'_2(x)$ have the forms:

$$y'_2(x) = x_{3,4,1}x_{5,5,1}x_{6,3,1} \vee x_{3,4,1}x_{5,5,2}x_{6,3,1},$$

$$\tilde{y}'_2(x) = x_{3,4,1}x_{5,5,1}x_{6,3,1} \vee x_{3,4,1}x_{5,5,2}x_{6,3,1}\bar{x}_{5,5,1}.$$

A transition from the Boolean function $\tilde{y}'_2(x)$ to the probability function $P'(t)$ and calculation of the scheme s_2 probability are implemented.

$$P'(t) = p_{3,4,1}p_{5,5,1}p_{6,3,1} + p_{3,4,1}p_{5,5,2}p_{6,3,1}(1 - p_{5,5,1}) = \\ = 0{,}9997$$

Since $P'(t) > p^*(t)$, the transformation process is finished.

V. CONCLUSION

The results of the study considered above suggest the following conclusions.

- Up to now, while logical-probabilistic analysis is carried out, occurs the problem of constructing Boolean functions, which describe the reliability logic circuit of a studied system. In this paper the new universal recurrent formulas for constructing such functions based on a conceptual model of the heterogeneous DCE are obtained.
- The use of the models and algorithm developed for logic-probabilistic analysis of the distributed computing reliability in multi-agent management system improves the quality of service for users' jobs in the DCE.
- An integrated application of the diagnostic methods and tools of computing nodes jointly with the tools for the problems solving processes analysis in these nodes on the scheme level provides a substantial increase reliability of the DCE on the whole.

ACKNOWLEDGMENT

The study was supported by Russian Foundation of Basic Research, projects no. 15-29-07955-ofi_m and no. 16-07-00931. The authors are grateful to G. A. Oparin for discussions and valuable comments.

REFERENCES

[1] D. P. Siewiorek, R. S. Swarz, Reliable Computer Systems: Design and Evaluation. Natick, MA: CRC Press, 1998.

[2] F. Cappello, A. Geist, W. Gropp, S. Kale, B. Kramer, M. Snir, "Toward exascale resilience: 2014 update," Supercomputing Frontiers and Innovations, vol. 1, no. 1, pp. 5-28, April 2014.

[3] E. Elmroth and J. Tordsson, "A standards-based Grid resource brokering service supporting advance reservations, coallocation, and cross-Grid interoperability," Concurrency and Computation-Practice & Experience, vol. 21, no. 18, pp. 2298-2335, June, 2008.

[4] E. Bauer, R. Adams Reliability and Availability of Cloud Computing, Wiley-IEEE Press, September 2012.

[5] C. Engelmann, G.Vallee, T. Naughton, "Proactive Fault Tolerance Using Preemptive Migration," Parallel, Distributed and Network-based Processing, 2009 17th Euromicro International Conference on, pp. 252 – 257, February 2008

[6] G. Da Costa, T. Fahringer, J. Rico-Gallego, I. Grasso, A. Hristov, H. Karatza, A. Lastovetsky, F. Marozzo, D. Petcu, G. Stavrinides, D. Talia, P. Trunfio, H. Astsatryan, "Exascale machines require new programming paradigms and runtimes," Supercomputing Frontiers and Innovations, vol. 2, no. 2, pp. 6-27, September 2015.

[7] I. V. Bychkov, G. A. Oparin, A. G. Feoktistov, V. G. Bogdanova, A. A. Pashinin, "Service-oriented multiagent control of distributed computations," Automation and Remote Control, vol. 76, no. 11, pp. 2000-2010, November 2015.

[8] D. A. Menasce, E. Casalicchio, "QoS in Grid Computing," IEEE Internet Computing, vol. 8, no. 4, pp. 85-87, July 2004.

[9] I. A. Ryabinin, "Logical-Probabilistic Calculus: A Tool for Studying the Reliability and Safety of Structurally Complex Systems," Automation and Remote Control, vol. 64, no. 7, pp. 1177-1185, July 2003.

[10] J. Li, "A Model of Resource Reservation in Grid," International Conference on Environmental Science and Information Application Technology, Wuhan, China: IEEE CS Pupl., pp. 199-202, July 2009.

[11] D. A. Zorin, V. A. Kostenko, "Algorithm for Synthesis of Real-Time Systems under Reliability Constraints," Journal of Computer and Systems Sciences International, vol. 51, no. 3, pp. 410-417, May 2012.

[12] S. Rai, M. Veeraraghavan, K. Trivedi, "A Survey of Efficient Reliability Computation Using Disjoint Products Approach," Networks, vol. 25, pp. 147-165, May 1995.

[13] J. Xing, C. Feng, X. Qian, P. Dai, "A simple algorithm for sum of disjoint products," Reliability and Maintainability Symposium, Reno: IEEE CS Publ., pp. 1-5, January 2012.

[14] A. Rauzy, E. Chatelet, Y. Dutuit, C. Berenguer, "A practical comparison of methods to assess sum-of-products," Reliability Engineering & System Safety, vol. 79, no. 1, pp. 33-42, January 2003.

[15] I. V. Bychkov, G. A. Oparin, A. P. Novopashin, I. A. Sidorov "Agent-Based Approach to Monitoring and Control of Distributed Computing Environment," Lecture Notes in Computer Science, vol. 9251, pp. 253-257, September 2015.

[16] M. R. Garey, D. S. Johnson, "Computer and Intractability. A Guide to the Theory of NP-Completeness," New York: W. H. Freeman and Company, 1979.

[17] V. G. Bogdanova, I. V. Bychkov, A. S. Korsukov, G. A. Oparin, A. G. Feoktistov, "Multiagent Approach to Controlling Distributed Computing in a Cluster Grid System," Journal of Computer and Systems Sciences International, vol. 53, no. 5, pp. 713-722, September 2014.

[18] A. A. Pospelov, "Logical Methods of Circuit Analysis and Synthesis," Moscow-Leningrad: Energiya, 1964. (in Russian).

Distributed Graph Reduction Algorithm with Parallel Rigidity Maintenance

Diego Sušanj* and Damir Arbula*
* Faculty of Engineering, Department of Computer Engineering, Rijeka, Croatia
dsusanj@riteh.hr, damir.arbula@riteh.hr

Abstract - Precise localization in wireless sensor networks depends on distributed algorithms running on a large number of wireless nodes with reduced energy, processing and memory resources. Prerequisite for the estimation of unique nodes locations is rigid network graph. To satisfy this prerequisite network graph needs to be well connected. On the other hand, execution of distributed algorithm in the graphs with large number of edges can present significant impact on scarce resources of wireless nodes. In order to reduce number of edges in the network graph, novel distributed algorithm for network reduction is proposed. Main objective of the proposed algorithm is to remove as many edges as possible maintaining graph rigidity property. In this paper a special case of graph rigidity is considered, namely the parallel rigidity.

I. INTRODUCTION

In today's era of rapid development of technology and human progress, concept of ubiquitous computing has become a reality in the form of computers, mobile phones, home appliances, and other, so called, smart devices. Directly driven by such progress is a need for simple and fast connection of all those devices without relying on any existing fixed infrastructure.

Wireless ad hoc networks were created in response to the above mentioned needs. They represent a self-configuring, decentralized dynamic networks in which devices or nodes, can move freely without depending on existing fixed infrastructure. In other words, devices are both the users and the infrastructure of those networks.

A. Wireless sensors networks

A certain type of ad hoc wireless networks consisting of spatially distributed autonomous sensors which are used to monitor various environmental conditions, such as temperature, pressure, illumination, etc. are called wireless sensor networks. These networks are made up of several to hundreds of interconnected nodes with cooperative relaying of the information through the network to the input/output gateway nodes.

Size and the price of these nodes are directly dependent on constraints such as the type of power source, the amount of memory, processing power and communication bandwidth. The ultimate goal is to have nodes that are as independent as possible which relates to low or no maintenance and no use of the existing infrastructure, most notably in terms of power.

Nodes are usually equipped with batteries and optionally they have some means of energy harvesting i.e. means of obtaining energy from the environment. Consequently, it is essential to pay attention to energy consumption both on hardware and on software level.

There are two basic rules: (1) energy needed to send a unit of data is far greater than the energy needed to process the same amount of data, and (2) messages sent in smaller steps with lower transmission power is more energy efficient than those sent in a single step with greater transmission power [1].

B. Problem definition

The motivation for this work primarily stems from the issues of scalability of wireless sensor networks, more precisely excessive network density. Problems arise when there is a need for more advanced calculations, such as localization of dense network on low end nodes and because of that, successful execution of algorithm is limited by the maximum number of neighbors for each node [2].

In this sense, an adequate selection of neighbors enables implementation of distributed algorithms in dense wireless networks without losing network rigidity property thus preserving information required for the estimation of unique nodes locations.

II. LOCALIZATION

Location service is one of the basic services of many emerging computing and networking paradigms [5]. In the case of wireless sensor networks, sensor nodes need to know their locations in order to put detected and recorded events into proper spatial context, or for example to enable more informed and thus more efficient routing.

Sensors are usually deployed without any previous knowledge of their absolute or relative location and there is no infrastructure that would enable them to localize themselves after deployment. One method of informing node of its location is manual measurement and location calculation. Such procedure is not feasible for larger networks or in the case of deployment in the inaccessible environment.

Another method involves the use of GPS, the global positioning system. To be able to localize the node using GPS, it is necessary to add a GPS receiver, which significantly increases costs of the hardware. Also its size and energy consumption are oftentimes prohibitive, as

well as the requirement of being in the range of satellite signals i.e. outdoors.

A. Location estimation

Localization methods are classified by measurements modality and two most popular are distance and azimuth between neighboring nodes. In this paper only methods based on azimuth measurements are considered.

Azimuth measurements between neighboring nodes are usually performed using few different techniques. One is based on using known radiation pattern of one or more directional antennas, measuring received signal strength. By comparing the intensity of the received signal received on multiple antennas the angle of arrival of the signal can be determined [3].

Another approach is based on the measurement of the radio signals phase differences using an array of omnidirectional antennas. There is also similar technique that is using ultrasound signals [4].

B. Anchored and anchor-free localization

The location information can be defined in the absolute coordinate system of the environment, or in the relative coordinate system – one that is being agreed upon between nodes themselves. Although the actual absolute locations in the latter case are not known, network topology is matching and, by using translation, rotation and scaling transformations, the relatively positioned network graph can be aligned with its absolute counterpart.

The basic premise of anchored localization is having set of nodes (anchors) that know their actual location in the environment i.e. by using GPS or manually. Using at least three (in the two-dimensional space - $\mathbb{R}^2$) or four (in case of $\mathbb{R}^3$) non-collinear anchors it is possible to localize the entire network.

In case of anchor-free localization there is no information about actual location of any node in the network. The goal of this localization is to construct new relative coordinate system usually pinned by placing one node in the origin and one of its neighbors at coordinate (1, 0). In this paper only two-dimensional case is considered.

III. GRAPH RIGIDITY

Basic definitions of graph theory and rigidity are given in [9] and they are the basis for the modern theory of rigidity and localization of wireless sensors networks that is developed further in [5], [6] and [7].

Wireless network topology is usually described using graph, but to be able to include nodes locations, it is necessary to augment this description with the positions of the graph vertices. This concept is called point formation (or framework) and it represents one of the possible spatial realization of the network graph.

Congruent formations are those that have equal distances between all vertices while *equivalent formations* have equal distances between connected vertices. Formation is defined as a *globally rigid* if, and only if, all equivalent formations are also congruent. The key property of networks with globally rigid formations is that it is possible to estimate unique location of all nodes.

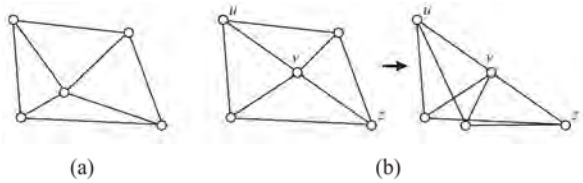


Figure 1. Generic (a) and non-generic (b) formation

Generic formations are point formations in which the coordinates of the vertices are algebraically independent over the set of rational numbers. If the formation is not generic, there is possibility that set of neighboring nodes is part of same $d-1$-dimensional space, which results in existence of equivalent, but not congruent formations.

Fig. 1 shows two different formations of the same graph with the location of vertices that are: (a) generic, (b) not generic because these vertices are collinear and it's possible to find equivalent non-congruent formation.

The problem of determining whether the individual formation is globally rigid is NP-hard problem, while the determination of generic formation is easier. In addition, as an important feature, and a reason why the graph theory is so thoroughly studied, is that global rigidness of generic formation depends only on the network graph.

A. Parallel rigidity

Equation (1) defines mapping of measured azimuths between nodes to a set of values which represents global azimuths under which node i see node j, (2). Two point formations β_p and β_q are parallel if their corresponding graphs $\mathbb{G}_N$ and azimuth functions β are equal for all vertices i and j provided $i \neq j$.

$$\beta : \mathfrak{L} \mapsto [0, 2\pi) \tag{1}$$

$$\beta(i, j) \triangleq \Theta_{ij} \tag{2}$$

Example of two parallel formations is shown in Fig. 2. Subfigure (a) shows the basic formation, while in (b) one can see translation of that formation. Subfigures (c) and (d) show preservation of azimuth measurements i.e. angles between edges of the graph, while scaling the network. Subfigure (e) shows a non-congruent realization of the formation in (a) while maintaining constraints set by measured azimuths. Formation which has more than one non-congruent solutions while using constraints based on azimuth measurements is called flexible, while the one that has only one congruent solution is called *parallel rigid*. One such parallel rigid formation is given in figure (f) [5], [9].

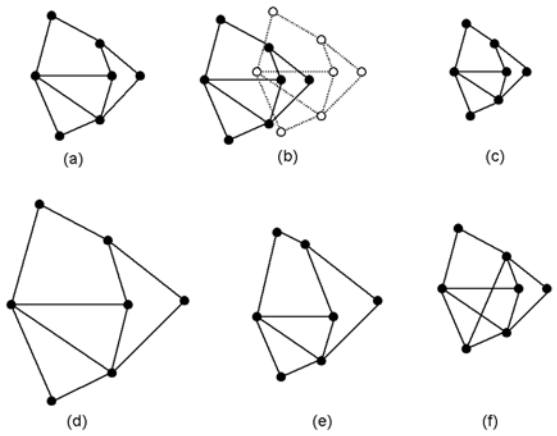


Figure 2. Parallel formations

IV. PROPOSED METHOD

In this paper we propose a method for edge removal based on modified depth first search.

Algorithm starts in the root node. Distributed election of the root node is specific problem that was not focus of proposed method thus the node with the lowest ID was always selected as the root node.

Root node sends message to all his neighbors requesting from them to response with number of their neighbors. This number is defined as node weight. After root receives weights from all of its neighbors, a special message called token is sent to the neighbor with lowest weight. In the case in which several nodes have the same weight token is sent to the one with lowest ID. Node receiving token starts the same process as root, and requests all his neighbors' weights.

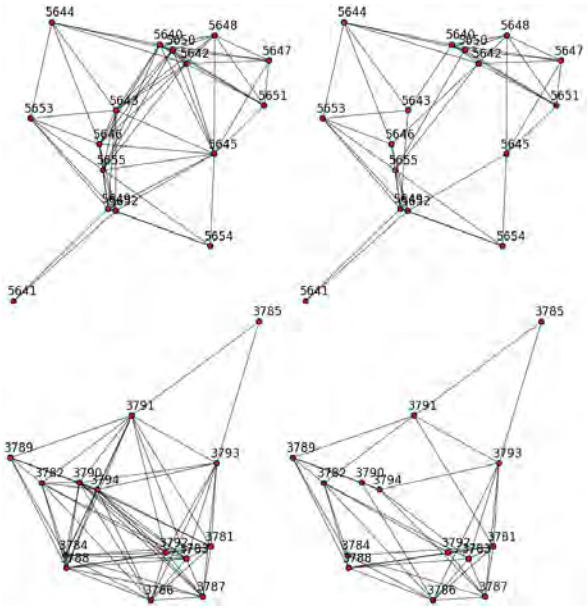


Figure 3. Network graphs before and after edge removal algorithm

Every node, after receiving token, records sender's ID in stack. In the case node doesn't have any more neighboring nodes from which it has not received token he returns token to the node on top of the stack. When node receives token from three different neighbors it is marked as rigid but continues to process information in the same way as before.

Algorithm terminates globally when root node receives return token and has no neighbors left to forward it to. In that moment, all nodes should be flagged as rigid with newly defined set of traversed edges $\mathfrak{L}_q$ which is subset of initial set $\mathfrak{L}_p$.

Fig. 3 shows two examples of network graphs before (on the left) and after (on the right) execution of edge removal algorithm. Network graphs after execution of the algorithm have far less edges in the dense parts of the graph.

V. RESULTS

Testing of this algorithm is done in three phases. First two phases are performed using Pymote simulator [8].

First phase includes automated testing on two sets of networks. First set is made of networks with specific topologies on which this, and other algorithms, didn't perform well. Second set is based on random generated networks with number of nodes between 8 and 128. In this phase proposed algorithm was tested on 8000 networks.

In the second phase benchmark testing was performed in order to compare few versions of proposed algorithm. Benchmark set was made up of 100 networks with number of nodes between 8 and 16. Comparison of algorithms is based on two parameters (1) maximum node weight after algorithm execution compared to the number before execution and (2) total number of remaining edges.

In the third phase five network topologies selected from the benchmark set were implemented and tested on real nodes in laboratory conditions.

Proposed algorithm passed all three phases always producing rigid networks.

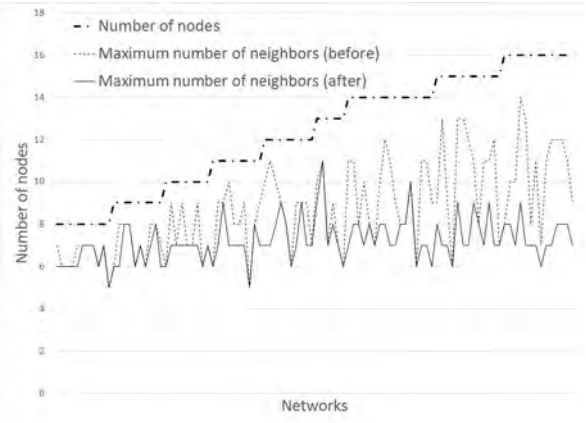


Figure 4. Maximum number of neighbours (weight) before and after edge removal algorithm for benchmark set networks

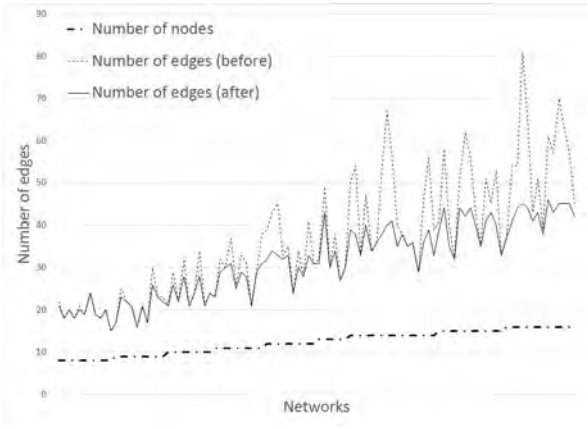


Figure 5. Number of edges before and after edge removal algorithm for benchmark set networks

Graph on Fig. 4 represents maximum node weight per network before and after execution of algorithm compared to the number of nodes in network. Algorithm is executed in the 100 networks from the benchmark set with random topologies. Important observation is that maximum weight after edge removal algorithm (full line) does not increase proportionally with the number of nodes in the network.

Another measure of algorithm success, is the number of remaining edges after execution of edge removal algorithm as shown in Fig. 5. Number of removed edges increases with increasing density of the network.

VI. CONCLUSION

Wireless sensor networks consist of a few dozens up to several thousand of spatially distributed autonomous sensors which are used to monitor various environmental conditions. In these networks location information is required for many services i.e. to provide measurement proper spatial context or to localize recorded events.

Localization methods have their limitations. One of them is a need for higher computing resources and higher energy demand on low end nodes. Energy demand is notable in dense networks, because of higher communication complexity. Solution for that problem is reducing the density i.e. by creating communication tree or reduced communication graph.

On the other hand, requirement to estimate unique location of all nodes is global rigidity of network formation.

To alleviate communication complexity while retaining the rigidness of network graph, distributed algorithm was designed and implemented. This edge removal algorithm was tested using both Pymote simulator and real wireless sensor nodes. The results obtained in the simulator and in the implementation on the real nodes in the laboratory have shown a significant reduction in the number of edges and maximum network weight while retaining graph rigidness.

The field of localization in wireless sensor networks and the development of new, and improving existing algorithms is extremely broad. Future work includes improvement of proposed algorithm by creating minimal rigid graph (graph with theoretic minimum number of edges), as well as creating method for selecting a subset of neighbors which results in as small as possible localization errors.

Furthermore, it would be interesting to study the rigidity with directed graph, i.e. network where mutual visibility of nodes is not guaranteed.

REFERENCES

[1] F. Zhao and L. J. Guibas, *Wireless Sensor Networks: An Information Approach*. Morgan Kaufmann Publishers, Elsevier Inc., 2004.

[2] D. Arbula, "Distributed algorithm for node localization in anchor free wireless sensor network," Faculty of Electrical Engineering and Computing, University of Zagreb, 2014.

[3] J.-R. Jiang, C.-M. Lin, F.-Y. Lin, and S.-T. Huang, "ALRD: AoA Localization with RSSI Differences of Directional Antennas for Wireless Sensor Networks," *International Journal of Distributed Sensor Networks*, vol. 2013, Mar. 2013.

[4] N. B. Priyantha, A. K. L. Miu, H. Balakrishnan, and S. Teller, "The Cricket Compass for Context Aware Mobile Applications," in *7th ACMConf. Mobile Computing and Networking (MOBICOM)*, 2001.

[5] T. Eren, W. Whiteley, and P. N. Belhumeur, "A Theoretical Analysis of the Conditions for Unambiguous Node Localization in Sensor Networks," Department of Computer Science, Columbia University, 2004.

[6] T. Eren, W. Whiteley, and P. N. Belhumeur, "Using Angle of Arrival (Bearing) Information in Network Localization," in *2006 45th IEEE Conference on Decision and Control*, 2006, pp. 4676–4681.

[7] D. Zelazo, A. Franchi, and P. R. Giordano, "Rigidity Theory in SE(2) for Unscaled Relative Position Estimation using only Bearing Measurements," *arXiv:1311.1044 [math]*, Nov. 2013.

[8] D. Arbula and K. Lenac, "Pymote: High Level Python Library for Event-Based Simulation and Evaluation of Distributed Algorithms," *International Journal of Distributed Sensor Networks*, vol. 2013, Mar. 2013.

[9] G. Laman, "On graphs and rigidity of plane skeletal structures," *J Eng Math*, vol. 4, no. 4, pp. 331–340, Oct. 1970.

Architecture of Virtualized Computational Resource Allocation on SDN-enhanced Job Management System Framework

Yasuhiro Watashiba*, Susumu Date†, Hirotake Abe‡, Kohei Ichikawa*,
Yoshiyuki Kido†, Hiroaki Yamanaka§, Eiji Kawai§, and Shinji Shimojo†
*Nara Institute of Science and Technology, Nara, Japan
Email: {watashiba, ichikawa}@is.naist.jp
†Osaka University, Osaka, Japan
Email: {date, kido, shimojo}@cmc.osaka-u.ac.jp
‡University of Tsukuba, Ibaraki, Japan
Email: habe@cs.tsukuba.ac.jp
§National Institute of Information and Communications Technology (NICT), Tokyo, Japan
Email: {hyamanaka, eiji-ka}@nict.go.jp

***Abstract*—Nowadays, users' computation requests to a high-performance computing (HPC) environment have been increasing and diversifying for requiring large-scale simulations and analysis in the various science fields. In order to efficiently and flexibly handle such computation requests, resource allocation of the virtualized computational resources on an HPC cluster system such as Cloud Computing service is attracting attention. Currently, we aim to realize a novel resource management system (RMS) that enable to handle various resources of an HPC cluster system, and have been studying and developing the SDN-enhanced Job Management System (JMS) Framework, which can manage an interconnect as network resources by integrating Software Defined Networking (SDN) concept into a traditional JMS. However, the current SDN-enhanced JMS Framework cannot allocate virtualized computational resources to a job because the computational resource management is performed by the mechanism of a traditional JMS. In this paper, we propose a mechanism to handle virtualized computational resources on the SDN-enhanced JMS Framework. This mechanism enables to deploy virtual machines (VMs) requested by the user to the computing nodes allocated to a job and execute job's processes in the VMs.**

I. Introduction

Nowadays, a high performance computing (HPC) center, which operates HPC environment to provide resources for users, is required the efficient resource management due to a fact that users' computations has been increasing and diversifying. In various science fields, the necessity of HPC resources to perform scientific simulations and analysis has been growing for the purpose of gaining high computing performance and solving complex and/or large-scale scientific problems. Since each of such computations has different resource usage pattern and resource requirements, the HPC center has to accommodate many and various user computations into its HPC environment. Thus, it is essential that the resources in an HPC environment are handled by considering the characteristics of each computation.

The dominant architecture of the HPC environment has been becoming a cluster system, which is composed of many computing node connected with a high-speed network called an interconnect. An general HPC cluster system has two type of resources: computational resources such as CPU and memory, and network resources in its interconnect. Thus, it is important to efficiently and flexibly manage these resources in order to provide an appropriate set of resources for user's computation. In many HPC cluster systems, Job Management System (JMS), such as NQS [1], PBS [2] and Open Grid Scheduler/Grid Engine (OGS/GE) [3], is widely adopted as Resource Management System (RMS) which has the role to administer users' computations as jobs and efficiently allocate resources to them for the purpose of efficient workload balancing. However, most traditional JMS available today can only handle computational resources. This causes the computing performance degradation under the influence of inefficient allocation of the other resources to a job.

In regard to the computational resource allocation to a job managed by a traditional JMS, today's diversified user computations may require more flexible resource provision from an HPC cluster system. Generally, the computational resources allocated to a job by a traditional JMS are physical computational resources, that is, the computation of the job is performed on computing nodes directly. Since the computing environment in a computing node, such as the version of the kernel and libraries, is static, it may not be suitable for a computation of job with a specified resource requirement and restriction. Therefore, a mechanism to provide computational resources with user-required computing environment is also required. In order to realize such flexible computational resource management, virtualized computational resources, which is allocated computational resources as virtual machines (VMs) to a job, have been focused on.

However, since virtualized computational resources have

some overhead coursed by virtualizing the physical computational resources, the computing performance on VMs may be lower than the case of physical resource allocation. Thus, a user who submits a job suitable for the allocation of physical computational resources hopes traditional resource allocation rather than the provision of virtualized computational resources. For achieving efficient and flexible resource management, a JMS on an HPC cluster system is desirable to allocate both physical and virtualized computational resources according to the features of jobs. However, a traditional JMS in an HPC cluster system does not have any mechanism to handle VMs as virtualized computational resources, and the RMS for Cloud Computing targets to allocate resources to users' computations by virtualized computational resources.

From these considerations above, we aim to realize a novel flexible and efficient JMS capable of handling various resources as well as computational resources on an HPC cluster system. So far, we have addressed an issues for managing an interconnect in an HPC cluster system as network resources, and have been studying and developing network-aware JMS, called SDN-enhanced JMS framework [4], [5]. The SDN-enhanced JMS framework can allocate an appropriate set of computational and network resources by integrating Software Defined Networking (SDN) into a traditional JMS. However, in the SDN-enhanced JMS framework, since a traditional JMS takes charge of the management of computational resources, the SDN-enhanced JMS framework does not have the functionality to control virtualized computational resources. In this paper, we propose the resource management mechanism to handle VMs as virtualized computational resources on the SDN-enhanced JMS framework as well as physical computational resources.

This paper is structured as follows. Section II briefly introduces the SDN-enhanced JMS framework, which can manage not only physical computational resource, but also the interconnect of HPC cluster system as network resources. Section III analyses the required functionalities for handling VMs as virtualized computational resources, and then explains the implementation of the function enhancement on the SDN-enhanced JMS framework. In Section IV, we evaluate the behaviors of the proposed mechanism to control virtualized computational resources. In Section V, we conclude this paper with the future work.

II. SDN-enhanced JMS Framework

In this section, we briefly mention the system structure and mechanism of the SDN-enhanced JMS framework [4], [5]. The SDN-enhanced JMS framework enables to manage and allocate both computational and network resources by integrating OpenFlow [6], which is a technology for realizing the SDN concept, into a traditional JMS. The OpenFlow enables the administrator to manage the whole network in a centralized and programmable manner through a SDN controller, which can be designed as software. Figure 1 shows the architecture of the SDN-enhanced JMS framework. In designing the SDN-enhanced JMS framework, the low-cost reusability and a high degree of program portability were considered for facilitating to cooperate with various traditional JMSs. As the result, the functionalities to handle network resources was implemented as an external module called Network Management Module (NMM). Therefore, the mechanism to manage computational resources in the SDN-enhanced JMS framework is provided by the traditional JMS as illustrated in Fig. 1.

In the implementation of the SDN-enhanced JMS framework, OGS/GE [3] has been adopted as a traditional JMS to integrate the functionalities of network resource management. In the SDN-enhanced JMS framework, an interface to cooperate with the NMM is required for a traditional JMS. The OGS/GE has the Parallel Environment Queue Sort (PQS) API, which allows an administrator to define the new policy to decide the allocating computational resources. The SDN-enhanced JMS framework realize the cooperation between the traditional JMS and the NMM by enhancing the PQS API.

The NMM has two components: the network control component and the brain component. The network control component has two functionalities to manage network resources in the interconnect of HPC cluster system: retrieving the information of the interconnect, and controlling the communication paths among computing nodes allocated to a job. In order to implement these two functionalities by leveraging the functions of OpenFlow, Trema [7], which is a development framework of the OpenFlow controller, has been included in the network control component.

The role of the brain component is to decide how to allocate resources to a job based on resource usage of the HPC cluster system and user requirement of the job. To gather the information of a target job and the usage of both computational and network resources, the brain component is connected with a traditional JMS and the network control component through the XML-RPC [8]. Moreover, a traditional JMS do not have any algorithm to determine appropriate resources for a job based on the usage and requirement of both computational and network resources. Thus, the brain component has the *resource assignment policy class module*, through which the administrator can define how to allocate resources to a job as a resource assignment policy. The *resource assignment policy class module* provides a set of APIs that facilitate to flexibly design a resource assignment policy in a Ruby script. In a job script, a user can require an appropriate resource allocation algorithm by indicating the name of resource assignment policy as a parameter.

These mechanisms enable the SDN-enhanced JMS framework to manage and allocate both computational and network resources. In the SDN-enhanced JMS framework, network resources are handled as *flow entries*, which are defined how to communicate between computing nodes. That is, all communication paths between computing nodes allocated to each job are administered as allocated network resources. Moreover, the information of the network resource allocation and the the usage of each link in the interconnect of the HPC cluster system retrieved by the network control component are stored in a database which is equipped in the NMM.

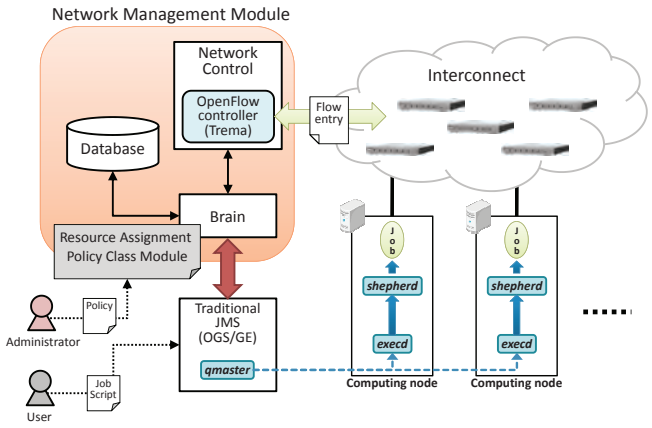


Fig. 1. Architecture of original SDN-enhanced JMS framework.

III. Architecture and Implementation

In this section, we derive the functionalities which are required for handling VMs as virtualized computational resources in the process flow of JMS, and then explain the implementation of the novel SDN-enhanced JMS framework with a mechanism to control virtualized computational resources.

A. Virtualized Computational Resources

The virtualized computational resources by leveraging a VM enables to flexibly control the computing environment in computational resources allocated to a job. In a VM, a user can construct arbitrary computing environment that does not depend on the structure of computing node in a HPC cluster system. Thus, a computation of a job can be perform on appropriate computing environment according to its resource requirement and restriction. Resource provision by leveraging VMs as virtualized computational resources has been adopted in Cloud Computing [9], [10]. In the HPC field, resource allocation by leveraging virtualized computational resources has attracted much attention due to the growth of virtual resource technology, and many researches to utilize virtualized computational resources have been conducted [11], [12], [13], [14].

For running a VM in a computing node, the hypervisor is generally required. The hypervisor works as the middleware for connecting between a VM and physical hardware (e.g. Kernel-based Virtual Machine (KVM), Xen, VMware vSphere Hypervisor, and Hyper-V). The VMs in a computing node are not shared the internal resources each other due to a fact that they are managed as different computing environments. Since a today's computing node has of many CPU cores and large memory, the characteristic is useful to guarantee the performance of computational resources allocated to a job from other jobs' behaviors. In the viewpoint of resource management, virtualized computational resources also have some advantages compared with the resource management based on physical computational resources.

B. Functionalities for Handling Virtualized Computational Resources

In this section, we analyze the required functionalities for controlling the virtualized computational resources as well as the physical computational resources and network resources on the SDN-enhanced JMS framework. To realize that virtualized computational resources are handled on the SDN-enhanced JMS framework, the following functionalities are considered to be necessary.

1) User interface to indicate an arbitrary user-required VM for executing a program in a job.
2) Controlling the behaviors of VMs in computing nodes allocated to a job.
3) Installing modules and configuration for deploying the job process in VMs.
4) Managing the information of virtualized computational resources allocated to jobs.

The first functionality is user interface to indicate a user-required VM for executing a computation in the job. This functionality is necessary to allocate arbitrary computing environment constructed by a user to a job in an HPC cluster system. Additionally, the parameters to define the number of CPU cores and the amount of memory in a VM should be included in this functionality. The second functionality is required for controlling the behavior of VMs because the SDN-enhanced JMS framework cannot operate the action of VMs in computing nodes. Before starting to assign job processes to VMs, they have to boot up. Moreover, after the job is finished, the VMs must shut down for releasing the resources allocated to a job. Since the behavior of a VM is generally controlled by a hypervisor, this functionality enables to operate the hypervisor from the SDN-enhanced JMS framework. The third functionality is a mechanism to set up an environment of VM as a computing node of the HPC cluster system. In the proposed virtualized computational resource management, since it is supposed that the computing environment in a VM is constructed by a user, the VM does not have the information and the functions to control job processes from the SDN-enhanced JMS framework (e.g. environment variables and module to manage job processes, user information for authentication and authorization, a path to access the user's home directory in the HPC cluster system, and so on). Thus, the functionality to construct computing environment capable of managing a VM from the SDN-enhanced JMS framework is essential. The fourth functionality is a mechanism to manage the information of virtualized computational resources allocated to jobs. Though the amount of used virtualized computational resources can be administered as the usage of physical computational resources, this information is necessary for administrators and users to check the status of virtualized computational resources.

C. Implementation of Virtualized Computational Resource Management

In this section, we describe the implementation of the novel SDN-enhanced JMS framework which equips the functional-

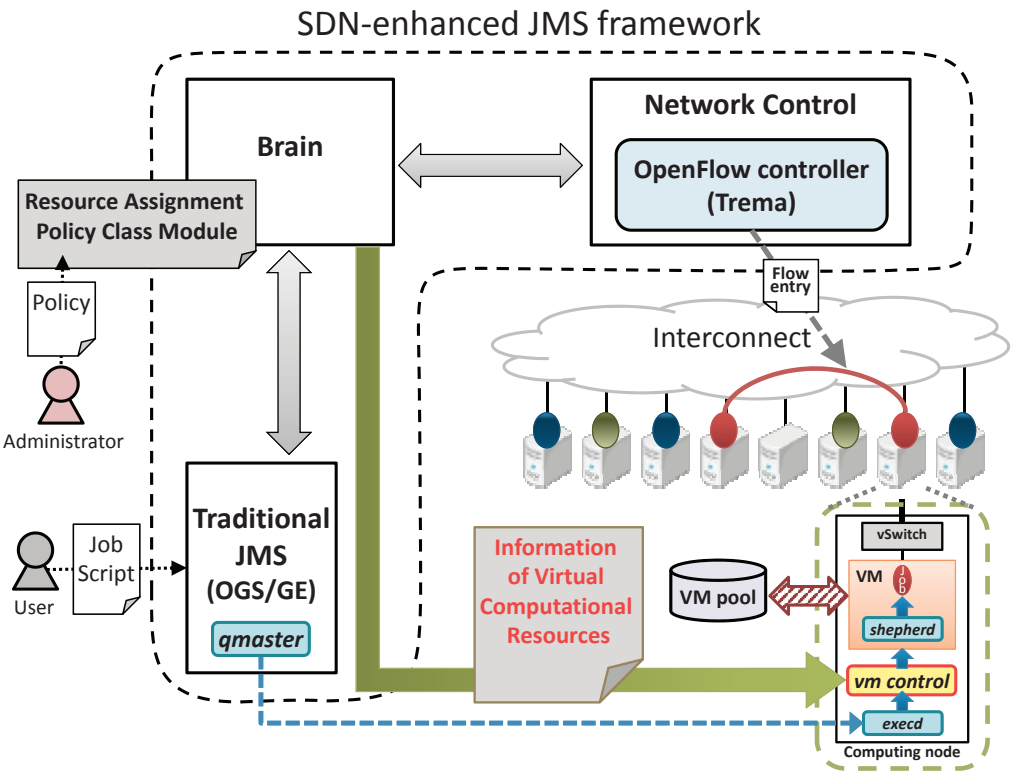


Fig. 2. Architecture of SDN-enhanced JMS framework with the functionalities virtualized computational resource management.

ities for handling virtualized computational resources. As for a hypervisor in a computing node, the KVM hypervisor is adopted due to the fact that the OS of a computing node in many current HPC cluster system is Linux. In order to manage virtualized computational resources as well as physical computational resources on the SDN-enhanced JMS framework, we have adopted a strategy of enhancing a mechanism of computational resource management on the OGS/GE. Figure 2 illustrates architecture of our proposed SDN-enhanced JMS framework with the functionalities for managing virtualized computational resource.

For the first functionality listed in Section III-B, we implemented a user interface by extending the parameters of job script offered by the OGS/GE with similar uses as the requirement of network resources. An example of extended job script is shown in Fig. 3. The extended job script allows a user to embed the job's requirements to virtualized computational resources as well as physical computational. In this example, the requirements of virtualized computational resources are composed of a path of VM image and parameters of VM setting such as the amount of memory in the VM. As for the number of CPUs configured in a VM, it is limited to one in this implementation for equalizing the number of allocating slots.

The second functionality listed in section III-B was achieved by operating the KVM hypervisor on each computing node from the SDN-enhanced JMS framework. In order to control

```
#!/bin/csh
#$ -q  QUEUE_NAME
#$ -pe  orte 32
#$ -l  netprio = policy_name
#$ -l  vm image = /vm/my-vm
#$ -l  vm memory = 8gb

mpirun  -np $NSLOTS ./a.out
```

Fig. 3. Example of extended job script in proposed resource management system.

the KVM hypervisor on each computing node, we developed a new control module, called "*vm control*", and deployed it in each computing node. This module is called by the "execd" module managed by the qmaster of OGS/GE, and then performs the start or stop control of VMs on a computing node. In the OGS/GE, "*execd*" module usually calls "*shepherd*" module for deploying job processes in a computing node. In virtualized computational resource management in the proposed SDN-enhanced JMS framework, "*execd*" module calls *vm control* module instead of "*shepherd*" module for managing the behavior of VMs. Since this enhancement of the process flow to control job processes in a computing node is performed by utilizing a setting in the OGS/GE, the source code of OGS/GE is not changed.

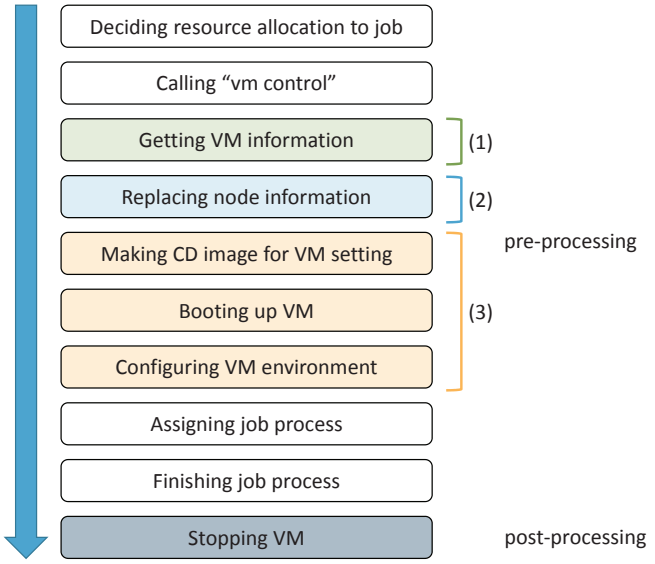
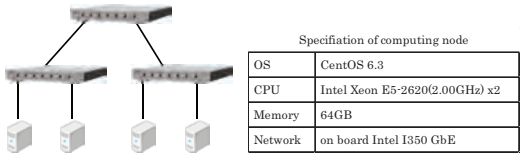


Fig. 4. Process flow of “*vm control*” module.



Specifiation of computing node

OS	CentOS 6.3
CPU	Intel Xeon E5-2620(2.00GHz) x2
Memory	64GB
Network	on board Intel I350 GbE

Fig. 5. Specification of cluster system for evaluation.

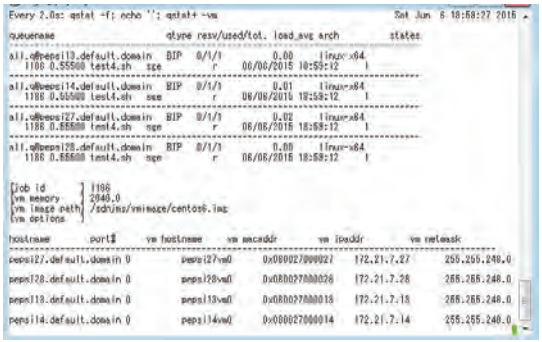


Fig. 6. Situation of resource allocation to a job.

The third functionality listed in section III-B is implemented as scripts included in the “*vm control*” module. The functions of the “*vm control*” module are composed of three steps: (1) Getting the VM information, (2) modifying a list of computing nodes allocated to the job, and (3) setting up an environment in a VM based on retrieved information. Figure 4 shows the process flow of the “*vm control*” module.

The step (1) is to get the hostname of VMs allocated to a job. In this implementation, the hostname of each VM is decided based on the one of allocating computing node. Since the VM hostname is not registered as the computing node of the HPC cluster system to the OGS/GE, the SDN-enhanced JMS framework cannot handle them. Thus, it is necessary to retrieve the hostname of all VMs allocated to the job. In the step (2), the list of allocating computing nodes generated by the OGS/GE is rewritten into the list of VMs. In the OGS/GE, the allocation of job processes is performed in accordance with the list of allocated computing nodes. In order to assign job processes to VMs by using the mechanism of the OGS/GE, replacing the list based on the hostname retrieved in the step (1) is required. The step (3) is to boot up allocated VMs on computing nodes and then reconfigure computing environment of the VM for executing a job process. In this step, a configuration process which is performed after a VM starting is prepared as a CD image with a script to set up it, and then the CD image is run in each VM starting. Moreover, since the CD image includes a script to execute “*shepherd*” module, the OGS/GE can assign a job process to the VM.

Regarding the fourth functionality listed in section III-B, the brain component of the SDN-enhanced JMS framework store the information of VMs allocated to each job into a database. Since a traditional JMS provides a command to show the status of resource allocation to jobs, a command to refer the information of VMs allocated to each job was implemented. Moreover, in this enhanced SDN-enhanced JMS framework, a user can also require the network resources through the same way.

IV. Evaluation

In this section, we mention the evaluation of the behavior of the proposed SDN-enhanced JMS framework with the mechanism to control virtualized computational resources in the developing environment. The evaluation environment is a cluster system which is composed of 4 computing nodes and 3 OpenFlow switches (NEC PF5240) as illustrated in Fig. 5

To observe the allocation of virtualized computational resource by the proposed SDN-enhanced JMS framework, we conducted the experiment in which multiple job with the different number of processes were submitted to the cluster system. In the experiment, as a requirement of virtualized computational resources, every jobs requested a same VM image, which consists of a single CPU, 2GB memory and CentOS 6.3. Moreover, a user’s home directory was mounted to the VM through Network File System (NFS) by the configuration performed by the “*vm control*” module.

Figure 6 shows the result of two commands to display the status of the allocation of physical and virtualized computational resources. The upper part of Fig. 6 is the result of *qstat* command, which is provided by the OGS/GE for displaying the usage of computing nodes. The lower part of Fig. 6 show the situation of the allocation of VMs by *qstat+* command, which is developed in this implementation of the fourth functionality described in Section III-C. From Fig. 6, it was confirmed that the proposed SDN-enhanced JMS framework makes it possible to assign job processes in user-requested VM like as the allocation to physical computational

TABLE I
OVERHEAD OF PROCESS TO CONTROL VIRTUALIZED COMPUTATIONAL RESOURCES.

Process Num	Pre-process (sec)	Post-process (sec)
1	1.299	0.004
2	1.315	0.004
3	1.140	0.004
4	1.196	0.004

resources. The proposed SDN-enhanced JMS framework allows a user to request either physical or virtualized computational resources to the HPC cluster system. Furthermore, a user can simultaneously require network resources by the functionality of the original SDN-enhanced JMS framework because a communication path between VMs is the same as a path between computing nodes in which the respective VMs are accommodated.

Next, we measured the overhead caused by the "*vm control*" module for controlling VMs on the proposed SDN-enhanced JMS framework. The overhead of the "*vm control*" module has the pre-process and the post-process as shown in Fig. 4: the process to boot up VMs on allocated computing nodes and then configure the computing environment in the VMs before the job execution starts, and shutting down the VMs allocated to the job after it finished. Since these processes increase a processing time to allocate resources compared with the original SDN-enhanced JMS framework, it is essential to evaluate the influence of overhead generated by the virtualized computational resource management. Table I shows the measurement of each process in the "*vm control*" module. The measurement results are classified according to the number of processes in jobs. From this result, it was confirmed that the additional resource allocation time to control the behavior of VMs is not much to influence efficient resource management. Generally, since the time required for calculating an appropriate set of resources is spent several tens of seconds, the effect to the system throughput is small. Moreover, since the number of processes did not affect the processing time of resource allocation, it is also considered to adopt the allocation of virtualized computational resources to the large-scale computation.

V. CONCLUSION

In the service of today's HPC center, to provide virtualized computational resources for a job like as Cloud Computing is required for handling various users' computational requests. In this paper, we proposed a mechanism of virtualized computational resource management on top of the SDN-enhanced JMS framework, which we have been studying and developing for efficiently and flexibly managing various resources and described the architecture of the novel SDN-enhanced JMS framework capable of handling the VM as virtualized computational resources. In the evaluation, we showed the behavior of virtualized computational resource allocation by the proposed SDN-enhanced JMS framework as well as physical computational resources.

As the future work, we will conduct the evaluation experiments on the larger computing environment and confirm the effectiveness of this resource management technology. Moreover, we will investigate more flexible network resource management by leveraging the functions of vSwitch.

ACKNOWLEDGMENTS

This research was supported in part by the collaborative research of the National Institute of Information and Communications Technology (NICT) and Osaka University (Research on High Functional Network Platform Technology for Large-scale Distributed Computing).

REFERENCES

[1] B. A. Kingsbury, *The Network Queuing System*, Sterling Software, Palo Alto, 1986.

[2] R. Henderson, "Job Scheduling under the Portable Batch System," in *Job Scheduling Strategies for Parallel Processing*, D. Feitelson and L. Rudolph, Eds. Springer, 1995, vol. 949, pp. 279–294.

[3] "Open Grid Scheduler: The official Open Source Grid Engine." [Online]. Available: http://gridscheduler.sourceforge.net/

[4] Y. Watashiba, S. Date, H. Abe, Y. Kido, K. Ichikawa, H. Yamanaka, E. Kawai, S. Shimojo, and H. Takemura, "Performance Characteristics of an SDN-enhanced Job Management System for Cluster Systems with Fat-tree Interconnect," in *Emerging Issues in Cloud (EIC) Workshop, The 6th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2014)*, December 2013, pp. 781–786.

[5] ——, "Efficacy Analysis of a SDN-enhanced Resource Management System through NAS Parallel Benchmarks," *The Review of Socionetwork Strategies*, vol. 8, no. 2, pp. 69–84, December 2014.

[6] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: Enabling Innovation in Campus Networks," *SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, Mar 2008.

[7] "Trema: Full-Stack OpenFlow Framework for Ruby/C." [Online]. Available: http://trema.github.com/trema/

[8] S. S. Laurent, E. Dumbill, and J. Johnston, *Programming Web Services with XML-RPC*. O'Reilly & Associates, Inc., 2001.

[9] B. P. Rimal, E. Choi, and I. Lumb, "A taxonomy and survey of cloud computing systems," in *2009 Fifth International Joint Conference on INC, IMS and IDC*, 2009, pp. 44–51.

[10] W.-T. Tsai, X. Sun, and J. Balasooriya, "Service-oriented cloud computing architecture," in *2010 Seventh International Conference on Information Technology: New Generations (ITNG)*, 2010, pp. 684–689.

[11] Y. Chen, T. Wo, and J. Li, "An efficient resource management system for on-line virtual cluster provision," in *Cloud Computing, 2009. CLOUD'09. IEEE International Conference on*. IEEE, 2009, pp. 72–79.

[12] X. Li, H. Palit, Y. S. Foo, and T. Hung, "Building an hpc-as-a-service toolkit for user-interactive hpc services in the cloud," in *Advanced Information Networking and Applications (WAINA), 2011 IEEE Workshops of International Conference on*, 2011, pp. 369–374.

[13] M. Taifi, A. Khreishah, and J. Y. Shi, "Building a Private HPC Cloud for Compute and Data-Intensive Applications," *the International Journal on Cloud Computing: Services and Architecture (IJCCSA)*, pp. 1–20, 2013.

[14] J.-L. Yu, C.-H. Choi, D.-S. Jin, J. R. Lee, and H.-J. Byun, "A Dynamic Virtual Machine Allocation Technique Using Network Resource Contention for a High-performance Virtualized Computing Cloud," *International Journal of Software Engineering & Its Applications*, vol. 8, no. 9, 2014.

Near Real-time Detection of Crisis Situations

Sylva Girtelschmid*, Andrea Salfinger†, Birgit Pröll‡, Werner Retschitzegger§, Wieland Schwinger¶
*‡ Inst. for Application Oriented Knowledge Processing, †§¶ Dept. of Cooperative Information Systems
Johannes Kepler University Linz
Altenbergerstr. 69, 4040 Linz, Austria
K0956640@students.jku.at*, {andrea.salfinger†, werner.retschitzegger§, wieland.schwinger¶}@cis.jku.at, bproell@faw.jku.at‡

***Abstract*—When disaster strikes, be it natural or man-made, the immediacy of notifying emergency professionals is critical to be able to best initiate a helping response. As has social media become ubiquitous in the recent years, so have affected citizens become fast reporters of an incident. However, wanting to exploit such 'citizen sensors' for identifying a crisis situation comes at a price of having to sort, in near real-time, through vast amounts of mostly unrelated, and highly unstructured information exchanged among individuals around the world. Identifying bursts in conversations can, however, decrease the burden by pinpointing an event of potential interest. Still, the vastness of information keeps the computational requirements for such procedures, even if optimized, too high for a non-distributed approach. This is where currently emerging, real-time focused distributed processing systems may excel. This paper elaborates on the possible practices, caveats, and recommendations for engineering a cloud-centric application on one such system. We used the distributed real-time computation system Apache Storm and its Trident API in conjunction with detecting crisis situations by identifying bursts in a streamed Twitter communication. We contribute a system architecture for the suggested application, and a high level description of its components' implementation.**

I. Introduction

During the last decade, social media channels have evolved to the point of becoming omnipresent in our everyday lives. Platforms, such a Twitter, play a vital role in sharing information fast. It should come to no surprise that the Twitter data pool is popular for information mining as it can reveal valuable insights, be it for crisis monitoring applications, or marketing, to monitor the perception of a new product. Moreover, being able to utilize Twitter data to promptly detect a situation that requires a responsive action from a rescue crew, e.g. during natural disasters, can often save lives. However, even the situation detection alone is non-trivial. For a human, processing Twitter conversations for this purpose would mean having to look into the content of each message and extract the interesting information from it. A more efficient approach, better suited for a computer, is to cluster similar posts without having to first consider the semantics of the content. To satisfy the real-time detection demands, techniques for bursty topics identification are applicable here.

Bursts are in this context collections of posts mentioning the same topic more often within the current period than within the previous one. The problem of detecting burst topics falls under a broader category of research dealing with Topic Detection and Tracking (TDT) that has been extensively addressed in academic articles.

Of the various TDT approaches, clustering is a popular one to identify bursts. Other methods for burst detection include finite state automata, Fourier transform, time series, or Wavelet transform [1]–[3]. For our first setup, the clustering approach appealed the most to us thanks to its inherent property of being readily parallelizeable.

However, the enormous throughput of the real-world streams of Twitter posts prohibits timely online processing of a sequential implementation of state-of-the-art methods for topic clustering. For an acceptable performance, this problem requires single-pass, and optimized methods, as well as scalable implementation. In the recent past there has been work on successfully applying Cloud technologies to the online clustering problem to guarantee scalability, and near real-time processing of streaming data [4]–[6]. In this paper we show that Cloud technologies have a potential in improving responsiveness during emergency situations through mining social media content.

Since our system also needs to work with persistent data, we decided to make use of the Trident API, a high-level abstraction of Apache Storm, as it makes stateful processing more manageable than Storm alone. Although Trident is gaining popularity, to the best of our knowledge, we are not aware of any other systems that employ Trident for the use of Burst Topic Detection. In our work, we evaluate Trident's applicability to detecting an outburst of a crisis situation in near real-time. We propose an architecture for such an online disaster detector system and contribute a high level description of a Trident topology implementation.

The rest of this paper is organized as follows: In the next section we discuss a number of research areas related to our work and identify the specific ideas from which our devised system borrows. In sec. III, we detail our approach to detecting new emergency situations from Twitter data streams. Sec. IV discusses our test cases. Finally, in sec. V, we conclude on our findings, and provide an outlook on future work.

II. Related Work

Due to the multi-disciplinarity of the envisioned application domain, related approaches, and valuable preparatory work need to be drawn from several areas. Techniques essential to address the requirements imposed by the crisis management application domain can be found in the research fields of Event Detection, First Story Detection, Burst Detection, Knowledge

Bases, Keyword and Topic Extraction, and Parallelization, which we will motivate and discuss in the following.

a) Event Detection: Finding a common topic within a set of documents has been covered extensively in research (TDT initiatives). In the subject of crisis monitoring, this research also finds its application, as event/story detection is in this subject's heart. There are various approaches to this problem: e.g. in [7], [8], the method of clustering documents based on word similarity using the Vector Space model (VSM) is presented. In [9], the clustering of the documents is improved by considering the locality information in the process of discriminating the events into the clusters. In [10], the similarity score calculation also incorporates the social network structure besides the content of the message. Yet another class of solutions for event detection, and tracking bases on probabilistic methods, such as presented in [11]–[13]. There, the actually encountered message density w.r.t. the specific topics is compared to an expected density. A valuable work summarizing research on event detection in the realm of Twitter can be found in [14].

b) Detection of Novel Events: Although organizing messages into clusters brings a useful insight to our work, it is not sufficient for our problem solution. We have to consider methods capable of identifying those topics that have not been discussed before within some long enough time period. In other words, we need to identify posts that are discussing a new story. First Story Detection (FSD) methods are therefore well applicable here, as they are designed to detect when a document discusses a previously unseen content. Examples of FSD implementation are presented in the works of [15], [16]. The authors propose an optimized approach to FSD, which makes it well suited for identifying events online, i.e. for identifying events from real-time streaming text such as tweets. The ideas set out in these works form the basis of our new event detection algorithm.

c) Identifying Bursts: Typically, whenever a newsworthy event occurs, many people tend to share information about it on social media. This causes a temporal burst of closely related messages which can be captured [1], [17]. Detecting bursts has also been well studied [1]–[3], [18]. For example, in [3], the author achieves real-time event detection by clustering wavelet-based signals. Wavelet transformation technique is applied to build signals for individual words, which are cross correlated and the signals are then clustered using a scalable eigenvalue algorithm. In our approach, we use a bucketing method, similar to that proposed by [15], [16], since it directly fits the parallel programming paradigm we adopted.

d) Knowledge-based Sensing: In our system, we also need to consider techniques that allow us to report only events which we are interested in. In this context, an event is defined as a topic that suddenly draws the attention of the public. As such, it may often be of no value for our purposes, since, for example, news about celebrity deaths are also causing abrupt increase in related posts being sent. Our application needs to be able to identify only those events that are related to disasters. To achieve this, a disaster ontology may be employed. There have been many attempts in building an ontology related to disaster. It is, however, often the case that existing disaster ontologies focus only on a specific type of disaster, which they explore and describe in detail [19], [20]. For our purposes, it is sufficient to have a high level disaster dictionary to identify the particular disaster situation referred to in the tweets. Working with multiple languages would, therefore, be also feasible.

e) Extracting Keywords: Document keyword and keyphrase extraction is another complex problem addressed in the research [21]. The frequent techniques include word frequency analysis, distance between words, or lexical chains to rank the keywords. When it comes to keyphrase extraction, graph-based ranking methods are being successfully used. However, such methods are proposed primarily for a single document or a document collection. This is not directly applicable when it comes to extracting representative words from such a short document as a tweet. The large variety of topics found in tweets also makes this task more challenging. In [22], the authors propose a method specifically for keyword extraction from Twitter. The approach is based on organizing keyphrases by topics learnt from Twitter. For our purposes, the best approach is to apply a content-sensitive keyword extraction, as our tweet collection from which we need to extract the keywords is already formed of similar content.

f) Meeting Real-time Demands: The algorithms that can achieve our task of detecting a new story from a real-time stream of data are comparatively computationally expensive. To safely achieve real-time response of a system processing the full streaming data flow from Twitter (the so called 'firehose'), parallelization of the algorithm is necessary. Some recent studies have shown that the implementation of clustering algorithms on the Storm distributed framework is reasonable [4], [5]. In our implementation, however, we utilize the higher level Storm API called Trident. This design benefits from the guarantee of exactly-once semantics[1], as well as the ability to process streaming messages in batches which performs better when querying a database system.

III. Implementation

Extracting useful information from Twitter data flow is no doubt a challenge. The posts are not only short (140 characters), and varied in topic but also highly noisy, in that most of the conversations contain a lot of useless babble. The task of identifying bursts offers itself as the best approach to finding information of value as more people will get drawn into a conversation about something vital than about a certain individual reporting on currently drinking the most delicious cacao.

In this project, we focus on utilizing one of the FSD technique which applies a Vector Space model and is optimized using Locality Sensitive Hashing. This technique is described in detail in [16], and we provide a brief overview

[1]During a node failure, only the messages that have not been fully processed will get resent. This is as opposed to Storm's at-least-once semantics, where some messages may get processed twice.

in subsection III-B1. This FSD algorithm outputs a similarity score for each incoming message together with the ID of the message to which it is the most similar. The next stage involves the actual clustering, or bucketing to identify bursts. This is performed based on the message's content similarity score and monitoring of the growth rate of the buckets in which similar tweets are gathered. The components of our system are described in sec. III-B.

A. Distributed Platform

Besides Spark Streaming[2], the Apache Storm Trident API is currently the best suited framework for our application. It simplifies the implementation of parallel tasks by providing a programming interface suitable for stream processing and continuous computation[3].

As already mentioned in sec. I, the successful use of Storm Apache in data stream clustering applications has recently been reported [4], [5]. The advantages of using the Trident API over the core Storm API in our application are threefold: First, Trident can guarantee exactly-once semantics as opposed to at-least-once semantics of Storm. This means, that we don't have to worry about already processed messages being resend in case of a failure in the computing cluster and therefore we will not be faced with potential erroneous bursty event reports.

Another advantage of using Trident is the fact that it handles streams of batches of tuples as opposed to streams of tuples. This achieves better performance for communication with a database.

Lastly, state persistence handling is incorporated generically in the Trident API which allows us to be flexible in the selection of our back-end technology. On the other hand, using Trident requires some additional checks for our algorithm, which we point out at the end of subsection III-B2.

B. System Components

Fig. 1 shows a component overview of the proposed system. In our setup, the topology is fed by data from a KafkaSpout, where the source of the stream is 1% of the Twitter firehose accessed through a Hosebird client[4].

The first component of our system implements the core parts of the open source project "First Story Detection on Twitter using Storm"[5] and adopts it for running on an actual distributed cluster with real-life data streams. For later queries, tweets are saved in the NoSQL distributed storage mechanism Apache Cassandra. The next component takes care of grouping of related tweets into buckets, which are monitored for bursts based on their growth rate. References to the bulk of tweets from the fastest growing bucket, i.e. the tweets form the burst, are also made persistent. The task of the third component is to decide whether the topic of the burst is of interest. The final component executes only if a crisis related new event was detected. Its task is to notify responsible operators from the area of the event occurrence, and to automatically spawn a new instance of a Twitter search tracking this event.

1) FSD using Locality Sensitive Hashing: In this section, we briefly explain the gist of the FSD component.[6] This component outputs additional information for each incoming message: the Twitter ID of the closest neighbour message (the most similar post), and a score of similarity in the range of $[0.0 - 1.0]$, where 1.0 identifies an identical post.

Every newly arriving message from the Twitter data stream is split into words and a number of preprocessing steps are applied (e.g. removal of URLs and mentions, replacement of some, often intentionally, misspelled words, as well as simplistic[7] word stemming). The cleaned up corpus is then submitted for calculation of the nearest neighbour.

First, this process involves determining the TF-IDF (Term-Frequency - Inverse Document Frequency) weighting for each term in the message to convert the representation of the tweet message into a vector normalized by Euclidean norm. The new vector must then be compared to the vectors of the preceding messages. An approximate near-neighbour among the seen documents can be found fast using the Locality Sensitive Hashing algorithm. It works on the assumption that similar tweets tent to hash to the same value. This allows for an efficient optimization where the number of documents to which comparison has to be made is greatly reduced. Namely, it uses hash tables to bucket similar tweets so that each incoming tweet will be compared with only the tweets that have the same hash.Finally, cosine similarly measure is applied to compute the distance of the nearest neighbour from the incoming tweet. If the calculated distance to the closest tweet is below a predefined threshold, this algorithm also uses an additional step of comparing the distance to a fixed number of latest tweets. This step alleviates the problem of possibly overlooking a closer tweet posted in the immediate time proximity. Such problem can easily arise given the method's randomness of selecting tweets for comparison in the first place.

2) Bucketing and Identifying bursts: In effect, the process of grouping the posts that are similar also identifies a new event. In other words, if an incoming message was found to have a low similarity score, this message will be marked as a potential new event (by placing it in a new bucket). Then, given that there will be enough related/similar posts following it, this event will grow in its own cluster. In [16], this process is referred to as threading, whereas in [10] it is called the cluster summary. We chose to describe this process in two subtasks - one of bucketing and the other of identification of the bursts.

The first task of this component is to gather similar tweets in the same bucket. Before bucketing, tweets whose similarity score is unfavorable (too low) are filtered out. I.e. if $(1 - cosineDistance) < threshold$ the tweet is included in

[2]http://spark.apache.org/streaming/

[3]http://storm.apache.org/

[4]A robust Java HTTP library for consuming Twitter's Streaming API

[5]https://github.com/mvogiatzis/first-stories-twitter

[6]For detailed explanation, refer to the original author's website at http://micvog.com/2013/09/08/storm-first-story-detection/

[7]For performance reasons, we only apply stemming to English words that have a common stem in all their inflected variants

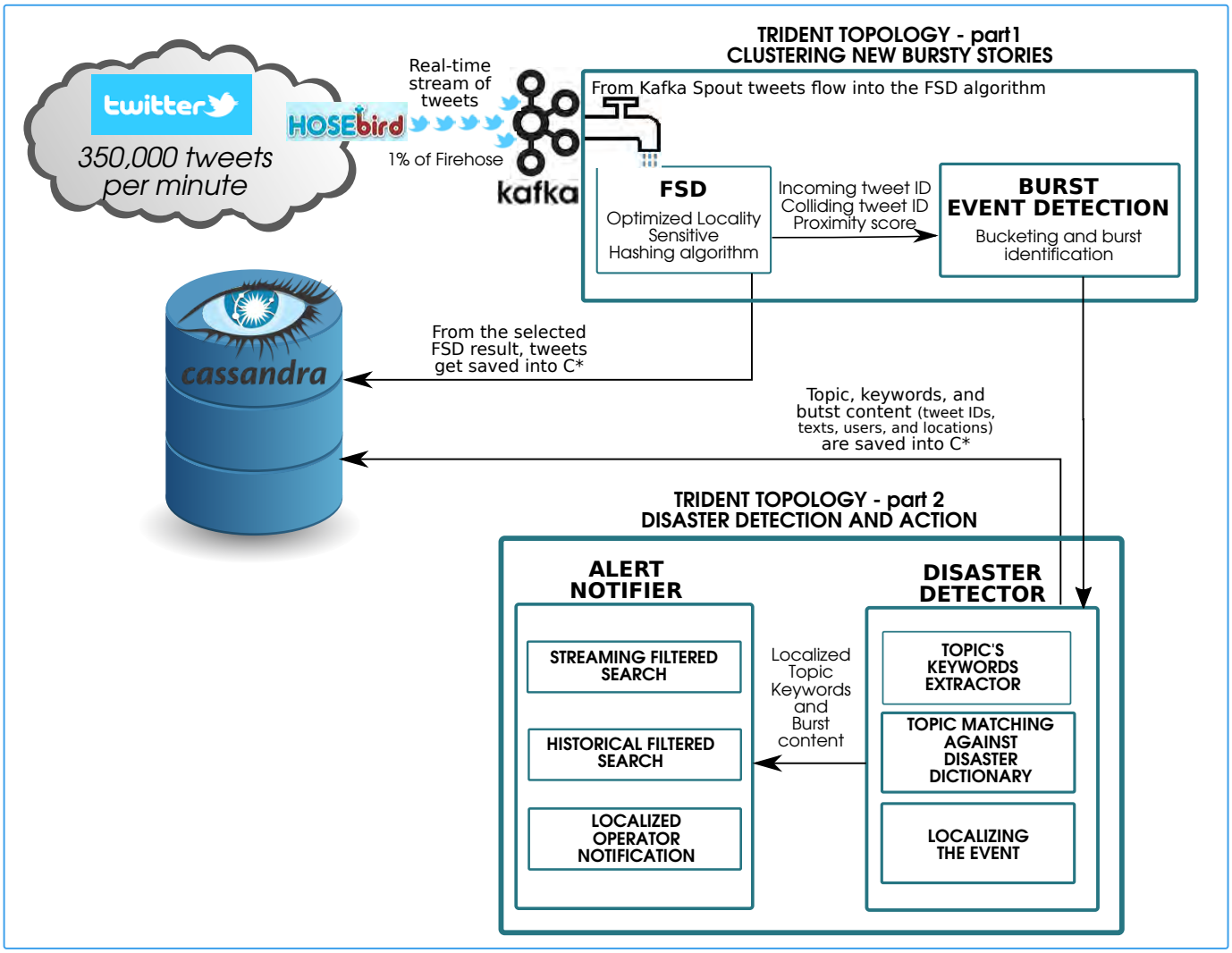


Figure 1: System architecture overview

further processing. The predefined threshold was found to give the best result if kept in the range of $[0.5, 0.6]$. As explained in [16], higher threshold values cause the topics in a bucket to be diverse and plentiful, while lower threshold makes the topics very specific and scarce. A passing tweet will then be placed either in a new bucket, if its nearest neighbour is not already in one of the existing buckets, or it will join the already existing cluster of similar tweets in their bucket. If it ends up in a new bucket, it is considered to potentially contribute to a new story discussing a previously unseen content. The number of buckets is finite, so whenever all buckets are occupied and a new story tweet streams in, the bucket that has the lowest timestamp of its most recent content update is freed up for reuse. Also the size of a bucket is limited. Whenever a new tweet is determined to belong to a full bucket, we simply remove the older half of the bucket to free up space. The assumption is that the tweets that the removed IDs refer to are not of importance (too far apart to be able to form a burst) since otherwise they would have been already reported as a burst.

An important Trident implementation detail lies in the necessity to first simulate the filling of the buckets. This is because Trident works on batches and, therefore, the state of the buckets gets updated only at the end of a batch. Let us consider the case when a new batch contains a new event as its first tweet, and the tweets following it are related only to that first tweet. Since that tweet is not placed in the bucket before the next tweet is processed, the next tweet would also be (incorrectly) identified as a new event and a new bucket would be assigned to it. This is a clear disadvantage of using batches in our algorithm, however, it pays off later when communication with a database system is required.

Finally, once in a while a fastest growing bucket will be detected. Before passing the burst content down the processing pipeline, the Burst Event Detector determines whether this bulk of tweets could be considered a burst by checking the time span within which the tweets were posted. The burst content is then consumed by the Disaster Detector component.

3) Disaster Detector: The task of the Disaster Detector is to find whether the captured event is an emergency situation and, if it is, where it is located. First, the keywords must be extracted and based on them, utilizing our disaster type dictionary, we identify the topic. Finally, if the topic matches a disaster situation, we localize it. At this stage, database query is necessary to access the tweet.

a) Extracting keywords: As keywords, we wish to find five words that best summarize the content of the burst. Our approach to finding them involves calculating TF within the burst. In other words, we only weigh each term positively for the number of times it occurs within the burst. The disaster dictionary is meant to contain words representative of a specific disasters. For our testing purposes, we populated our dictionary by representative keywords of a hurricane and snowstorm disasters. The datasets were compiled from querying Twitter's historical stream for the specific disaster types during their known occurrences.

b) Identifying disaster type: The topic is identified by matching any of the keywords to the part of the dictionary describing a given (predefined) topic. As an additional check, needed to prevent overlapping terms causing us to select the wrong topic, we weigh the keywords negatively relative to the number of times it occurs in the dictionary entry for other topics. The entries for topic, keywords, IDs of the contributing tweets and their text, location (if given) and users' information are saved to Cassandra. The database can be queried by other topologies (possibly spawned later to do retrospective analysis on the tweets) without affecting the performance of our system.

c) Geo-localizing the event: In order to inform only those crisis response agencies that are directly concerned with the occurring crisis situation, we need to be able to determine where the event is happening. Spacial grounding of the tweets in the burst requires to consider not only the coordinates, if given, but also the user's location as well as the places mentioned directly in the text of the tweet. This non-trivial task is carried out by the Geo-Tagger subcomponent we describe in our previous work [23].

4) Alert Component: Once a disaster event is detected and its location, type and representative keywords are determined, we pass the information to the Alert component. The keywords are used to start up new data collection sessions from Twitter both from recent history as well as from a real-time stream. Finally, the responsible operators can be notified who, for the duration of the disaster, are assumed to administer searches also from other social media channels using our CrowdSA application presented in [23]–[25].

IV. Results and Evaluation

Due to the difficulty of determining the "ground truth" and the appropriate metrics, a fully functional evaluation of our system (the correct and near real-time detection of a crisis situation) is beyond the scope of the present work. However, we devised three types of tests to show that our proposal of implementing the detection of crisis situations from Twitter content in Trident is reasonable. All tests carried out are run in an unsupervised mode, where we do not assume to know a priori the type of event being detected. The data collected for this purpose are from hurricane Iselle reaching Hawaii in August 2014, snowstorm affecting the US East cost in January 2016, and hurricane Patricia hitting Mexico in autumn 2015.

In the first test case, we published our collected historical data to the Kafka queue and measured the processing speed in terms of the number of tweets processed in a second. We were able to reach an average (for the three different datasets) processing speed of about 4000 messages per second. This is below what the full Twitter firehose could serve (on average, there are about 6000 messages posted on Twitter in one second[8]). However, after increasing the parallelization and utilizing a larger cluster, the performance can be boosted to surpass the firehose requirements.

In the second testing scenario, we evaluated the effectiveness of the burst detection capability and the keyword extraction by consuming the historical datasets from the hurricane Iselle. Our system could detect a number of bursts within the datasets occurring relative to the time of the day. Most importantly, the largest burst was detected on the 10th of August, which corresponds to our graphical inspection of the dataset.

For the third test, we fed the system by real-time Twitter streaming data. These account for about 1% of the Twitter Firehose. Since it was not expected to encounter any hurricane or snowstorm disaster during the run of the test, we let the system report on all bursts which it found, essentially skipping the Disaster Detector component. We observed that even with only six nodes, the system was performant in that it processed tweets fast enough without getting congested and falling behind.

All of the above tests were run on a six node virtual cluster with 8 cores each, 64 bit CentOS, and 16GB of RAM. The cluster was provisioned using Ambari[9] and runs Storm, Kafka broker, and a Cassandra server. The 1st node runs Nimbus (Storms master daemon) while the rest are the worker nodes (running Storm's Supervisors). Greater parallelization within Storm was selectively applied for intensive tasks such as the dot product calculation in similarity estimations. As a future evaluation strategy, we plan to experiment with different parallelization setups within Storm while processing the same data stream of tweets to better understand the capabilities of the distributed framework.

V. Conclusion and Future Work

In this paper, we reported on our work that focused on the problem of near real-time (online) Burst Topic Detection from streaming Twitter data in application to detecting newly occurring crisis situations. Our approach involves clustering tweet messages based on content similarity and implementing the algorithms in a parallel fashion using the Apache Storm

[8]http://www.internetlivestats.com/one-second/

[9]https://ambari.apache.org/

Trident API. Our results have shown that the use of First Story Detection in combination with capturing temporal bursts of similar messages streamed from a Twitter firehose, and mapping the captured events to a predefined disaster dictionary enables fast reporting of a crisis situation when implemented in a distributed fashion.

As a future work, we are interested in looking at the potential improvements if the structure of the social network is also considered in the similarity measurements. Additionally, we plan to compare the results of the unsupervised event detection (as currently performed by our setup) with a supervised approach in which we would use our dictionary to filter tweets before they are passed to the clustering component. Last but not least, we would like to experiment with incorporating a more sophisticated model for filtering out the tweets before bucketing. Namely, since in the case of a large burst, it might be more effective to also include the very near tweets in the processing, we'd like to use a dynamic threshold value updated based on the current state of the stream.

ACKNOWLEDGMENT

The authors would like to thank Matthias Steinbauer and the Telecooperation Institute of JKU Linz for providing cloud computing resources and support during provisioning our cluster.

This work has been funded by the Austrian Federal Ministry of Transport, Innovation and Technology (BMVIT) under grant FFG BRIDGE 838526 and under grant AD WTZ AR10/2015.

REFERENCES

[1] J. Kleinberg, "Bursty and hierarchical structure in streams," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, (New York, NY, USA), pp. 91–101, ACM, 2002.

[2] Q. He, K. Chang, and E.-P. Lim, "Analyzing feature trajectories for event detection," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, (New York, NY, USA), pp. 207–214, ACM, 2007.

[3] J. Weng and B.-S. Lee, "Event detection in twitter.," *ICWSM*, vol. 11, pp. 401–408, 2011.

[4] X. Gao, E. Ferrara, and J. Qiu, "Parallel clustering of high-dimensional social media data streams," in *Cluster, Cloud and Grid Computing (CCGrid), 2015 15th IEEE/ACM International Symposium on*, pp. 323–332, May 2015.

[5] G. Wu, O. Boydell, and P. Cunningham, "High-throughput, web-scale data stream clustering," in *Proceedings of the 4th Web Search Click Data workshop (WSCD 2014)*, 2014.

[6] K. XIANGSHENG, "Microblog mining based on cloud computing technologies: Mesos and hadoop.," *Journal of Theoretical & Applied Information Technology*, vol. 48, no. 3, 2013.

[7] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power, "Using social media to enhance emergency situation awareness," *Intelligent Systems, IEEE*, vol. 27, pp. 52–59, Nov 2012.

[8] J. Rogstadius, M. Vukovic, C. A. Teixeira, V. Kostakos, E. Karapanos, and J. A. Laredo, "Crisistracker: Crowdsourced social media curation for disaster awareness," *IBM J. Res. Dev.*, vol. 57, pp. 1:4–1:4, Sept. 2013.

[9] M. Nagarajan, K. Gomadam, A. P. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav, *Web Information Systems Engineering - WISE 2009: 10th International Confee rence, Poznań, Poland, October 5-7, 2009. Proceedings*, ch. Spatio-Temporal-Thematic Analysis of Citizen Sensor Data: Challenges ann d Experiences, pp. 539–553. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[10] C. C. Aggarwal and K. Subbian, "Event detection in social streams," in *In Proceeding of the Twelfth SIAM International Conference on Data Mining*, (Anaheim, California, USA), pp. 624–635, SIAM / Omnipress, April 26-28 2012.

[11] H. Smid, P. Mast, M. Tromp, A. Winterboer, and V. Evers, "Canary in a coal mine: Monitoring air quality and detecting environmental incidents by harvesting twitter," in *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, (New York, NY, USA), pp. 1855–1860, ACM, 2011.

[12] T. Sakaki, F. Toriumi, and Y. Matsuo, "Tweet trend analysis in an emergency situation," in *Proceedings of the Special Workshop on Internet and Disasters*, SWID '11, (New York, NY, USA), pp. 3:1–3:8, ACM, 2011.

[13] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), pp. 851–860, ACM, 2010.

[14] F. Atefeh and W. Khreich, "A survey of techniques for event detection in twitter," *Comput. Intell.*, vol. 31, pp. 132–164, Feb. 2015.

[15] H. Becker, M. Naaman, and L. Gravano, "Learning similarity metrics for event identification in social media," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pp. 291–300, ACM, 2010.

[16] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to twitter," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pp. 181–189, Association for Computational Linguistics, 2010.

[17] W. Xie, F. Zhu, J. Jiang, E.-P. Lim, and K. Wang, "Topicsketch: Real-time bursty topic detection from twitter," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pp. 837–846, Dec 2013.

[18] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, "Finding bursty topics from microblogs," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, (Stroudsburg, PA, USA), pp. 536–544, Association for Computational Linguistics, 2012.

[19] S. Liu, D. Shaw, and C. Brewster, "Ontologies for crisis management: a review of state of the art in ontology design and usability," in *Proceedings of the Information Systems for Crisis Response and Management conference (ISCRAM 2013 12-15 May, 2013)*, 2013.

[20] D. De Wrachien, J. Garrido, S. Mambretti, and I. Requena, "Ontology for flood management: a proposal," *Flood Recovery, Innovation and Response III*, vol. 159, p. 3, 2012.

[21] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," *International Journal on Artificial Intelligence Tools*, vol. 13, no. 01, pp. 157–169, 2004.

[22] W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li, "Topical keyphrase extraction from twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pp. 379–388, Association for Computational Linguistics, 2011.

[23] A. Salfinger, W. Retschitzegger, W. Schwinger, and B. Pröll, "Crowd sa – towards adaptive and situation-driven crowd-sensing for disaster situation awareness," in *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2015 IEEE International Inter-Disciplinary Conference on*, pp. 14–20, IEEE, 2015.

[24] A. Salfinger, S. Girtelschmid, B. Pröll, W. Retschitzegger, and W. Schwinger, "Crowd-sensing meets situation awareness: A research roadmap for crisis management," in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, pp. 153–162, IEEE, 2015.

[25] A. Salfinger, W. Retschitzegger, W. Schwinger, and B. Pröll, "Mining the disaster hotspots – situation-adaptive crowd knowledge extraction for crisis management," in *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2016 IEEE International Inter-Disciplinary Conference on, in press*, 2016.

Automatic protocol based intervention plan analysis in healthcare

Miklos Kozlovszky[1,2], Levente Kovács[3], Khulan Batbayar[1], Zoltán Garaguly[1]
[1] Biotech Knowledge Center/Obuda University, Budapest, Hungary
[2] MTA SZTAKI/Laboratory of Parallel and Distributed Computing, Budapest, Hungary
[3] Physiological Controls Group/Obuda University, Budapest, Hungary

{kozlovszky.miklos@nik, kovacs.levente@nik, khulan batbayar@biotech,garaguly.zoltan@biotech}.uni-obuda.hu,

Abstract - Evidence and protocol based medicine decreases the complexity and in the same time also standardizes the healing process. Intervention descriptions moderately open for the public, and they differ more or less at every medical service provider. Normally patients are not much familiar about the steps of the intervention process. There is a certain need expressed by patients to view the whole healing process through intervention plans, thus they can prepare themselves in advance to the coming medical interventions. Intervention plan tracking is a game changer for practitioners too, so they can follow the clinical pathway of the patients, and can receive objective feedbacks from various sources about the impact of the services. Resource planning (with time, cost and other important parameters) and resource pre-allocation became feasible tasks in the healthcare sector. The evolution of consensus protocols developed by medical professionals and practitioners requires accurate measurement of the difference between plans and real world scenarios. To support these comparisons we have developed the Intervention Process Analyzer and Explorer software solution. This software solution enables practitioners and healthcare managers to review in an objective way the effectiveness of interventions targeted at health care professionals and aimed at improving the process of care and patient outcomes.

Keywords

health care intervention, process analyzer

I. INTRODUCTION

A growing demand can be seen in the healthcare sector for value-added, premium category medical services, which involves continuous medical monitoring and care.

Another aspect within Europe is the increasing tendency of patient tourism and the ever growing number of medical services for foreigners at premium service providers. Interventions at a premium medical service provider has always well defined parameters (e.g.: cost, duration, etc.) and the whole intervention plan of the patients can be easily represented by timed graph structures. Intervention plan is part of the clinical pathway and in larger scale it is an important part of the patient's life path.

An intervention plan is basically builds up from generic consensus protocol(s), additionally it is always patient-centered thus it contains some patient related personalized parameter set as well. If we could compare the planned interventions (intervention plan) with the occurred interventions we could objectively assess the differences in medical services, analyze their impact, the effectiveness of the core consensus protocol, and the real usage scenarios of the underlying protocol. For such motivation we have created an Intervention Process Analyzer and Explorer software solution, which is automatically able to assess deviations from the intervention plan in an objective way. Our paper builds up as follows: In the first section we give an overview about our term definitions, in the next section we detail the pre-defined requirements of the solution and the identified data sources. This will be followed by an overview about the intervention plan analysis (about the analysis levels, and the used methods), then we provide a short descript about the internal architecture of the solution and finally summarize our results.

II. STATE-OF-THE-ART

In recent years Electronic Medical Record and medical billing process is merging together, and provides solid foundation for more effective tracking of patients, medical services and interventions. Patient Health Records (PHR) are evolving into digital format. HIMSS Analytics developed [13] the Electronic Medical Record Adoption Model (EMRAM) in 2005 as a methodology for evaluating the progress and impact of electronic medical record systems for hospitals and hospital ambulatory facilities. In this model eight stages (0-7) has been identified, that measures a hospital's implementation and utilization of digital information technology applications. The Electronic Health Record (EHR) is basically a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. It contains information about patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and various reports.

Electronic Health Records can be stored at medical service providers or accessed online by public [10, 11, 12]. Workflows are invading many medical fields. For example surgical workflows and their analysis is nowadays a hot topic, thus procedure modeling, workflow optimization, and skill assessment can provide more effective and precise procedures in the operating room (OR). Other medical workflows, treatment plans are usually accessible only as consensus protocols, however the real implementation of such protocols (orders, used materials, drugs, timelines, etc.) differs significantly from hospital to hospital. Some generic solutions are available both at the academic side (quantitative and qualitative workflow and EHR system evaluations), and at the commercial side (e.g.: the HealthCloud from Salesforce [8], and cloud based Practice Management systems).
A reliable and robust coding scheme and naming convention system is inevitable to build up future proof intervention plan databases. Recently there are still problems in the healthcare sector:

- coding schemes differs at country level (here we should note that most of the codes based on ICD-x published by WHO [14])
- coding schemes are by default rigid structures, however due to the new medical interventions their scope (list) is always growing (WHO's most recent version is ICD-10, ICD-10 was endorsed by the 43th World Health Assembly in May 1990 and came into use in WHO Member States as from 1994. ICD is currently under revision, and planned release date for ICD-11 is 2018.
- new national level coding and naming convention scheme versions are in many cases not backward compatible ,
- different codes can be mapped to the same element.

A. *Used term definitions*

In the followings we describe shortly the most important basic terms for our paper defined/adapted from literature [1, 4, 5]:

- Protocol: generic set of interventions (events and activities), and rules of a certain healthcare domain for a well defined group of patients, which was developed as a consensus agreement by a team of experts. It can have a constant evolution over time (can have version, can be expired after a well defined period of time), it based on the best practices of the healthcare professionals (and also on patient expectations) and the recorded patient statistics. It contains a set of events and activities with time and spatial constraints. Its internal structure can be represented as formal process description or a graph (protocol graph), which can contain iterations, conditional alternative paths.
- Intervention plan: Belongs to a single patient and contains events, activities, medical services with time, location, and resource parameters in a personalized manner. It can contain multiple protocols and also can contain complex control patterns (such as iterations, conditions, etc.)
- Intervention plan graph: Visualization of the intervention plan as a directed graph.
- Clinical pathway: Contains a set of events and activities with time and spatial constraints described in minimum one (personalized) intervention plan.
- Realized interventions: Multisource dataset, which belong to a single patient, and contains event, activity and medical service logs, with time, location and patient data and various resource parameters. It is a clear reflection what was historically occurred with the patient during its clinical pathway.

III. AIMS AND REQUIREMENTS

Premium medical service providers are only viable, if they able to run their "business" with high efficiency. The market filters out all the medical service providers, which are not able to assess their potential and the quality of their services. These data can only collected with accountable, objective measurements, and continuous high resolution service monitoring. On the patient side new requirements appeared, such as pre-evaluation and interactive monitoring of health services and intervention modeling/virtualized service models with high accuracy. Our Intervention Process Analyzer and Explorer solution provides effective solution to do comparative analysis between the intervention plans and the occurred interventions, assess the difference and provide large scale statistics about the frequently used intervention scenarios. As identified requirements the realized solution should support combination with external healthcare databases and data mining applications to reveal more aspects of healthcare services, and their effectiveness on patient health. The analysis process should support offline running mode, totally independent from any interventions, without any intervention plan status restrictions (if the targeted intervention plan is not in a finished or closed state).

IV. DATA SOURCES

Our software solution collects information from various data sources, and stores these data in a semi structured data repository for further processing. The simplified system overview is shown in Figure 1.

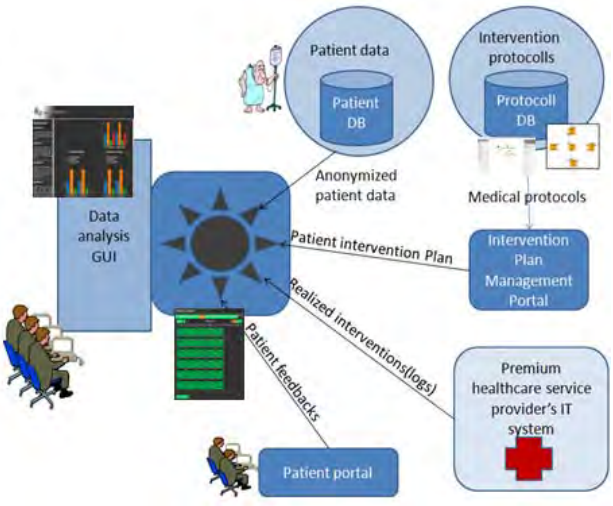


Figure 1. System overview

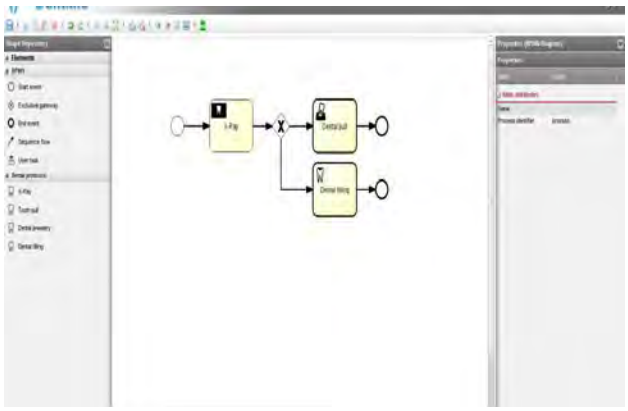


Figure 2. Intervention plan editor and visualizer GUI

The collected main data sets are the following:

- Patient data
  - Patient base data
  - Patient survey data (satisfaction survey, etc.)
  - Anamnesis data
- Healthcare service data
- Historical logs received from the medical service provider or any healthcare systems
- Protocol data
- Intervention plan data
- Realized intervention data (from the healthcare system)

V. BUILDING UP AND ANALYSIS OF INTERVENTION PLANS

On the intervention plan editor GUI (shown in Figure 2.) the intervention plan can be described as (an arbitrary complex) workflow, where the small circles are representing start/stop events, X represents alternative pathways with conditions, large boxes with labels and small icons are representing pre-defined intervention processes, and arcs define the direction of the path.

Intervention plan analysis can be done from many viewpoints. In the GUI different user groups (healthcare manager, medical expert/practitioner) can do analysis both on individual, or on a large set of intervention plans.

A. *Analysis process*

Two types of analysis have been identified: simple analysis (individual intervention plan vs. realized intervention set of a single patient) and population scale analysis (complex, multi-parameter level intervention plan comparative analysis of a set of patients) shown in Figure 3.:

- Simple analysis compares the intervention plan's graph structure with almost 150 different, pre-defined performance indicators to the occurred interventions (and the measured parameters). We try not only to compare the parameters, but assess the impact of the difference
- Population scale analysis can be a handy tool for practitioners and healthcare managers to create statistical analysis about protocol usage, average statistical parameters and expected patient outcomes.

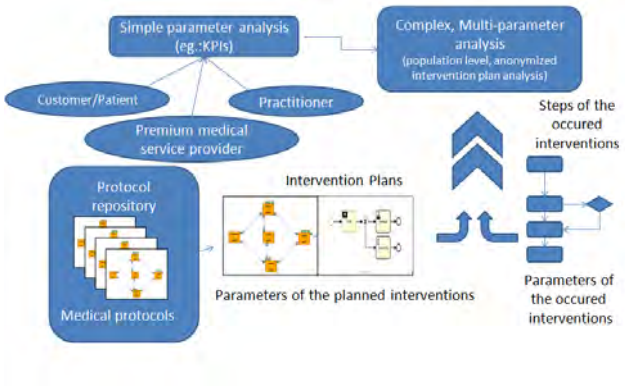


Figure 3. Intervention plan vs. realized intervention comparison

B. *Analysed parameters*

We have identified a large set of intervention task parameters (*Pi*), which support during evaluation to

objectively measure how the plans are matching the real world:

- Sequence of the intervention steps
- Amount of intervention steps
- Intervention step sub-parameters
- Conditional path decision accuracy
- Iterations
- Service resource consumption

Logically, an intervention step parameter can be arbitrary complex, and can contain undefined number of sub-parameters in a recursive form.

C. *Used algorithms*

We are using different type of algorithms to do comparison between intervention plan and realized interventions. :

- Boyer-Moore algorithm [6], which is basically a string searching algorithm.
- Needleman-Wunsch algorithm [7], which is an algorithm used mainly in bioinformatics to align protein or nucleotide sequences using dynamic programming. It provides global sequence alignment with penalties.
- Smith-Waterman algorithm [8], which is also a bioinformatics oriented matching algorithm. It provides local sequence alignment using penalties.

D. *Parameter difference evaluation*

After the analysis process we are evaluating the parameter value differences. We denote the P intervention process's i-th parameter as P_i .We have defined impact score matrix and assigned an impact score ($0<K_{Pi}<1$) to each member of the parameter set. We are using a simplified linear evaluation function to calculate the impact of the differences (I).

$$\Delta P_i = P_i - P'_i, \tag{1}$$

where P_i holds the planned and P'_i is the really occurred intervention values.

$$I = K_{Pi} * \Delta P_i, \tag{2}$$

simply calculates the weighted difference of the planned and occurred intervention tasks. Even if we can note here the subjective impact score value definitions. The larger I means larger distance between the plans and the real world scenarios. We can use the calculated values to search for alternative intervention graph paths or simply to optimize between the alternatives. Multiple impact score matrices can be used to analyze the intervention graphs from different viewpoints (medical service provider has easily different subjective impact values, than patients). In the current version of our solution only similar interventions process parameter types can be compared. This is caused by the fact, that we are not using any substitution matrix, to define mapping possibilities between the different intervention processes and their parameters. The evaluation results provide us a handy solution to compare and analyze intervention plans with the occurred intervention records. Similarities are helping to define real consensus intervention paths. The measurable differences are interesting investigation points where a premium medical service provider can gain vital information about sits service quality, and patients requirements.

VI. SUMMARY

The aim of our research was to define a medical intervention plan analyzer software solution, equipped with accurate intervention evaluation algorithms, combined with a user friendly graphical interface. With the designed and developed software solution we are able to compare arbitrary complex intervention graph structures with occurred interventions received as patient health records or intervention logs. Planned intervention task parameters are mapped to the occurred intervention parameters in an automatic way. We have defined the weight of each intervention parameters, and calculate an absolute distance of the two parameter values as the impact of the occurred differences

From the parameter evaluation a lot of "hidden" information can be extracted such as information about the planning accuracy of the medical professionals, the difference between the implemented intervention plans and the official so called consensus intervention protocols or the correlation factors between the (really) occurred interventions and the patient outcomes, or the user satisfaction level and the occurred interventions. The developed framework solution is a generic intervention plan analyzer, however it can be used or adapted easily to a large set of medical domains. We have validated successfully the system within the premium diabetes and dental care medical service domains in Hungary.

VII. ACKNOWLEDGMENTS

The projects have been supported by the European Union. The authors would like to thank GOP-1.1.1-11-2012-0055- Dialogic „DIALOGIC – Mathematical model based decision support system for diabetes monitoring", for its financial support. The authors would like to thank PPT Ltd. for their research support many advices and useful contributions.

REFERENCES

[1] Surján György, Borbás Ilona, Gődény Sándor, Juhász Judit, Mihalicza Péter, Pékli Márta, Kincses Gyula, Varga Eszter, Juhász

Judit, Nagy Attila, Szabó Dóra, Vargáné Lőrincz Ildikó; Egészségtudományi fogalomtár, http://fogalomtar.eski.hu/index.php/Kezd%C5%91lap

[2] Carry M. Renders,, Edward H. Wagner,, Gerlof D. Valk,, Jacques Thm. Eijk Van,, Simon J. Griffin,, Willem J.J. Assendelft; Interventions to Improve the Management of Diabetes in Primary Care, Outpatient, and Community Settings, A systematic review; Diabetes Care, Volume 24, Number 10, pp.:1821-1833, October 2001

[3] Dr. Vanessa Diaz, Marco Viceconti, Veli Stroetmann, Dipak Kalra et al. ;Radmal for the Digital Patient; DISCIPULUS project report, 2013 march

[4] A betegút szervezés módszertana és európai példái; HealthOnLine Hírlevél; 2012/7 – Különszám; 2012. augusztus 27.

[5] De Bleser L; Depreitere R; Waele K; Vanhaecht K; Vlayen J; Sermeus W (2006): Defining pathways. Journal of Nursing Management. 2006 october, 14. vol., 7.issue, pp.: 553-563. 11p.

[6] Boyer, Robert S.; Moore, J Strother (October 1977). "A Fast String Searching Algorithm.". Comm. ACM (New York, NY, USA: Association for Computing Machinery) 20 (10): 762–772. doi:10.1145/359842.359859. ISSN 0001-0782.

[7] Needleman, Saul B.; and Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of Molecular Biology 48 (3): 443–53. doi:10.1016/0022-2836(70)90057-4. PMID 5420325

[8] HealthCloud from Salesforce: http://www.salesforce.com/industries/healthcare/health-cloud/?d=70130000002DuoI , accessed on 20. March. 2016

[9] Mala Ramaiah, Eswaran Subrahmanian, Ram D Sriram, Bettijoyce B Lide; Workflow and Electronic Health Records in Small Medical Practices; Perspect Health Inf Manag. 2012 Spring; 9(Spring): 1d.Published online 2012 Apr 1.PMCID: PMC3329208

[10] EHR Portal http://www.practicefusion.com/phr/ , accessed on 20. March. 2016

[11] ChARM PHR Portal https://charmphr.com/login.sas , accessed on 20. March. 2016

[12] IntelliChart Portal http://www.intelichart.com/ , accessed on 20. March. 2016

[13] EMRAM http://www.himssanalytics.org/provider-solutions#block-himss-general-himss-prov-sol-emram , accessed on 20. March. 2016

[14] ICD-10, Classification of Diseases and Related Health Problems 10th Revision, Volume 2, Instruction manual, World Health Organization, 2010 Edition, ISBN 978 92 4 154834 2 (NLM classification: WB 15)

Using Fourier and Hartley Transform for Fast, Approximate Solution of Dense Linear Systems

Željko Jeričević* and Ivica Kožar**
* Department of Computer Engineering/Engineering Faculty, Rijeka, Croatia
** Department of Computer Modeling/Civil Engineering Faculty, Rijeka, Croatia
zeljko.jericevic@riteh.hr

Abstract - The solution of linear system of equations is one of the most common tasks in scientific computing. For a large dense systems that requires prohibitive number of operations of the order of magnitude n^3, where n is the number of equations and also unknowns. We developed a novel numerical approach for finding an approximate solution of this problem based on Fourier or Hartley transform although any unitary, orthogonal transform which concentrates power in a small number of coefficients can be used. This is the strategy borrowed from digital signal processing where pruning off redundant information from spectra or filtering of selected information in frequency domain is the usual practice. The procedure is to transform the linear system along the columns and rows to the frequency domain, generating a transformed system. The least significant portions in the transformed system are deleted as the whole columns and rows, yielding a smaller, pruned system. The pruned system is solved in transform domain, yielding the approximate solution. The quality of approximate solution is compared against full system solution. Theoretical evaluation of the method relates the quality of approximation to the perturbation of eigenvalues of the residual matrix. Numerical experiments illustrating feasibility of the method and quality of the approximation, together with operations count are presented..

I. INTRODUCTION

The solution of a system of linear equations is one of the basic methods in scientific computing. In the case of dense systems the order of magnitude N^3 multiplications is required (N is dimension of a square matrix)...

This paper extends the previous work [1],[2] in which we developed a framework for efficient linear least squares problems. We also supplement previous analysis with classical forward and backward error analysis [3]. Some equations developed previously [1] are briefly repeated here as final results without the details. The basic idea of constructing the approximate solutions for large, dense systems using the Fourier or Hartley space representation remains the same.

II. THEORY

The Fourier transform is formally done by multiplication with Fourier matrix but in actual computations FFT is used whenever possible.

$$b = Ac \tag{1}$$

Applying the Fourier transform on system of linear equations (1) is done along the columns. Premultiplying the (1) with Fourier matrix F and inserting $F^{-1}F=I$ after the matrix A, the transform along the column vector y and along the columns of matrix A, as well as the inverse Fourier along the rows of A and the Transform of solution vector c is performed (2). It is a good idea that before the transformation, the elements of vector b and corresponding row and column vectors in matrix A are sorted in such a way that the transform has close to the most compact representation. That would concentrate the energy in the Fourier transform in as little number of frequencies as possible.

$$Fb = \left(FAF^{-1}\right)\left(Fc\right) \tag{2}$$

To avoid computations with complex numbers in the case of real systems the Hartley transform [4] can be used. This transformation of original system with the terms grouped as shown using the parenthesis will yield the Fourier or Hartley transform of vector c as a solution for system (2).

III. METHOD

After the transformation the system (2) is of the same size as original system but it could be pruned down by deleting rows and columns containing "insignificant" information, yielding the smaller system. The information is termed insignificant in the signal processing sense: the frequencies whose magnitude is a smallest percentage of total magnitude are discarded. The selection of significant frequencies is accomplished by computing the magnitudes for frequencies in vector b and sorting them

in decreasing order. For columns, the sum of frequencies for each column is used for that purpose. Using that information, the transforms of vector *b* and matrix *A* can be shortened into significant parts to be retained and insignificant parts to be discarded. The solution of pruned system will yield the Fourier interpolation of solution vector c. This approach decreases the size of the system and represents the filtering out of non significant frequencies in order to build a smaller model system with a less equations and less unknowns then the original system [5]. In this respect our approach is different then Beylkin's [6] whose idea was to increase the sparsity of the matrix by using the wavelet transform. Using residual matrix *R* and Taylor order expansion [1] it was shown that approximate inverse for system converges faster to exact equations inverse the smaller eigenvalues (λ) for (*I-R*) matrix are (*R* is residual matrix).

$$y = Bx \qquad \left(y = Fb \quad B = FAF^{-1} \quad x = Fc\right)$$

$$B_p^{-1} \approx B^{-1}$$

$$R \equiv I - B_p^{-1}B \Rightarrow B_p^{-1}B = I - R$$

$$B = B_p\left(I - R\right)$$

$$B^{-1} = \left(I - R\right)^{-1} B_p^{-1}$$

$$\delta x = \left(I - R\right)^{-1} B_0^{-1}\delta y$$

$$\left(I - R\right)^{-1} = \left(QQ^{-1} - Q\Lambda Q^{-1}\right)^{-1}$$

$$= \left[Q\left(I - \Lambda\right)Q^{-1}\right]^{-1} = Q\left(I - \Lambda\right)^{-1}Q^{-1}$$

$$\lim_{\lambda \to 0}\left[Q\ diag\left(\frac{1}{1-\lambda_i}\right)_{i=1,\ldots,n} Q^{-1}\right] \to I \qquad (3)$$

$$\left(I - R\right)^{-1} = I + R + R^2 + R^3 + \ldots$$

$$B_n^{-1} \equiv \left(I + R + \ldots + R^n\right)B_p^{-1}$$

$$B_\infty^{-1} \to B^{-1}$$

Where B_p is pruned approximation of *B* based on the most significant frequencies. Solving the model system will yield the approximate solution of original system (1). The approximate solution ($x+\delta x$) is different from true solution *x* for difference vector δx. It is important to show how difference vector behaves depending on size of matrix B_p. The (3) shows that *R* should be a contraction mapping: its eigenvalues should all be non-negative, and smaller then one. Smaller the eigenvalues are, closer $(I-R)^{-1}$ is to identity matrix as can be seen from the limit (4) and consequently closer B_p^{-1} is to B^{-1}.

The system of linear equations based on Hilbert matrix (an ill conditioned, but non-singular matrix) is used here to illustrate the change in condition number for matrix of decreasing size. The numerical experiments with forward and backward error analysis are summarized in Table I. Table I contains the results for rearranged Hilbert matrix. The rearrangement was done to reduce the Gibbs rigging due to difference between the first and the last elements in each column and row. The forward (4) and backward (5) error analysis are defined here:

$$\Delta_f = \frac{\left\|c - c_p\right\|_2}{\left\|c\right\|_2} \qquad (4)$$

$$\Delta_b = \frac{\left\|b - b_p\right\|_2}{\left\|b\right\|_2} \qquad (5)$$

Where index *p* denotes solution of the pruned system.

TABLE I

N	*% freq*	*Con.#*	$\|c\|_2$	Δ_f	$\|b\|_2$	Δ_b
5		$4.7\text{x}10^5$	0.64		0.63	
5	100	$4.7\text{x}10^5$				
4	83	$1.2\text{x}10^5$	0.62	0.29	0.63	$6.9\text{x}10^{-6}$
3	68	874	0.70	0.22	0.63	$5.3\text{x}10^{-5}$
2	43	24	1.09	1.20	0.62	$4.3\text{x}10^{-3}$
1	25	1	0.43	1.03	0.60	$1.9\text{x}10^{-1}$

Where N is the dimension of the square system, second column defines how much of total frequency magnitude is used in constructing the solution, and *Con.#* is condition number of the matrix computed as a ratio of the largest and smallest singular value. Forward and backward errors are denoted Δ_f and Δ_b, respectively.

Solutions for Hilbert system (vector c) are shown in Figure 1. and restored right hand side (vector b) in Figure 2.

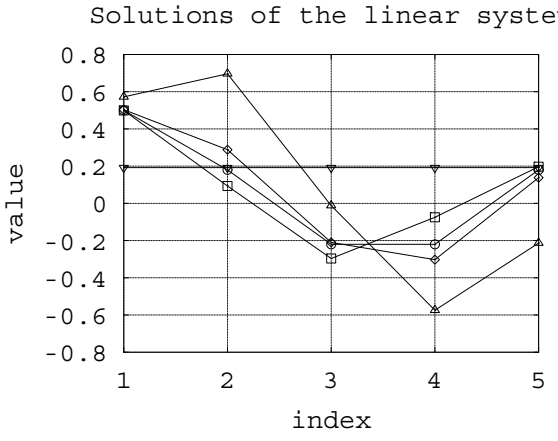


Figure 2. Exact solution (circle), N=4 (square), N=3 (diamond) N=2 (triangle up), N=1 (triangle down).

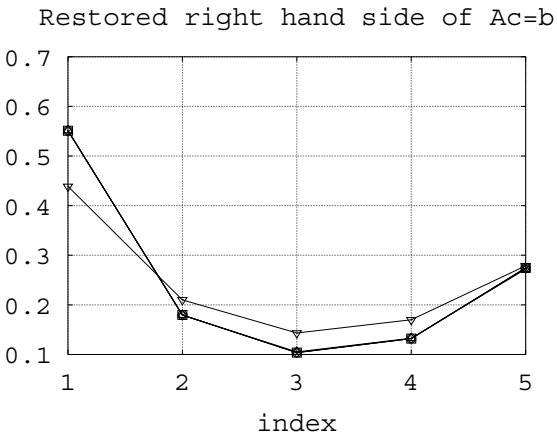


Figure 2. Exact vector *b* (circle), N=4 (square), N=3 (diamond) N=2 (triangle up), N=1 (triangle down).

Some interesting conclusions can be drown from the figures. The restoration of *b* is fairly good even for the worst of computed solutions. Consequently, the backward error analysis is too optimistic. The results of Forward error analysis are more realistic, but are not easy to compute, requiring the correct answer to evaluate an approximate one. Approach with eigenvalues (3) seems overly expensive, but to evaluate the convergence, only the highest eigenvalue is required. If it is less then 1, the iterative improvement will converge.

IV. CONLUSION

The approximate method for a fast solution of linear problems has been proposed. The quality of the approximation can be tested and controlled. We developed the equation which shows how the eigenvalues of residual matrix determine quality of solution and extent of corrections to be applied in order to approach the true solution. The proposed method offers more detailed assessment for the quality of solution then classical forward and backward error analysis.

REFERENCES

[1] Jeričević Ž., Kožar I, "Faster Solution of Large, Over-Determined, Dense Linear Systems", DC VIS MIPRO (2013) 228-231.

[2] Jeričević Ž., Kožar I, "Theoretical and statistical evaluation for approximate solution of large, over-determined, dense linear systems", DC VIS MIPRO (2015) 227-229.

[3] O'Leary, D.P., "Scientific Computing With Case Studies", 2009, SIAM, Philadelphia, pp 383

[4] Bracewell, R.N., "The Hartley Transform", Oxford University Press, New York, 1986 p 160.

[5] Jeričević, Ž., "Approximate Solution of Linear Systems" Croatica Chemica Acta, Vol. 78 (2005) 601-615

[6] Beylkin, G., Coifman, R., and Rokhlin, V, "Fast Wavelet Transforms and Numerical Algorithms", Com. Pure App Math. Vol. 44 (1991) 141–183

Procedural Generation of Mediterranean Environments

N. Mikuličić* and Ž. Mihajlović*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
niko.mikulicic@gmail.com, zeljka.mihajlovic@fer.hr

Abstract - This paper describes an overall process of procedural generation of natural environments through terrain generation, texturing and scattering of terrain cover. Although described process can be used to create various types of environments, focus of this paper has been put on *Mediterranean* which is somewhat specific and has not yet received any attention in scientific papers. We present a novel technique for procedural texturing and scattering of terrain cover based on cascading input parameters. Input parameters can be used to scatter vegetation simply by slope and height of the terrain, but they can also be easily extended and combined to use more advanced parameters such as wind maps, moisture maps, per plant distribution maps etc. Additionally, we present a method for using a satellite image as an input parameter. Comparing results with real-life images shows that our approach can create plausible, visually appealing landscapes.

Keywords: procedural modeling, landscape generation, virtual environments, natural environments

I. INTRODUCTION

Reproducing realistic environments has always been a challenge in computer graphics mainly due to a large number of data needed to be simultaneously processed and displayed on screen. With many optimization methods and ever advancing graphics hardware, realistic landscapes have become achievable in real time and are widely used in computer games, military simulators and film industry. Recent trends show that virtual landscapes have also found their usage in touristic promotions and since environmental research has become an important topic in the last years, it can be expected for virtual environments to have a major role in simulating how changes in different environmental parameters influence the surrounding environment.

Manually creating a whole environment would be a long lasting, tedious task, therefore many procedural techniques have been developed to aid the process. Although procedural, these methods usually have many parameters and finding the right values is often a time-consuming trial and error process. In recent years, various methods for inverse procedural modeling have been introduced [7][22]. Rather than setting parameter values manually, these methods provide ways to learn them. Nevertheless, procedural modeling remains the most common approach to generation of environments.

Authors discussing generation of natural virtual environments usually reproduce mountain meadows and dense forests. Although visually appealing, tall and dense trees efficiently hide everything behind them so many objects can be safely culled and popping effects can be easily hidden. In this paper we focus on natural Mediterranean environments which have not yet been discussed. Mediterranean environments often lack tall trees and instead have very dense shrubbery. Since shrubs often cannot hide the landscape behind, various optimization techniques have to be implemented with extra care for popping effect to be properly hidden and for the environment to be efficiently rendered in real time.

Our generation process is being done in four steps: preparation, terrain mesh generation, texture mapping and scattering of terrain cover. In preparatory step, a user defines a set of textures and models to be used and assigns them input parameter values. The second step is used to create a terrain triangle mesh from heightmap which is followed by procedural generation of splat map and texturing of the terrain in the third step. The last step places the terrain cover using scattering algorithms to achieve natural randomness. Although divided into four steps, the user often returns to preparatory step to adjust parameter values and then repeats the process until he is satisfied with results. The whole process needs to be repeated only when user changes the terrain geometry. Otherwise, user can focus on just one step.

Although the focus of this paper is on Mediterranean, the described process can also be used to reproduce any other natural environment simply by using different textures and models.

After presenting related work in chapter two, chapter three continues by explaining techniques used to represent terrain, textures and terrain cover as well as the optimizations required for real time performance. Fourth chapter focuses on algorithms used for procedural texturing and scattering of terrain cover. We present our results in chapter five and give conclusion and some guidelines for future work in chapter six .

II. RELATED WORK

This work belongs to procedural modeling of natural environments, and as such, it connects various fields of computer graphics.

Procedural modeling automates the process of content creation by using set of rules and algorithms rather than creating content manually. It is often applied where manual creation of content would be too cumbersome a task. Procedural modeling has been successfully used to create various types of content some of which include textures [6], plants [10], buildings [12], cities [18], and terrains [8].

Terrain generation is often being done in two steps: generation of heightmap and creation of 3D mesh. Heightmaps are usually generated using fractal algorithms [1][11], noises [1][6] and erosions [1][13][17]. After creating a heightmap, terrain mesh is generated and optimized using level of detail algorithms. Static LOD algorithms reduce computational at the cost of memory demands and often need nontrivial and rigid preprocessing stage [23]. Continuous LODs are generated dynamically and are therefore more flexible and scalable but also computationally more demanding [15][21].

Forest representation includes representations of individual trees as well as the process of achieving their natural distribution in ecosystem. Trees have been represented with level of detail [3], billboards [20], parallel billboards [9], volumetric textures [16] and billboard clouds [5]. To create an ecosystem, trees are scattered using *global-to-local* approaches [4][14] that define some global rule by which the scattering occurs. Another approach is *local-to-global* [2][4][14] that models interaction between individual plants from which the global distribution arises.

III. RENDERING

Natural environment consists of several somewhat independent layers: terrain, textures and terrain cover. In order to achieve efficient rendering, optimization should be done in each layer.

Terrain is usually represented with a heightmap which can be generated procedurally or downloaded from Internet in case real-world data is needed. Fig. 1 shows heightmap of Croatian island Ist (left) and heightmap generated using Perlin noise and Fractional Brownian motion (right). To generate terrain triangle mesh from heightmap we use Lindstrom-Koller simplification [15] on a single terrain region. Creating multiple regions with levels of detail is left for future work.

Textures are applied to terrain by simple planar texture mapping. Splat mapping technique is used to provide more surface details close to observer while simple color mapping is applied to more distant areas where details are not visible anyway.

Terrain cover is split into two subcategories: details and high cover.

Details are small objects, like grass, visible only from short distances. Due to their size, many objects are needed to cover any portion of a terrain so they tend to use up most of the available computational resources. In order to achieve real-time rendering of large fields covered with grass, various optimization methods have to be implemented. Modeling each blade of grass independently would quickly reach a maximum of vertices and polygons

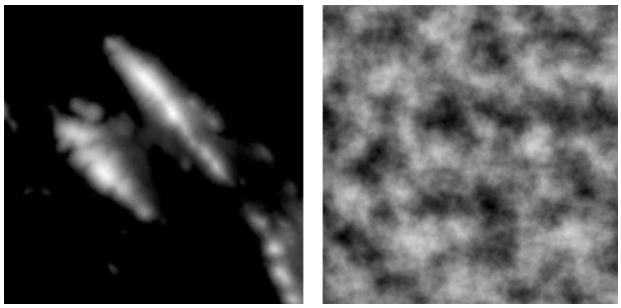


Figure 1. Heightmap of Mediterranean island Ist, Croatia (left) and procedurally generated heightmap (right)

a graphics card can process in real time and for that reason, more efficient approaches have been developed. We represent multiple blades of grass with one axial billboard. Many billboards are needed to cover an area with grass and sending each billboard to GPU independently would mean too many draw calls and therefore, slower rendering. Instead, multiple billboards are grouped together and sent to GPU as point clouds in which every vertex represents the position of a single billboard. Billboard geometries are then created at runtime in geometry shader at positions defined by vertices in the given point cloud. Additionally, since details are visible only from short distance, all billboards with distance to observer greater than some user defined value are culled. Terrain split to regions would fasten the culling process, allowing large groups of billboards to be discarded with a single distance check. To avoid sudden popping effects, a transitional area is used in which a billboard goes from completely invisible to fully visible using alpha cutout technique.

High cover includes all objects that can be viewed from greater distances like trees or larger rocks. To optimize the number of polygons in a scene, levels of detail have been used. Objects close to the observer are rendered using high quality models, while lower quality models are used on objects that are further away. To avoid popping effects while transitioning between different models of the same object, a cross-fade technique is used. Both models are rendered for some small amount of time, while one is slowly fading in and the other is fading out. For the lowest quality model of a tree we use a simple axial billboard.

IV. PROCEDURAL GENERATION

In this chapter we focus on procedurally transforming input parameters into texture weights and terrain object placement probabilities. Texture weights are used to create splat maps and color maps for terrain texturing while terrain object placement probabilities are used to generate terrain cover using scattering algorithms.

A. Weights Calculation

Input parameters are defined in preparatory step. For each input parameter, every texture and terrain object defines minimum and maximum values and a weight function to describe parametric area it resides in. For example, a texture can be given a slope input parameter. To define a parametric area it resides in, a texture has to specify minimum and maximum slope. Weight function is used to describe texture's weight or preference in area

between minimum and maximum. Fig. 2 represents a common weight function although any curve can be used. Weight functions are defined in $[0, 1]$ interval and scaled to fit specified range.

Once we have defined parametric areas for all textures and terrain objects, their weights can be easily calculated. For texture or terrain object t with parameters p_i and weight functions w_{p_i}, the weight w_t is defined as:

$$w_t(\boldsymbol{r}) = \prod_{i=1}^{n} w_{p_i}(f_{p_i}(\boldsymbol{r})), \quad (1)$$

where n is the number of parameters and f_{p_i} is parameter dependent function which maps terrain position $\boldsymbol{r}$ to parameter value. It can represent calculating surface slope, elevation, sampling value from texture etc. Put simply, final weight is calculated by multiplying weights of individual parameter values at given position.

B. Texturing

Texture weights are calculated to give us information how much each texture belongs to specified area. This information can be used to procedurally texture a terrain. For terrain texturing we used splat mapping and color mapping techniques.

Color mapping technique textures a whole terrain with just one texture. One pixel of color map usually covers not as small portion of terrain so this technique usually results with dull and blurry terrains when observed from close. It would require extremely large textures to create detailed surface using this technique which would be quite inefficient.

To add more surface detail, splat mapping technique is used. Splat mapping uses multiple high detailed surface textures that are mapped to a small part of terrain and then tiled to fit the rest. One additional texture called splat map is mapped to whole terrain and used to provide information on how much each surface texture contributes to the final color of surface. Final color is determined dynamically in fragment shader by sampling all surface textures and multiplying sampled colors with their contributions sampled from splat map.

Splat map can be directly generated from weights defined in previous chapter. One pixel in splat map corresponds to terrain area at position $\boldsymbol{r}$. Texture weights are then calculated using (1), scaled to sum of 1 and stored in a single splat map pixel, one channel per surface texture. This means that one splat map can store weights for four surface textures. We can always use a second splat map for four new surface textures, but that can become inefficient. Usually, four surface textures are more than enough.

Color map is generated similarly. Instead of storing texture weights and calculating surface color dynamically in fragment shader, colors are sampled, mixed and baked into a color map. To avoid high frequencies in resulting texture, colors are not sampled from original surface textures but rather from their last mipmap.

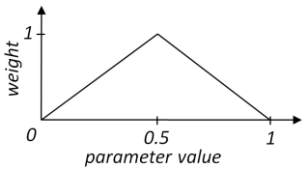


Figure 2. Weight function

C. Scattering of Terrain Cover

After procedurally texturing a terrain, next step is to provide it with grass and high cover.

To achieve natural randomness and uniform distribution of grass, we use a regular grid with random displacement algorithm. Terrain is split into a regular grid with one grass billboard assigned to the center of each cell. Each billboard is then displaced in random direction by amount small enough to remain inside the given cell. For each billboard position $\boldsymbol{r}$, grass weights are calculated using (1). Grass type is then chosen using roulette wheel method where weight of grass type is proportional to associated probability.

Potential collisions of grass on cell borders can be ignored as they visually do not make much difference. Collisions between trees, however, should be avoided. Trees can have different sizes and placing them in a regular grid would not be as easy as it was with grass. Therefore, for scattering trees we use simple random scattering with collision detection. We randomly choose a position, calculate tree weights and choose a tree sort using roulette wheel method. If chosen sort cannot fit onto position without colliding with surroundings, we remove the sort from consideration and try with another. The algorithm ends when all trees have been placed or the maximum number of iterations has been reached.

In reality, forest is not just trees being randomly scattered. They are grown from seeds, fighting for space and resources. Stronger trees prevail giving birth to new ones and the process repeats. New seeds fall in the vicinity of mother tree which results in trees than are not just randomly scattered but rather grouped. To simulate this process we have implemented a survival algorithm with environmental feedback similar to algorithm described in [2]. At the beginning of each iteration, all plants generate seed at random positions inside seeding radius. Seed that falls outside of terrain or on a ground that is not inside specie's parametric area is removed. Next step is detecting collisions between plants. Plants in collision are fighting for survival and weaker plant, a plant with smaller viability [2], is eliminated. At the end of iteration, all plants increase in age and plants that have reached their maximum age are removed from the ecosystem. The algorithm ends after defined number of iterations.

Fig. 4 shows comparison of these three scattering algorithms where different tree species are being represented by circles of different colors and radii. Grid displacement algorithm (left) gives good uniform distribution of objects of the same size. Random scattering with collision detection (middle) can successfully scatter objects of different sizes, however, it does not produce clusters of same species like survival algorithm (right) which gives forest a more natural look.

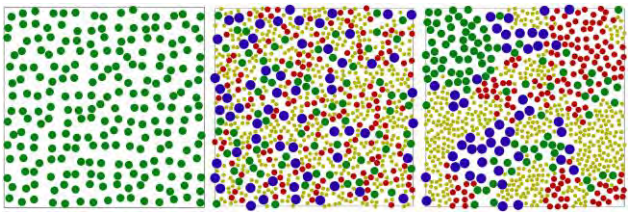


Figure 4. Scattering algorithms. Scattering on regular grid with random displacement (left), random scattering with collision detection (middle) and survival algorithm (right)

D. *Classifying Satellite Image*

For procedural texturing and scattering of terrain cover, we can also use a satellite image as an input parameter. To translate image into weights we use a simple, color based classification. To calculate weight at certain terrain position, satellite image is sampled and given color mapped to HSV space where boundaries between colors are more obvious.

Fig. 3 shows a satellite image with its classification. Due to atmospheric scattering, images taken from great distances have their colors shifted towards blue. This makes classifying harder and more advanced terrain cover classifier would be needed to fully indentify the coastal line or tell difference between sea and vegetation. More advanced methods for classifying surface cover have been developed [19], however this approach is simple and gives good enough results in natural, uninhabited areas.

After color sampled from satellite image is classified to surface cover type, we can treat that information as weight equal to 1 if terrain object is of specified type, or 0 if it isn't.

V. RESULTS

To evaluate our approach we will make comparisons in different scales between real and procedurally generated Mediterranean landscape.

Fig. 5 shows Croatian island Ist textured using different input parameters. Fig. 5 (left) represents terrain textured using slope and height. This method is generic and usually gives good results on both real and procedural terrains. Fig. 5 (middle) represents the same landscape textured using terrain cover classification from satellite image. Terrain is textured more precisely using this method, however it can only be used on real-world

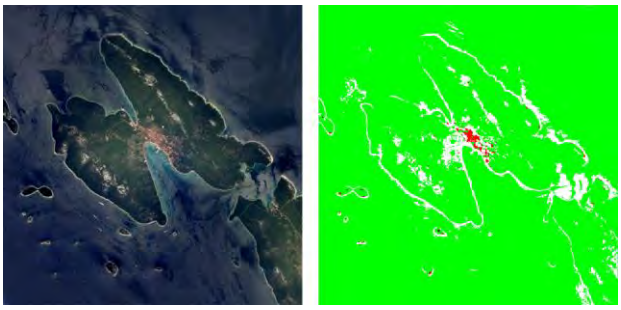


Figure 3. Satellite image of Ist (left) and classification of terrain cover (right)

landscapes. Additionally, procedurally generated terrain is never a perfect representation of real one due to inaccuracies in heightmap, interpolations and simplifications of terrain geometry. For that reason, coastal line is often misaligned as it can be seen on Fig. 5 (middle). Fig. 5 (right) tries to solve the coastal line problem using combination of previous two approaches. The coastal line height is defined and everything above that level is textured using terrain cover classification while everything below using only slope and height as input parameters. This makes coastal line more monotonous but consistent.

Fig. 6 shows satellite image of smaller, uninhabited part of Ist and its procedural representation generated using already mentioned coastal line reconstruction approach. The technique proves to be more reliable in uninhabited areas with image taken from closer distance. In addition to textures, Fig. 6 (right) also contains trees scattered according to terrain cover classification.

Following image (Fig. 7) shows the same terrain from the ground. To represent Mediterranean grass we used textures that are similar to following Mediterranean plants: *Lactuca serriola, Trisetum flavescens, Conyza sumatrensis, Eupatorium cannabinum, Urtica dioica* and *Melilotus sp*. Also, a model of *Arbutus unedo* has been used as a bush. To have more control over distribution of each species we included additional spread parameter for every grass and tree type.

Finally, Fig. 8 displays a comparison between real and procedurally generated landscape. Every object on the image is procedurally scattered except for the two bushes in front which are placed manually for easier comparison.



Figure 5. Comparison of procedural texturing by slope and height (left), by classifying satellite image (middle) and using coastal line reconstruction approach (right)

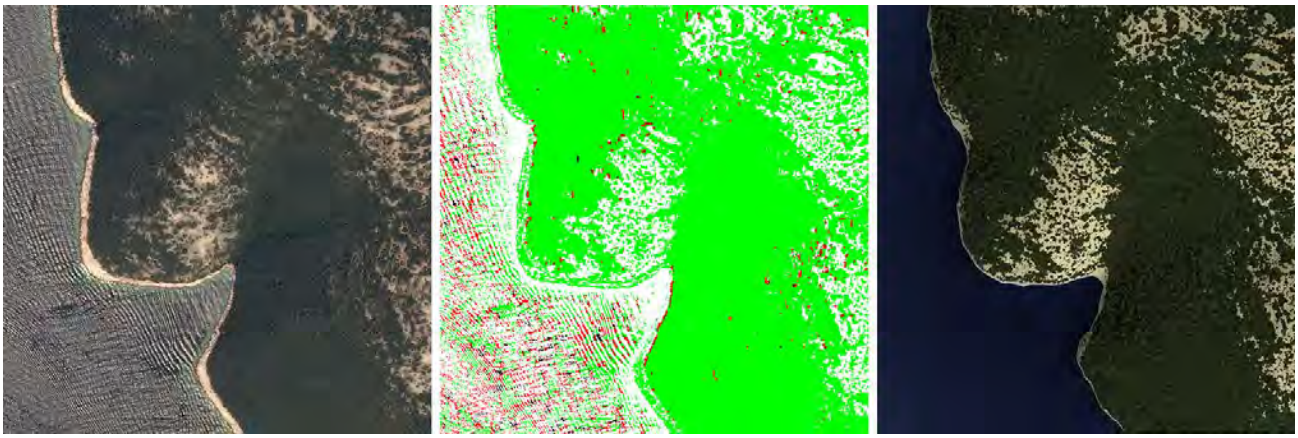


Figure 6. Satellite image of Dumboka cove at Ist (left), terrain cover classification (middle) and procedurally generated landscape (right)



Figure 7. Procedurally generated Dumboka cove from ground



Figure 8. Comparison between real (left) and procedurally generated landscape (right)

VI. Conclusion and Future Work

We presented a method for procedural generation of environments through terrain generation, texturing and scattering of terrain cover using cascaded input parameters. By calculating and multiplying weights of individual input parameters we obtain information about how much each texture or species wants to reside on given terrain position. That information can be used as texture's contribution to terrain color when texturing the terrain, or as probability when scattering the terrain cover. Height and slope of the terrain have proven to be reliable input parameters for generic uses. In case of real-world terrains, simple terrain cover classification from satellite image can be used to provide basic information for texturing and scattering of terrain cover. To fix coastal line issues caused by geometry misalignments between real and generated terrain, we introduced a coastal line reconstruction approach. Additionally, we used a spread parameter to have more control over distribution of each species. Comparing real life images with those generated procedurally, we believe that our method can create plausible and visually appealing Mediterranean landscapes.

Many suggestions for future work can be made. Terrain should be split into regions for faster view frustum culling of terrain geometry, grass and trees. Level of detail algorithms should be used to dynamically reduce the polygon count of faraway terrain regions. Those regions could also use cheaper texturing technique, for example color mapping instead of splat mapping. Closer, high detail regions could use normal mapping or even fractal displacement of geometry to achieve 3D look of Mediterranean karst. Scattering terrain cover based on satellite image would benefit from more advanced terrain cover classifier. Although height and slope as input parameters give good generic results, they are not primary factors in shaping Mediterranean environment and it would be interesting to see how would additional parameters, like wind map and resistance to wind and salt, affect the final look of the environment.

References

[1] T. Archer, "Procedurally generating terrain", 44th annual midwest instruction and computing symposium, Duluth, 2011., pp. 378–393.

[2] B. Beneš, "A stable modeling of large plant ecosystems", In Proceedings of the International Conference on Computer Vision and Graphics, Association for Image Processing, 2002, pp. 94–101.

[3] C. Colditz, L. Coconu, O. Deussen, C. Hege, "Real-time rendering of complex photorealistic landscapes using hybrid level-of-detail approaches", 6th Interational Conference for Information Technologies in Landscape Architecture, 2005.

[4] O. Deussen, P. Hanrahan, B. Lintermann, R. Měch, M. Pharr, and P. Prusinkiewicz, "Realistic modeling and rendering of plant ecosystems", In Computer Graphics (Proceedings of ACM SIGGRAPH) (1998), pp. 275–286.

[5] X. Décoret, F. Durand, F. X. Sillion, J. Dorsey, "Billboard clouds for extreme model simplification", In ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH) (2003), pp. 689–696.

[6] D. S. Ebert, S. Worley, F. K. Musgrave, D. Peachey, and K. Perlin, Texturing & Modeling, a Procedural Approach, Elsevier, 3rd edition, 2003.

[7] A. Emilien, U. Vimont, M.-P. Cani, P. Poulin, and B. Beneš, "WorldBrush: Interactive example-based synthesis of procedural virtual worlds", ACM Transactions on Graphics (SIGGRAPH 2015), vol. 34, issue 4, pp. 11.

[8] J.-D. Génevaux, É. Galin, É. Guérin, A. Peytavie, and B. Beneš, "Terrain generation using procedural models based on hydrology", ACM Transactions on Graphics (SIGGRAPH 2013), vol. 32, issue 4, 143:1–13.

[9] A. Jakulin, "Interactive vegetation rendering with slicing and blending", In Eurographics 2000 (Short Presentations), 2000.

[10] R. Měch, and P. Prusinkiewicz, "Visual models of plants interacting with their environment", In Computer Graphics (Proceedings of ACM SIGGRAPH) (1996), pp. 397–410.

[11] G. S. P. Miller, "The definition and rendering of terrain maps", In Computer Graphics (Proceedings of ACM SIGGRAPH) (1986), vol. 20, pp. 39–48.

[12] P. Müller, P. Wonka, S. Haegler, A. Ulmer, and L. Van Gool, "Procedural modeling of buildings", In ACM Transactions on Graphics (SIGGRAPH, 2006) , vol. 25, issue 3, pp. 614–623.

[13] F. K. Musgrave, C. E. Kolb, and R. S. Mace, "The synthesis and rendering of eroded fractal terrains", In Computer Graphics (Proceedings of ACM SIGGRAPH) (1989), pp. 41–50.

[14] B. Lane and P. Prusinkiewicz, "Generating spatial distribution for multilevel models of plant communities", In Proceedings of Graphics Interface ’02., vol. 1, pp. 69–80.

[15] P. Lindstrom, D. Koller, W. Ribarsky, L. F. Hodges, N. Faust, and G. A. Turner, "Real-time, continuous level of detail rendering of height fields", In Computer Graphics (Proceedings of ACM SIGGRAPH) (1996), pp. 109–118.

[16] F. Neyret, "Modeling, animating and rendering complex scenes using volumetric textures", IEEE Transactions on Visualization and Computer Graphics, vol. 4, issue 1 (1998), pp. 55–70.

[17] J. Olsen, "Realtime procedural terrain generation", Technical Report, University of Southern Denmark, 2004.

[18] Y. I. H. Parish, and P. Müller, "Procedural modeling of cities", In Computer Graphics (Proceedings of ACM SIGGRAPH) (2001), pp. 301–308.

[19] S. Premoze, W. B. Thompson, and P. Shirley, "Geospecific rendering of alpine terrain", In Rendering Techniques (1999), pp. 107–118.

[20] J. Rohlf, J. Helman, "IRIS performer: A high performance multiprocessing toolkit for real-time 3D graphics", In Computer Graphics (Proceedings of ACM SIGGRAPH) (1994), pp. 381–394.

[21] F. Strugar, "Continuous distance-dependent level of fetail for rendering heightmaps (CDLOD)", In Journal of Graphics, GPU, and Game Tools, 2009, vol. 14, issue 4, pp. 57–74.

[22] O. Šťava, S. Pirk, J. Kratt, B. Chen, R. Měch, O. Deussen, and B. Beneš, "Inverse procedural modelling of trees", Computer Graphics Forum, 2014, vol. 33, issue 6, pp. 118–131.

[23] T. Ulrich, "Rendering massive terrains using chunked level of detail control", SIGGRAPH Super-Size It! Scaling Up to Massive Virtual Worlds Course Notes, 2002.

Energy-Aware Power Management of Virtualized Multi-core Servers through DVFS and CPU Consolidation

Hamed Rostamzadeh Hajilari *, Mohammad Mehdi Talebi * and Mohsen Sharifi *
*Iran University of Science and Technology/School of Computer Engineering,Tehran, Iran
hamedrostamzade@comp.iust.ac.ir, mehdi_talebi@comp.iust.ac.ir, msharifi@iust.ac.ir

Abstract— **Considerable energy consumption of datacenters results in high service costs beside environmental pollutions. Therefore, energy saving of operating data centers received a lot of attention in recent years. In spite of the fact that modern multi-core architectures have presented both power management techniques, such as dynamic voltage and frequency scaling (DVFS), as well as per-core power gating (PCPG) and CPU consolidation techniques for energy saving, the joint deployment of these two features has been less exercised. Obviously, by using chip multiprocessors (CMPs), power management with consideration of multi-core chip and core count management techniques can offer more efficient energy consumption in environments operating large datacenters. In this paper, we focus on dynamic power management in virtualized multi-core server systems which are used in cloud-based systems. We propose an algorithm which is effectively equipped by power management techniques to select an efficient number of cores and frequency level in multi-core systems within an acceptable level of performance. The paper also reports an extensive set of experimental results found on a realistic multi-core server system setup by RUBiS benchmark and demonstrates energy saving up to 67% compared to baseline. Additionally it outperforms two existing consolidation algorithms in virtualized servers by 15% and 21%.**

***Index Terms*— CPU Consolidation, Energy Saving, DVFS, PCPG, Multi-core Servers.**

I. INTRODUCTION

Due to large amount of energy consumed in today's servers, there is a growing need for energy-aware resource management. Large volume of energy consuming has implications on electricity costs causing many datacenters reporting millions of dollars for annual usage. 30$ billion is reported for enterprise power and cooling in the world annually. Additionally, a more recent report from the Natural Resources Defense Council (NRDC) claims waste and inefficiency in U.S. data centers that consumed a massive 91 billion kWh of electricity in 2013 will increase to 140 billion kWh by 2020, the equivalent of 50 large (500 megawatt) power plants. The amount of energy consumed by the world's data centers will triple in the next decade, putting an enormous strain on energy supplies and dealing a hefty blow to efforts to contain global warming, experts say. It is probably safe to say that most data center managers are dealing with the challenge of increasing data growth and limited IT resources and budgets. With "save everything forever" strategies becoming more prevalent for many organizations.

Traditionally active servers were focused for energy efficiency. The former approaches aim at reducing the number of active servers in a datacenter by consolidating all the incoming workloads into fewer server machines whereas the latter attempts to keep the performance of each active server in accordance with the assigned workload so that energy can be saved at the level of each server.

A data center is designed to provide the required performance based on its service level agreements (SLAs) with clients even during workload hours, and hence, its resources are vastly under-utilized at other times. As an example, the minimum and the maximum utilization of the statically provisioned capacity of Facebook's data center are 40% and 90%, respectively [1]. Therefore, a great amount of energy costs can be reduced by consolidating workloads into as few server machines as possible and turning off the unused machines. The server consolidation has been considered in many studies; and virtual machine migration (VMM) has been used as a means of server consolidation in many researches [2], [3], [4], [5] and [6].

Due to overheads and limitations associated with the server consolidation like high network latency, large system boot time, other approaches still can be concerned for energy saving. Because of these limitations, consolidation decisions are made in long periods of times. The longer the period is, the more server machines are still under-utilized which implies that there is still hoping to use other techniques to improve energy efficiency.

Focusing on each active sever is another state-of-art technique. While focusing on each active server; two widely accepted and employed techniques are used which are sleep states and Dynamic Voltage and Frequency Scaling (DVFS). Obviously these new approaches can be used as a complementary with the previous techniques.

DVFS was introduced decades ago and is one of the well-known and common energy-aware CPU management techniques which has been studied in many researches [2] [7] [8].

DVFS has its own limitations. First the supply voltages have already become quite low and therefore a small amount of further supply voltage reductions is possible. In addition, typically in datacenter servers there are two or more processor chips containing multiple CPUs with a single on-chip power distribution network, shared by all the CPUs, which makes it impossible for CPUs to operate in different supply voltage level and hence various clock frequencies. This limitation will result in under-utilized CPUs where the available performance level is higher than what is actually needed, hence energy is wasted.

Another well-known CPU energy management technique is Dynamic Power State Switching. In Many modern processors, it is possible to turn on or off the cores or CPUs. The workload for each CPU can be measured by OS and it is possible to Turn CPUs on or off. In other words each CPU is placed in its own power domain and therefore the power to each such domain can be independently gated. You should notice that the suggestion from the OS may not be a good one that's why it may be cancelled by Power Control

Unit (PCU) which resides in the processor chip. PCPG is the deepest sleep state in modern multi-core processors which is capable to shut down a core completely. So a gated core has nearly zero energy consumption received a lot of attention.

CPU consolidation is a system level technique to consolidate load to fewer active cores and the remained cores can be turned off. Traditionally, clock gating was used for the unloaded cores which also leads to energy saving. But, since modern processors support PCPG and CPU consolidation, they lead to significant energy saving. Obviously this approach beside DVFS can report better results.

In this paper, our algorithm is equipped with both of the mentioned approaches. We suppose virtualized multi core servers that make our approach appropriate for cloud computing environments. In our approach we present a method to find the best choice of the number of active cores and appropriate frequency according to the related workload and then power gate remained cores. All of the results are obtained from the actual hardware measurements and not simulations. Experiments with implementation of our algorithm on a realistic multi-core server system setup have shown 67% percent improvement in energy consumption.

The rest of the paper is organized as follows. Section II introduces related work on using PCPG and DVFS for energy efficient power management in datacenters. Section III describes problem definition for multicore processors of servers. In section IV our energy consumption model is presented and in section V the proposed energy aware mechanism based on energy model is described. We evaluate proposed mechanism compared to existing works in section VI. Finally in section VII conclusion and future works are presented.

II. RELATED WORKS

Despite various studies in power management of processors which exploit DVFS [2], [7], [9] and [10] rather than sleep states, recently presenting PCPG as deepest sleep state spread out the use of sleep states in datacenters [11] [12], [13], [14] and [15]. The Jacob Leverich and et al. research was the first one in this area which evaluates the use of PCPG in datacenters and demonstrated the importance of PCPG in power managements systems [12]. Their results show that PCPG solely can save energy up to 40% which is 30% more than using DVFS as a knob. They also show that the joint deployment of these two knobs together could save energy up to 60% in datacenters.

As respect to considerable energy saving by PCPG, gating the power of cores has energy and performance overheads [16], [14] and [17]. Therefore, frequent mispredictions may result in inefficiency power management and power dissipation. Niti Madan et al. demonstrate that it is vital to attend PCPG by guarded mechanisms to hinder such negative consequences on intra-core level [17]. Reference [16] demonstrate that in addition to intra-core algorithms, inter-core algorithms also suffer from such negative outcomes and may lead to power dissipation, unlike expectations. Therefore, it is crucial to exploit guard mechanisms to prevent such negative outcomes. In [16] the authors use a guarded mechanism to decide when to gate the cores in both intra-core and inter-core gating algorithms. We also considered power and performance overheads of PCPG in our mechanism. Power overhead is considered in power model and due to the interval between the high threshold and maximum capacity of active cores; there is always room to mitigate performance overhead.

In the datacenters, due to power capacity overload and overheating caused by high server densities, system failures may occur. There is a need for robust and definite power managements to avoid system failures. As gated core doesn't consume energy, it has high potential to widely be used in power capping problems. Kai Ma et al. propose an integration design of using PCPG and DVFS/overclocking in such problems to not only satisfy power constraints but also to optimize performance [13]. They were trying to optimize the performance of a Chip Multiprocessor (CMP) within a given power constraint (i.e., power capping) but we are minimizing the energy consumption of CMPs in virtualized servers within an acceptable performance.

In the datacenters experiencing low utilization, CPU consolidation technique equipped by PCPG present significant energy saving [18], [11] and [19]. With commercial hardware support of Core-level Power Gating, consolidation becomes a promising power management technique in datacenters [20]. In [1] the authors presented a technique called Core Count Management (some variant of the CPU consolidation technique) and reported 35% energy saving. However, they reported both power and performance results based on simulations performed by using simple power and performance models. Their work only uses PCPG but we propose a joint deployment of PCPG and DVFS.

The most related works to ours are [11] and [18] which use joint techniques of DVFS and CPU consolidation together in virtualized environments. In spite of aforementioned researches, they implement a consolidation algorithm on a realistic multi-core server system setup. The authors first investigated effect of CPU consolidation on power dissipation and performance (latency) of such systems, and concluded by presenting two new CPU consolidation algorithms for virtualized multi-core servers. Their algorithms blindly consolidate CPUs based on the predicted utilization requirement, but, in our work we consolidate and choose frequency level of active cores based on their energy consumptions.

III. PROBLEM DEFINITION

In most of the todays power management researches, for the sake of simplicity in design due to different performance overhead of DVFS and PCPG, a decoupled approach is used in which the power management knobs operate independently. Considering these power management knobs independently have substantial drawback which leads inefficient simultaneous functions.

In [19] this issue has been discussed that DVFS hinders PCPG for the workloads of multi thread applications like Fluidanimate from PARSEC benchmark. DVFS and PCPG are applied in different periods of time independently and due to lower overhead, DVFS is used in smaller periods and PCPC is used in longer periods in contrary. While workload is low, DVFS which is running in shorter periods of time decides to reduce the frequency to keep cores in higher utilization. But high utilization makes PCPG to keep the cores working and so none of them is turned off. Authors in [19] indicate that using DVFS beside PCPG not only doesn't improve energy saving but also results higher energy consumption than PCPG solely. Therefore there is a need for an approach using both of the techniques without hindering each other.

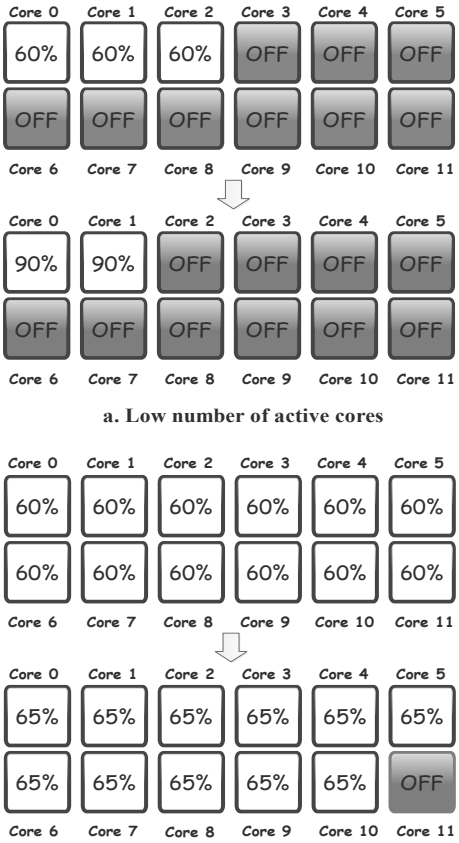


Fig1. Effect of CPU consolidation on workload.

CPU consolidation tries to resolve aforementioned issues based on utilization and frequency of cores. Now this question is raised that "Does consolidation always maximize energy saving?". Our experimental result show that the answer is no. In low number of cores while turning off one core, the workload of the others is increased dramatically and makes them to work in higher frequency which results more energy consumption (Fig1.a). Conversely in High number of cores, reducing one of the active cores doesn't affect the workload of the others much, so there is no tangible change in the frequencies (Fig1.b). In this case turning off active cores is an effective approach. Our approach in this paper brings dynamicity to consolidation approaches to choose best number of active cores and frequency for minimizing energy consumption.

IV. ENERGY CONSUMPTION MODEL

In this section we present our approach for energy consumption model which the approach is based on. We first model the relation between core frequencies, number of active cores and energy consumption. We use presented model in our approach to set appropriate number of active cores and frequency in order to save energy.

Obviously the frequency, performance state and energy consummation are tightly coupled elements of our model. In this part these relationships are calculated. In Fig.2 there are different performance states (P-State). Each P-State corresponds to a frequency supported by the processor.

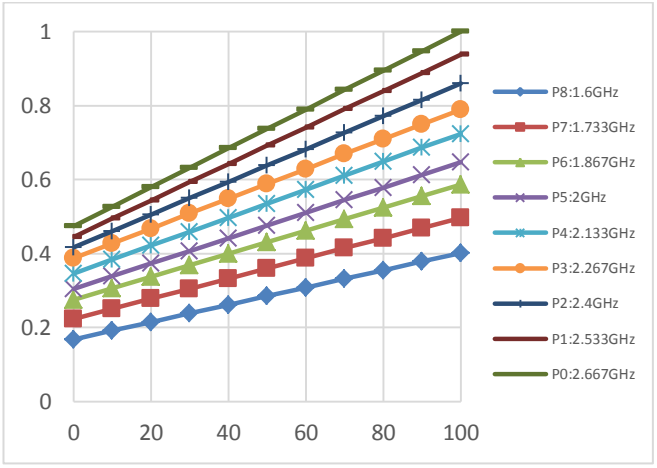


Fig2. Linear function for energy consumption respect to utilization

Our processor supports 9 P-states which P0 is the maximum performance state corresponding maximum frequency and P9 is lowest possible frequency that processor supports. You should notice that lower P-state results higher frequency and computation and in consequence higher energy consumption. The relations described in this section are derived from data collected in benchmarking experiments that CPU intensive workloads with different levels of CPU demand were applied to the servers under test.

The power consumption of the servers was collected for all the operation frequencies. The data from the experiments suggest that for a fixed operating frequency, the power consumption of the server is approximately linear function of the utilization of the processor which is given by (1) where U_i is the utilization of the ith core and a_i and b_i are the constant value calculated from the collected data.

$$Processor_power = a_i * U_i + b_i \text{ , for each frequency } I \quad (1)$$

In our approach constant parameter of linear relation between frequencies and utilizations are calculated beforehand. We use the parameters in our energy consumption model to select the number of cores and its performance state which minimize the energy consumption.

In order to control the number of active cores, PCPG is used which is a circuit-level technique to cut off the power supply to a core. Obviously there are overheads for switching on and off a gated core. In [16] and [17] have discussed the overhead costs of the per core power gating, and demonstrate importance of guard mechanisms in using PCPG in order to ensure power savings. We also consider the overhead of the PCPG to achieve power saving by turning any core off.

In this paper we use DVFS and PCPG simultaneously, so our approach must consider both of them. In each iteration we should decide about the number of active cores and related frequency, therefore energy consumption for target active cores and target frequency by considering the overhead for core gating should be calculated. Energy consumption is calculated by (2) which E_p is the consumed energy in the specified performance state and E_g is the overhead for turning cores on or off .

$$E = E_p + E_g \quad (2)$$

Assume a server with „N' CPUs which all VMs have been consolidated to run on „n' active CPUs and that the remaining CPUs are power gated. The average CPU consumption of the ith CPU (E_i) is intrinsically related to the average CPU utilization (U_i) and the time interval. It can be modeled as follows:

$$E_p = \sum_i^n P_i * \mathrm{T} = \sum_i^n (a_i * U_i + b_i) * \mathrm{T} \quad (3)$$

Where the T is the control period. For the calculation of E_g if $core_{target}$ is bigger than $core_{current}$ we have (4) otherwise E_g equals zero.

$$E_g = (core_{target} - core_{current}) * E_{overhead} \quad (4)$$

V. Energy aware mechanism

In this section we present an energy aware algorithm to decide about the number of active cores and appropriate frequency simultaneously. So, due to assess required resources in different frequencies, there is need to find the relationship between utilization and frequency. In the case of a core, increasing the frequency will result decrease in utilization. In fact there is linear relationship between frequency and utilization which can be illustrated as:

$$U_i F_i = U_j F_j \quad (5)$$

Where U_i is the utilization of the core in frequency of Fi. Web servers are multi-thread applications which its load can be distributed to more CPUs. So, for the web servers we can extend (5) to (6).

$$\sum_{\mathrm{N}=0}^{C_i} U_{in} F_{in} = \sum_{n=0}^{C_j} U_{jn} F_{jn} \quad (6)$$

Where C_i and C_j is the number of cores in state i and j, respectively. For the simplicity we assume all the cores work on the same frequency and having the same utilization.

$$U_i F_i C_i = U_j F_j C_j \quad (7)$$

Additionally as we demonstrated in the previous works [21] and [22], in virtualized servers there is contention between VMs in getting the physical resources. Our research demonstrated that for the ratio of vCPUs to physical CPUs bigger than a threshold, contention of VMs increase significantly. So, in this algorithm $c_{minPossible}$ is the minimum number of cores for preventing significant VM performance degradation due to resource contentions in an overbooked server.

Input: *N (total number of CPUs), f_n (frequency), c_n (the number active CPUs), T_h (high threshold), and u_n (average utilization)*
Output: *f_{n+1} (frequency), c_{n+1} (the number of active CPUs)*

```
energy = max;
for core c_{n+1} in [c_minPossible, c_minPossible+1, ... , N]
    if U_{n+1}(f_max) < T_h then //need more core
        for frequency f in all frequencies (ascending order) do
            if u_{n+1} > T_h then //need more frequency
                continue;
            energy_i = calculateEnergy(f, u_{n+1}, c)
            if energy_i < energy then
                energy = energy_i
                f_{n+1} = f
                c_{n+1} = core
```

Fig3. Energy aware algorithm

In each iteration we assume that the amount of needed resources is the same with the previous cycle. In each cycle we choose the best number of active cores and specified frequency based on the workload amount in current iteration (Predicting the amount of consumed energy is out of the scope of this paper) in order to minimize E in (2). All of the possible states (frequency and number of active cores) are calculated and the best one is chosen. In each iteration, based on the predicted workload we can calculate the utilization of each core based on (7). The calculation is done by the presented energy model in previous section in which the overhead of changing the number of active cores in addition to energy consumption of active cores. You can find our algorithm on Fig 3.

$U_{n+1}(f_{max})$ represents utilization for the max frequency of current cores, if it is bigger than up threshold, then current cores cannot satisfy the workload, Hence we have to increase number of cores. We calculate all of the possible states based on the number of cores and possible p-states to minimize energy consumption. The method *calculateEnergy* calculates the energy consumption based on the selected frequency, load and the number of active cores and if the calculated energy is less than the previous states then the parameters are set.

VI. Experimental results

In this section, first we present our experimental testbed and benchmark which generates workload to evaluate the algorithm then three existing consolidation approaches are introduced and then the experimental testbed results are proposed. Next our algorithm is compared with the approaches which due to equal comparison conditions we also implemented them in Xen kernel. Finally, the physical testbed results are presented.

Our server is equipped with two Intel Xeon X5650 processor and 8 GB RAM. Each one of the CPUs has 6 physical cores having 9 frequency levels: 2.667GHz, 2.533Gz, 2.4GHz, 2.267GHz, 2.133GHz, 2GHz, 1.867GHz, 1.733GHz and 1.6GHz. Our processors support PCPG as the deepest sleep state.

Xen [23] is the widely used virtual machine manager in big companies like Amazon. We used Xen 4.4 latest available version as the hypervisor to host VMs. Our algorithm is implemented in Xen kernel which have direct access to resources for measurement and manipulating power management knobs. To avoid limitations of the network in experiments, we put the workload clients on the server, too. Each server hosts 4 virtual machines as clients and 4 web server virtual machine altogether with dom0. All the virtual machines run Ubuntu 14.04 as OS with 20GB of hard disk space. Each web server virtual machine has 2 vCPUs and 800 MB RAM but each client machine has 1 vCPU and 600 MB RAM.

We chose RUBiS (Rice University Bidding System) [24], an auction website benchmark modeled after eBay to simulate real web servers. In RUBiS multiple clients which their number are defined beforehand, connect to the web server and simultaneously do selling, browsing and bidding operations. Each client starts a session and does aforementioned operations randomly after getting the response of each request from the servers. The benchmark starts with up ramp phase which it progressively increases the number of sessions to reach peak load due to not

Table I. Energy efficiency and throughput results.

Approaches	Throughput %	Energy improvement %
Base	1	n/a
Consolidation1	0.97	58
Consolidation2	0.94	55
Our algorithm	0.94	67

creating massive sudden load which is not tolerable by the server. In the steady state that is the second phase, the peak load will be maintained and finally in down ramp the load will be decreased to reach the end of the benchmark.

Our approach is compared with three existing approaches. The first approach is the baseline which does not use any power management knobs. Both the frequency and number of on cores in server is maximum. The two other approach that use PCPG and DVFS simultaneously, are threshold based algorithms. In the first consolidation approach if the required resource is bigger than up threshold, it first checks that there is any gated core or not. If there is, the algorithm will bring the gated core up and otherwise, it will increase the frequency of the cores. For the required resources lower than down threshold, it firstly check lower frequency to satisfy the workload and finally if there is no lower frequency, it will turn the cores off to satisfy the workload [11]. All in all, this algorithm tries to keep the cores as lowest possible frequency.

The second consolidation algorithm, conversely tries to keep the number of cores at the lowest possible value. In the case of need for more resources, first it tries to increase the frequency to meet the required resource and if the frequency is in the highest possible frequency, it will increase the number of cores. In the case of lower resource need than down threshold, it will try to decrease the number of cores first. Then if lower frequency could not be found, the number of cores will be decreased.

Table I illustrates the energy efficiency and throughput of each test proportion to the baseline result. Throughput of each test is defined as the ratio of the requests to duration of the benchmark in the experiment. The first row illustrates results of the baseline approach defined beforehand. The second and third row are the experimental results of the consolidation algorithms described and the last one is ours.

Even though blindly consolidation algorithms will have significant energy savings but there is still room for minimizing energy consumption. In Table I, since the baseline approach uses maximum resource to response the workload; we normalized values proportion to its results. Our algorithm reduces energy consumption up to 67% and 15% more than consolidation1 algorithm which is the best existing algorithm for consolidation. Note that although our algorithm has maximum energy saving but its throughput is less than consolidation1, because consolidation1 increases core and unfold more resources in facing growth in workload. Obviously we can mitigate the throughput degradation by exploiting accurate prediction mechanism but it is out of the paper scope. The reason why our algorithms and consolidation 2 throughputs are equal is that based on the energy consumption model, both the algorithms operate similar to each other in facing the increase in workload. However, in lowering workload they have different approach and our algorithm have 21% more energy saving respect to consolidate 2.

VII. Conclusion and Future Works

Due to limitations of server consolidation approaches, the servers also suffer from low utilizations of resources. Hence there is a need for an efficient energy management for power of processors. Modern chip multi-processors provide two options to reduce energy consumption: DVFS and PCPG which made CPU consolidation as an effective knob in power management of servers.

Even though CPU consolidation hinders decoupled use of DVFS and PCPG but our experiments on modeling the CPU power consumption in different frequencies and different number of cores demonstrate that for low number of cores scaling out and operating in low frequencies there has more energy savings. Therefore, although blindly consolidation approaches significantly improve energy consumption but the energy saving is not maximized.

Our proposed algorithm provide dynamically consolidate CPUs to minimize energy consumption. We also propose a novel energy model which considers energy overhead of PCPG, too.

Finally, up to 67% energy saving without significant performance degradation is reported in our experimental results. Additionally the proposed algorithm outperforms two existing CPU consolidation algorithms by 15% and 21% in energy saving.

In this paper we didn't consider any specific order for ON cores in our consolidation mechanism. Choosing the right order of active and inactive cores is planned as future work. Since higher density of ON cores results in higher temperature, more energy may be needed for cooling system, and hence different order of active cores is of a great importance.

VIII. References

[1] O. Bilgir, M. Martonosi and Q. Wu, "Exploring the Potential of CMP Core Count Management on Data Ceter Energy Savings," in *Workshop on Energy Efficient Design*, 2011.

[2] G. Dhiman, G. Marchetti and T. Rosing, "vGreen: a System for Energy Efficient Computing in Virtualized Environments," in *Proceedings of the 2009 ACM/IEEE international symposium on Low power electronics and design*, 2009.

[3] R. Nathuji and K. Schwan, "VirtualPower: Coordinated Power Management in Virtualized Enterprise Systems," in *ACM SIGOPS Operating Systems Review*, 2007.

[4] N. Bobroff, A. Kochut and K. Beaty, "Dynamic placement of Virtual Machines for Managing SLA Violations," in *Integrated Network Management, 2007. IM'07. 10th IFIP/IEEE International Symposium on,*, Munich, 2007.

[5] N. Van, F. D. T. Hien and J.-M. Menaud, "Autonomic Virtual Resource Management for Service Hosting Platforms," in *Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing*, 2009.

[6] C. Clark, S. H. Keir Fraser, J. G. Hansen, E. Jul, C. Limpach, I. Pratt and A. Warfield., "Live Migration of Virtual Machines," in *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation-Volume 2*, 2005.

[7] G. Von Laszewski, L. Wang, A. J. Younge and X. He, "Power-aware Scheduling of Virtual Machines in DVFS-enabled Clusters," in *In*

Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on, 2009.

[8] P. Pillai and K. G. Shin, "Real-time Dynamic Voltage Scaling for Low-power Embedded Operating Systems," in *ACM SIGOPS Operating Systems Review*, 2001.

[9] J. Chen, T. Wei and J. Liang., "State-Aware Dynamic Frequency Selection Scheme for Energy-Harvesting Real-Time Systems," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on ,* vol. 22, no. 8, pp. 1679 - 1692, 2014.

[10] Y. Wang, X. Wang, M. Chen and X. Zhu, "Power-Efficient Response Time Guarantees for Virtualized Enterprise Servers," in *Real-Time Systems Symposium, 2008*, Barcelona, 2008.

[11] I. Hwang, T. Kam and M. Pedram, "A study of the Effectiveness of CPU Consolidation in a Virtualized Multi-Core Server System," in *Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design*, 2012.

[12] J. Leverich, M. Monchiero, V. Talwar, P. Ranganathan and C. Kozyrakis, "Power Management of Datacenter Workloads Using Per-Core Power Gating," in *Computer Architecture Letters*, 2009.

[13] K. a. X. W. Ma, "PGCapping Exploiting Power Gating for Power Capping and Core Lifetime Balancing in CMPs," in *Proceedings of the 21st international conference on Parallel architectures and compilation techniques*, 2012.

[14] H. Jiang, M. Marek-Sadowska and S. R. Nassif., "Benefits and Costs of Power-Gating Technique," in *Computer Design: VLSI in Computers and Processors, 2005. ICCD 2005. Proceedings. 2005 IEEE International Conference on*, San Jose, CA, USA, 2005.

[15] J. Lee and N. S. Kim, "Optimizing Throughput of Power- and Thermal-Constrained Multicore Processors Using DVFS and Per-Core Power-Gating," in *Design Automation Conference, 2009. DAC '09. 46th ACM/IEEE*, San Francisco, CA, 2009.

[16] M. Annavaram, "A Case for Guarded Power Gating for Multi-Core Processors," in *High Performance Computer Architecture (HPCA), 2011 IEEE 17th International Symposium on*, San Antonio, TX, 2011.

[17] N. Madan, A. Buyuktosunoglu, P. Bose and M. Annavaram, "Guarded Power Gating in a Multi-Core Setting," in *Computer Architecture*, 2010.

[18] I. Hwang and M. Pedram, "CPU Consolidation versus Dynamic Voltage and Frequency Scaling in a Virtualized Multi-Core Server," in *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 2013.

[19] A. Vega, A. Buyuktosunoglu, H. Hanson, P. Bose and S. Ramani., "Crank It Up or Dial It Down: Coordinated Multiprocessor Frequency and Folding Control," in *Proceedings of the 46th Annual IEEE/ACM International Symposium on Microarchitecture*, 2013.

[20] J. B. Leverich, "Future Scaling of Datacenter Power-efficiency," Stanford University, 2014.

[21] M. Sharifi, H. Salimi and M. Najafzadeh., "Power-efficient distributed scheduling of virtual machines using workload-aware consolidation techniques," *The Journal of Supercomputing,* vol. 61, no. 1, pp. 46-66, 2012.

[22] H. Salimi and M. Sharifi, "Batch scheduling of consolidated virtual machines based on their workload interference model," *Future Generation Computer Systems,* vol. 29, no. 8, pp. 2057-2066, 2013.

[23] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt and A. Warfield, "Xen and the Art of Virtualization.," in *ACM SIGOPS Operating Systems Review*, 2003.

[24] "http://rubis.ow2.org/," RUBiS: Rice University Bidding System. [Online].

Human Posture Detection Based on Human Body Communication with Muti-carriers Modulation

Wenshu Ni[1], Yueming Gao[2], Zeljka Lucev[3], Sio Hang Pun[4], Mario Cifrek[5], Mang I Vai[6] , Min Du[7]
[1,2,4,6,7]Fujian Key Laboratory of Medical Instrument and Pharmacy Technology, Fuzhou University, Fuzhou,China
[3,5]Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
[1]595317255@qq.com, [2]fzugym@163.com, [3]zeljka.lucev@fer.hr

Abstract – Multi-node sensors for human posture detection, by acquiring kinematic parameters of the human body, helps to further study the laws of human motion. It can be used as reference for the quantitative analysis tool for some specific application, such as healthcare, sports training and military affairs, etc. Compared with the traditional optical method, posture detection based on the inertial sensors has smaller limitation of space, lower cost, and easier implementation. In this paper, a human posture detection system was introduced. Utilizing the parameter data obtained by the inertial sensors, three-dimensional angles of the human hand movement could be calculated via quaternion algorithm for data fusion. The angles data transmission among the sensor nodes was successfully realized by the human body communication(HBC) transceivers based on capacitive coupling of multi-carriers OOK modulation at the data rate of 57.6kbps. The bit error rate(BER) was less than 10^{-5}. The human posture could be reconstructed on the PC host. Ultimately, the implementation of the overall result showed the feasibility of the system.

Keywords - Inertial sensors, Human body communication, Multi-carriers OOK modulation, Posture reconstruction

I. INTRODUCTION

In recent years, human action recognition combined with specific applications such as virtual reality, health care, sports training, has caught much attention. Human posture is closely related to some disease characteristics. Detection will be further helpful for some specific condition as a quantitative tool. Physical therapy with quantitative approaches for geriatrics training or Parkinson's patients for example[1]. Due to the improvements of Micro-Electro-Mechanical system (MEMS)[2][3], the size, accuracy, robustness and dynamic response are greatly improved. Compared with the traditional optical method, inertial sensors for posture detection has lower cost, less space requirement, and easier implementation.

However, traditional means of communication among the inertial sensors are often based on Bluetooth, RF, Zigbee or WIFI[4], which have high power consumption and are likely to cause radiation and wireless signal aliasing. To reduce these impacts, the human body communication(HBC) has been introduced. HBC utilizes the human body as signal channel to transmit the data among the nodes. Capacitive coupling method of HBC transmits signals by generating an electric field so that the sensors attached on or implanted in the human body could share the information. Methods of HBC system are divided into two types. The first is the direct transmission of digital square wave signal without modulation[5][6], and the second is a modulation scheme with carrier frequency[7]. While the latter one is more conducive to achieve high data rate[8] and multichannel transmission[9].

In this paper we proposed a novel approach to the human posture detection system design. In order to save the channel resources and reduce the radiation impact, taking the direct measurement of the human body posture into account as well, capacitive coupling method of HBC with multi-carriers on-off keying(OOK) modulation was considered to achieve the two ways transmission of digital signals. The circuit of OOK modem has been proposed, which was capable to transmit the digital signal at the rate of 57.6kbps and the BER was less than 10^{-5}. Besides, multi-carriers modulation method was adopted to realize the two channels parallel transmission in real time.

II. SYSTEM DESIGN

Body posture detection system can detect the precise angle information of human action, and play an important role as a quantitative tool for some specific application. In this paper, a human gesture detection system design was proposed. Three inertial sensors was capitalized on to collect and process information of the posture angle. HBC was responsible for the two ways sensor data transmission among the sensors at the data rate of 57.6kbps. Eventually, the reconstruction operation was implemented on the PC.

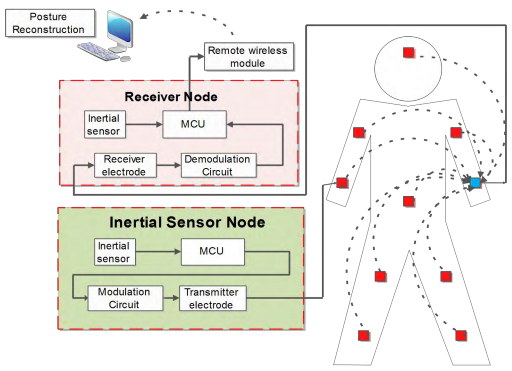


Fig.1 Diagram of the proposed human posture detection system

The structure of the system can be showed in Fig. 1, it comprises of: (1) Inertial sensor nodes, (2)HBC

device, (3)Reconstruction of body posture. To begin with, the inertial sensor node, which consists of accelerometer, gyroscope and magnetometer, will detect the kinematics data of the body. Fig.1 shows the structure of the sensor. In this paper, we had three sensors put on the up arm, lower arm and hand respectively to get the kinetic information of the three parts. Then algorithm of data fusion will figure out the three-dimensional angles of three parts in the free space. The capacitive coupling HBC was used for wireless transmission. Only the signal electrodes of TX and RX are attached to the body while the ground (GND) electrodes are left floating. The TX establishes the quasi-static electric field, while the RX detects the electric potential at the remote end in this application[10], therefore, the information could be shared between TX and RX, which performs less attenuation than the galvanic coupling. As shown in the Fig.2. In this mechanism, the node 1 got the upper arm posture data and then transmitted it with carrier 1 modulation. The node 2 received the data via the HBC receiver 1 and transmitted the data of node 2 and node 1with carrier 2 frequency modulation. Finally, the node 3 received former two channel signal via HBC receiver 2, and then the received signals together with the data of node 3 would be used to reconstruct the posture on the PC. Therefore a tree structure was formed to detect the arm and hand posture. The transceivers position setup on the human body is as Fig.5.

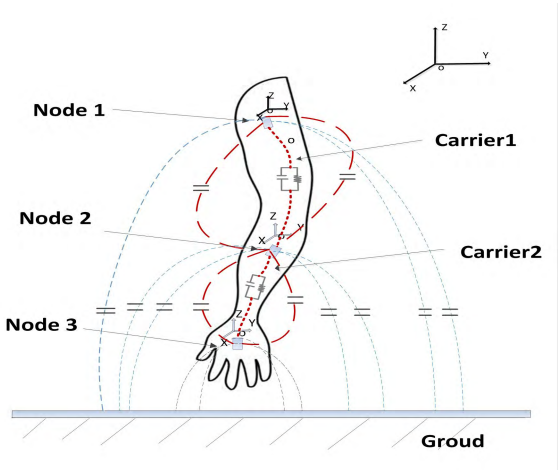


Fig.2 The HBC mechanism

III. METHOD

A. Calculation of Angles

In this system, inertial sensor module MPU6050, was adopted considering its convenience in wearable property. With sensor attached, the arm movement for example can be regarded as the coordinate change of the sensor model. Therefore the three-dimensional angles of arm is equivalent to the module's, which could be expressed by quternion as shown in (1). Equation (2) is the carrier's transformation matrix. With the data obtained by the accelerometer, gyroscope and magnetometer, the transformation matrix could be corrected using error vector products. Then the quaternion is refreshed by the numerical integration method, and the RungeKutta method is described as (3). Finally, due to the transformation of quaternion being equivalent to the Euler's, the Roll, Pitch and Yaw angles can be calculated eventually, following the (4)-(6).

$$Q=q_0+q_1i_0+q_2j_0+q_3k_0=\cos\frac{\theta}{2}+(li_0+mj_0+nk_0)\sin\frac{\theta}{2}, \quad (1)$$

$$C_b^R(q)=\begin{bmatrix}1-2(q_2^2+q_3^2) & 2(q_1q_2-q_0q_3) & 2(q_1q_3+q_0q_2)\\ 2(q_1q_2+q_0q_3) & 1-2(q_1^2+q_3^2) & 2(q_2q_3-q_0q_1)\\ 2(q_1q_3-q_0q_2) & 2(q_2q_3+q_0q_1) & 1-2(q_1^2+q_2^2)\end{bmatrix}, \quad (2)$$

$$\begin{bmatrix}\dot{q}_0\\ \dot{q}_1\\ \dot{q}_2\\ \dot{q}_3\end{bmatrix}=\frac{1}{2}\begin{bmatrix}0 & -g_x & -g_y & -g_z\\ g_x & 0 & g_z & -g_y\\ g_y & -g_z & 0 & g_x\\ g_z & g_y & -g_x & 0\end{bmatrix}\begin{bmatrix}q_0\\ q_1\\ q_2\\ q_3\end{bmatrix}, \quad (3)$$

$$Yaw=\arctan\frac{\sin\varphi}{\cos\varphi}=\arctan\left[\frac{2q_1q_2+2q_0q_3}{1-2(q_2^2+q_3^2)}\right], \quad (4)$$

$$Pitch=\arcsin(2q_1q_3-2q_0q_2), \quad (5)$$

$$Roll=-\arctan\left[\frac{2q_2q_3+2q_0q_1}{1-2(q_1^2+q_2^2)}\right], \quad (6)$$

B. Human body communication of OOK multi-carriers modulation

In order to transmit multiple signals, we use a multi-carrier OOK modulation method based on frequency division multiplexing principle. As shown in the Fig.3, the total bandwidth of the human body channel was divided into many sub-bands(or subchannels) to transmit multiple signals with different carrier frequency modulation, while the receiver end would separate the channels, so that multi-channel parallel transmission of signals without interference could be possible. The energy of electromagnetic wave is mainly in the form of electric field when its frequency is below 10MHz in the capacitive coupling of HBC, which performs less radiation and better transmission characteristics[11]. In this system, we use 6MHz, 9MHz as carriers frequency. The narrow band pass filters with center frequency of the carriers separate the two channel signals and removed the unwanted noise.

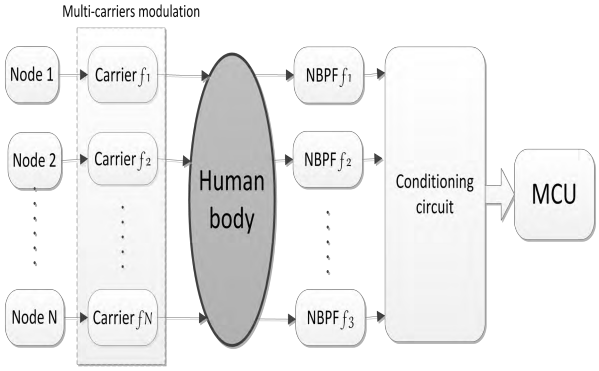


Fig.3 Multi-carriers modulation of HBC

HBC on-off keying modulation scheme is shown in Fig.4. The HBC transceiver realizes the OOK modulation by sinusoidal oscillator circuit and analog switch. We produce the frequency of 9 MHz or 6MHz 3V amplitude sine wave. The digital signal is input to the TXIN. When the TXIN is '1', frequency sine wave is output, and when the TXIN is '0', no signal is output. Thus the OOK modulation is realized. In the receiver, signals suffered attenuation and distortion. Chebyshev passive band pass filter has been employed because of its fast attenuation of the transition band to remove the unwanted signals. The center frequency of the filter is carrier frequency. Then the envelop detector together with the comparator could demodulate the signal and control the electrical level in the range of TTL level, making sure the MCU of the receiver could get the digital signal via the serial port directly.

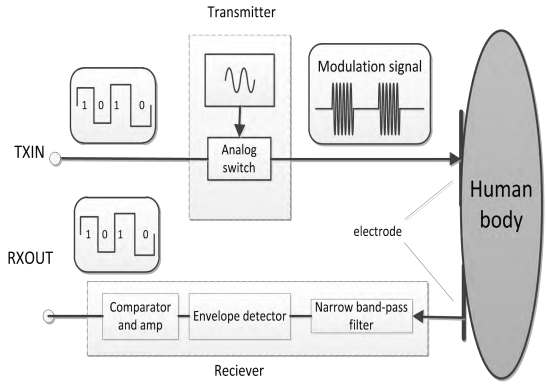


Fig.4 The OOK modulation scheme

IV. Result and Discussions

A. Experiment of HBC

In the measurement experiment, all of the circuits board were powered by battery in order to separate the transmitter node and the receiver node; the Textronix2024 oscilloscope was used in the experiment, which is self-powered by its own battery inside, could keep apart two passageways in the round as well. A 24-years old male's left arm was used as experimental subject. The medical electrode's size was 4cm×4cm.

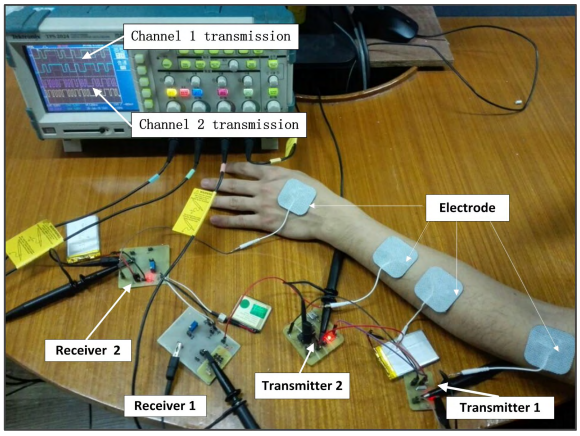


Fig.5 HBC devices experiment

Fig.5 shows the measurement setup and the wave result. The electrodes of transceivers were attached on the human arm and hand back as described in session II. The data was modulated and transmitted though the human body to the receiver board. The test signal was NRZ type generated by a STM32 serial port with a rate of 57.6kbps. The measured output waveform of recovered data could be seen clearly on the screen of the oscilloscope, showing that the data receivers recovered the modulation signals and separated the two channel signals successfully. The Fig.6 shows the output of the modulation signal and the received signal of transceiver 1 whose carrier frequency is 6MHz. After analysis, the bit error bit was less than 10^{-5} when the data rate was 56.7kbps, indicating the HBC system could work successfully.

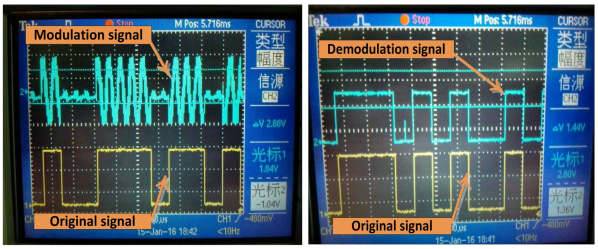


Fig.6 Modulation and demodulation wave form compared with original signal

B. Reconstruction of the posture of human arm and hand

As shown in Fig.7, three inertial sensors are respectively put in the upper, lower arm and the hand back to obtain the attitude angles of the three parts. As a result, the three dimensional angles could be acquired by the computer. The result of reconstruction could be seen in Fig.7, there are three cylinders regarded as the human arm and hand, the red part is the upper arm, and the yellow one is the lower arm, while the purple one is the hand. The result showed that the system based on the inertial sensor was able to measure and reconstruct the human posture.

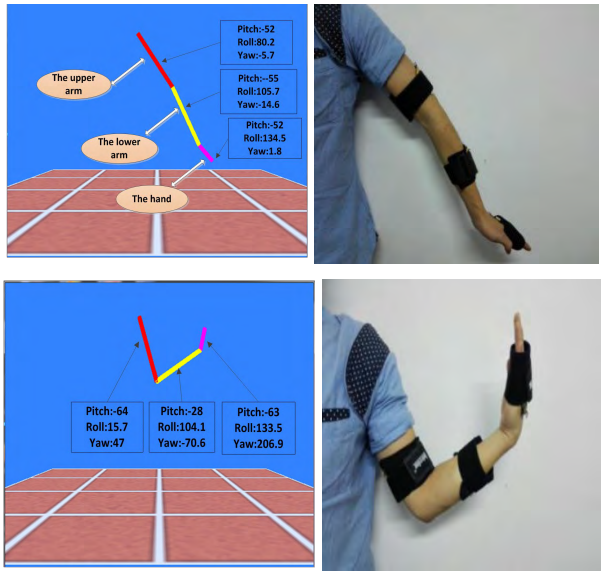


Fig. 7 The reconstruction of human posture

V. CONCLUSION

In this paper, a human posture detection system based on inertial sensors and human body communication was proposed and implemented. Using algorithm of quaternion to calculate the three-dimensional angles of the human arm and hand, the posture information could be obtained. The human body was selected as propagation medium for two channel data transmission. Capacitive coupling method of HBC was considered, besides, the frequency of 6MHz and 9MHz were chosen as multi-carriers frequency in HBC devices of OOK modulation. The transceivers were implemented and the BER was less than 10^{-5} when the data rate was 57.6kbps as system designed. The arm and hand posture could be reconstructed on computer so that the results are visible and quantitative for some specific application.

ACKNOWLEDGMENT

The authors would like thank to the Ministry of Science Foundation of China 2013DFG32530, the National Natural Science Foundation of China 61201397 and the Funds of the Department of Education of Fujian Province, China, JA13027.

REFERENCES

[1] Fay B. Horak, PhD, PT, and Martina Mancini, PhD. Objective Biomarkers of Balance and Gait for Parkinson's Disease Using Body-worn Sensors. BALANCE AND GAIT BIOMARKERS, Vol.38, NO.11, pp.1544-1551,2013.

[2] N. Barbour and G. Schmidt, "Inertial sensor technology trends," IEEE Sensor J., vol. 1, no. 4, pp. 332–339, Dec. 2001.

[3] J. Barton, A. Lynch, S. Bellis, B. O'Flynn, F. Murphy, K. Delancy, and S. C. O'Mathuna, "Miniaturized inertial measurement units (IMU)for wireless sensor networks and novel display interfaces," in Proc. Electronic Components and Technol. Conf., 2005, pp. 1402–1406.

[4] Yao-Chiang Kan, Chun-Kai Chen. "A Wearable Inertial Sensor Node for Body Motion Analysis." IEEE SENSORS JOURNAL, VOL. 12, NO. 3, MARCH 2012.

[5] C. H. Hyoung, J. B. Sung, J. H. Hwang, J. K. Kim, D. G. Park, S. W.Kang, "A novel system for intrabody communication: Touch-And-Play," Proc. ISCAS, pp. 1343–1346, May.21-23, 2006.

[6] Linlin Zhang, Yueming Gao. "Design of Human Motion Detection Based on the Human Body Communication," In Proc, IEEE TENCON 2014, Macau SAR, China. Nov. 2015. pp. 1-4.

[7] K. Hachisuka, A. Nakata, T. Takeda, Y. Terauchi, K. Shiba, K. Sasaki,H. Hosaka, and K. Itao, "Development and performance analysis of an intra-body communication device," in Proc. 12th IEEE Int. Conf. Solid-State Sens., Actuators, Microsyst., pp. 1722–1725, Jun, 2003.

[8] T. Leng, Z. Nie, W. Wang, F. Guan, and L. Wang, "A human body communication transceiver based on on-off keying modulation," In Proc., IEEE isbb 2011, China, Nov. 2011, pp. 61-64.

[9] Ž. Lčev, I. Krois, and M. Cifrek, "A multichannel wireless EMG measurement system based on intrabody communication," XIX IMEKO World Congress, Fundamental and Applied Metrology, Lisbon, Portugal, pp. 1711-1715, September 2009.

[10] Ruoyu Xu, Hongjie Zhu. Electric-Field Intrabody Communication Channel Modeling With Finite-Element Method. IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 58, NO. 3, MARCH 2011

[11] Joonsung Bae, Hyunwoo Cho, Kiseok Song, Hyungwoo Lee and Hoi-Jun Yoo. The Signal Transmission Mechanism on the Surface of Human Body for Body Channel Communication, IEEE Transactions on Microwave Theory and Techniques, 2012, 60(3): 582-593.

SAT-based Search for Systems of Diagonal Latin Squares in Volunteer Computing Project SAT@home

Oleg Zaikin, Stepan Kochemazov, Alexander Semenov
Matrosov Institute for System Dynamics and Control Theory SB RAS, Irkutsk, Russia
Email: zaikin.icc@gmail.com, veinamond@gmail.com, biclop.rambler@yandex.ru

***Abstract*—In this paper we considered the problem of finding pairs of mutually orthogonal diagonal Latin squares of order 10. First we reduced it to Boolean satisfiability problem. The obtained instance is very hard, therefore we decomposed it into a family of subproblems. To solve the latter we used the volunteer computing project SAT@home. In the course of 10-month long computational experiment we managed to find 29 pairs of described kind, that are different from already known pairs. Also we considered the problem of search for triples of diagonal Latin squares of order 10 that satisfy weakened orthogonality condition. Using diagonal Latin squares from the known pairs (the most of them were found in SAT@home) we constructed new triples of proposed kind. During this computational experiment we used a computing cluster.**

I. Introduction

The combinatorial problems related to Latin squares pose interest for mathematicians since Leonard Euler. A lot of general information about these problems can be found in [1]. Latin square of order n is a square $n \times n$ table filled with elements from some set M,$|M| = n$ in such a way that each element from M appears in each row and each column exactly once. Initially Leonard Euler used the set of Latin letters as M, therefore the corresponding combinatorial designs were named Latin squares. In this paper for convenience we will use as M the set $\{0, \ldots, n-1\}$. The Latin square is called diagonal if both its primary and secondary diagonals contain all numbers from 0 to $n-1$. In other words, the constraint on the uniqueness is extended from rows and columns to two diagonals.

A pair of Latin squares of the same order is called orthogonal if all ordered pairs of the kind (a, b) are different, where a is the number in some cell of the first Latin square and b is the number from the same cell in the second Latin square. If there are m different orthogonal Latin squares, from which each pair is orthogonal, then it is called the system of m mutually orthogonal Latin squares (MOLS). One of the most well-known unsolved problems in this area is the following: to answer the question whether there exists a triple of MOLS of order 10.

There are many different approaches to solving problems regarding combinatorial designs. In the present paper we use the so-called SAT approach [2]. Basically, it means that we reduce the original problem to Boolean satisfiability problem (SAT) and then apply state-of-the art SAT solvers to an obtained instance. The attractiveness of this approach is justified by the fact that a lot of problems from different areas (for example, software verification, cryptography or bioinformatics) can be effectively reduced to SAT. Despite the fact that SAT is NP-hard, and all known algorithms for solving SAT are exponential in the worst case scenario, state-of-the-art heuristic algorithms manage to solve SAT instances encoding practical problems from various areas in reasonable time. The majority of such fast SAT solvers are based on the Conflict-Driven Clause Learning paradigm (CDCL) [3].

Evidently, among SAT instances there are very hard variants and to solve them in reasonable time it is necessary to involve significant amounts of computational resources. That is why the improvement of the effectiveness of SAT solving algorithms, including the development of algorithms that are able to work in parallel and distributed computing environments, is a very important direction of research. In 2011 we launched the volunteer computing project SAT@home aimed at solving hard SAT instances [4]. One of the aims of the project is to find new combinatorial designs based on the systems of orthogonal Latin squares.

Let us briefly outline the paper. In the second section we describe the volunteer computing project SAT@home. In the third section we detail how we apply SAT approach to finding pairs of mutually orthogonal diagonal Latin squares (MODLS) of order 10 in SAT@home. In the fourth section we show how the pairs of MODLS found in SAT@home can be used to search for triples of diagonal Latin squares of order 10 that satisfy weakened orthogonality condition and discuss the results of our computational experiment, performed using a computing cluster.

II. Volunteer Computing Project SAT@home

Volunteer computing [5] is one of the types of distributed computing. Its defining characteristic is that it uses computational resources of volunteers PCs. The majority of volunteers here are usual people, not affiliated with a single organization or group of companies. Usually, one volunteer computing project is designed to solve one or several closely related hard problems. It is important to note that when volunteers PC is connected to the project, all the calculations on it are performed automatically and do not inconvenience user since for this purpose only idle resources of PC are employed.

Another distinctive feature of volunteer computing projects is that they can use only embarrassing parallelism [6], i.e. the original problem should be decomposed into a family of subproblems that can be solved independently from each other. Volunteer computing is very cheap — to maintain a project one needs only a dedicated server and several client applications fitting into one infrastructure. Empirically, the main difficulty here lies in software development and database administration. Also it is crucial to provide a feedback to project users via web site and special forums. An attractive consequence of such structure is that volunteer project can spend several months or even years to solve one hard instance.

The majority of currently functioning volunteer computing projects are based on the Berkeley Open Infrastructure for Network Computing (BOINC) [5], developed in Berkeley in 2002. Overall, at the present moment there are about 70 active volunteer projects. Total performance of these projects exceeds 11 PFLOPs. Structurally, volunteer computing project can be considered as a sum of the following parts: server daemons, database, web site and client applications. Here daemons include work generator (it generates tasks to be processed by volunteers PSs), validator (that checks the correctness of the results received from volunteers PCs) and assimilator (it processes correct results). The client applications should have versions for the widespread computing platforms.

In 2011 the authors of the paper in collaboration with colleagues from IITP RAS developed the volunteer computing project SAT@home [4]. On February 7, 2012 this project was added to the official list of BOINC projects[1] with alpha status. In 2015 the status was upgraded to beta. The main goal of the project is to solve hard SAT instances from various subject areas. SAT@home is based on the BOINC platform. Currently (as of January 22, 2016) the project has about 3784 active PCs from users all over the world, and has the average performance of 8.4 TFLOPs.

Let us consider the basic features of the SAT@home project in more detail. The work generator daemon is located on the server. It decomposes the original SAT instance into a family of subproblems. For this purpose it uses decomposition parameters, found via the special Monte Carlo method [7] on a computing cluster. On this stage it is necessary to use a computing cluster because the corresponding computational scheme uses a lot of interprocessor exchange (essentially, it employs fine-grained parallelism). According to the concept of redundant calculations used in BOINC, the work generator creates 2 copies of each task to be processed by 2 users from different teams (it decreases the possibility of cheating). The validator daemon checks if both copies of a single task yielded the same result and then gives verified results to the assimilator daemon. If there was found satisfying assignment then the assimilator checks it, and if it is correct, then it marks the original SAT instance as solved. The client application in SAT@home is based on the slightly modified CDCL solver MiniSat [8].

[1] http://boinc.berkeley.edu/projects.php

In May, 2012, the six-month computational experiment, aimed at solving 10 instances of A5/1 keystream generator cryptanalysis, successfully ended in SAT@home [7]. Note that in this experiment there were considered only instances, that could not be solved via known Rainbow tables. For each instance the goal was to find unknown initial values of generator registers. In 2014 in SAT@home there were successfully solved five weakened cryptanalysis instances for the Bivium keystream generator [9]. In each instance the values of 9 out of 177 bits corresponding to initial values of generator registers (at the end of initialization phase) were known to make them solvable in reasonable time.

In general, the SAT@home project is a powerful tool for solving hard SAT instances. In the next sections we will describe how we apply it to the search for combinatorial designs.

III. Finding Pairs of Mutually Orthogonal Diagonal Latin Squares of Order 10 in SAT@home

The existence of a pair of MODLS of order 10 was proved in 1992 — in the paper [10] three such pairs were presented. In 2012 we started in SAT@home the computational experiment aimed at finding new pairs of MODLS of order 10. The experiment was finished in 2013 and its results were published in [11]. First we constructed a propositional encoding for this problem. The obtained CNF had 2000 Boolean variables and 434440 clauses. The size of the corresponding Conjunctive normal form (CNF) in the DIMACS format was 10 Mb. We used the so-called "naive" encoding (for example, see [12]). We decomposed the obtained SAT instance as follows. The first row of the first diagonal Latin square was fixed to be equal to "0 1 2 3 4 5 6 7 8 9". It does not lead to the loss of generality because of the properties of Latin squares. After this we processed all possible values of the first 8 cells of the second and the third rows. As a result we obtained a family of subproblems, in which for each subproblem for the first diagonal Latin square the values were fixed for 26 out of 100 cells (10 from the first row and 8 from the second and the third rows). In terms of CNF it translated into assigning values to 260 out of 2000 variables in each SAT instance. In SAT@home experiment each job batch contained 20 such SAT instances. For each instance MiniSat had a limit of 2600 restarts that is more or less equal to 4 minutes on one core of state-of-the-art CPU. To process 20 million subproblems generated for the experiment it took about 9 months of work of the SAT@home project (from September 2012 to May 2013). As a result we found 17 new pairs of MODLS of order 10 (in addition to three previously known pairs from [10]).

In April 2015 we started in SAT@home another experiment on the search for new pairs of MODLS of order 10. Note that in the experiment held in 2012 (see [11]) we used the decomposition chosen according to some rational considerations. In the new experiment we decided to use a more systematic approach to choosing the decomposition. In Table 1 we show the data for 7 variants of decomposition (including one suggested in [11]) and the results obtained for them. In

TABLE I
THE RESULTS OF EXPERIMENTS FOR DIFFERENT DECOMPOSITIONS.

Decomposition	Solutions found	Time
1 row, 9 cells	1	1 month in 2015
2 rows, 2 cells each	-	1 day in 2015
2 rows, 3 cells each	-	3 days in 2015
2 rows, 4 cells each	-	2 weeks in 2015
2 rows, 5 cells each	23	4 months in 2015-2016
2 rows, 6 cells each	5	3 months in 2015-2016
2 rows, 8 cells each	17	9 months in 2012-2013

total from April 17, 2015 to February 7, 2016 we managed to find 29 new previously unknown pairs of the considered kind (compared to 3 pairs from [10] and 17 pairs from [11]). All found solutions are available online at the web site of the SAT@home project. In this experiment we used the same client application with the same limit on the amount of restarts as in 2012. In all decompositions we considered first cells from the left (it means that since we always fix the first row, then we vary the rows starting from the second).

Let us comment the Table 1. The decomposition "2 rows, 8 cells each" corresponds to the decomposition from [11]. All the other 6 decompositions have been tested in 2015-2016. The decomposition "2 rows, 2 cells each", "2 rows, 3 cells each" and "2 rows, 4 cells each" did not yield any results under the experiment rules. The decomposition "2 rows, 5 cells each" turned out to be more effective than "2 rows, 8 cells each" used in [11], because it made it possible to find more new pairs per time unit (even if we take into account that the performance of the SAT@home project in 2015 was twice that in 2012). It also was the most effective compared to all other decompositions we tested in 2015-2016. Using "2 rows, 6 cells each" we managed to find several new pairs, but ifs effectiveness compared to "2 rows, 5 cells each" turned out to be low, that is why the corresponding experiment was suspended.

IV. FINDING TRIPLES OF MUTUALLY PARTIALLY ORTHOGONAL DIAGONAL LATIN SQUARES OF ORDER 10 USING COMPUTING CLUSTER

Let us consider a triple of Latin squares of the same order. Among all possible sets of ordered pairs of elements, that form when we look at this in the same manner as when we check the orthogonality condition for all three pairs of squares simultaneously, we choose the set with maximal power. Let us refer to the power of this set as characteristics of the considered triple of Latin squares. Currently the record triple of mutually partially orthogonal Latin squares of order 10 in this notation is the one published in [13]. In this triple square A is fully orthogonal to squares B and C, but squares B and C are orthogonal only over 91 pairs of elements out of 100. It means that in our notation the characteristics of this triple is 91. We are focused on finding triples of diagonal Latin squares of order 10, that have high value of characteristics.

In [11] we suggested to consider the problem of constructing triples of MODLS of order 10 in the form of SAT problem. We proposed a propositional encoding, in which by editing a special clause it was possible to set up the desired value of characteristics. When we add to this CNF the known values of two known Latin squares that form an orthogonal pair, it is reduced to finding values of remaining unknown Boolean variables in such a way that the resulting third diagonal Latin square forms with the considered known pair the partially orthogonal triple with desired value of characteristics. For each known pair (at that moment there were only 20) we constructed a separate CNF by assigning values to Boolean variables corresponding to the elements of known pair. We employed the multi-threaded CDCL solver TREENGELING [14], that was launched on every obtained SAT instance with time limit equal to 1 hour on one computing node of "Academician V.M. Matrosov" computing cluster of ISC SB RAS[2]. One node of this cluster contains 32 CPU cores and 64 Gb RAM. Using this approach it was possible to find the triple of proposed kind of squares of order 10 with characteristics equal to 73 (i.e. only for 1 considered SAT instance TREENGELING managed to find a satisfying solution, calculations on the other SAT instances were interrupted). It should be noted, that the triple of MODLS of order 10 from [10] has characteristics 60, but in that triple 2 pairs of squares out of 3 are orthogonal (opposite to 1 out of 3 in the triple from [11]).

Since in 2015-2016 we found 29 more new pairs of MODLS of order 10, we constructed 29 more SAT instances in addition to 20 considered in [11]. UsingTREENGELING with the same time limit (1 hour) we found two more partially orthogonal triples with characteristics equal to 73 (see (1) and (2)).

$$\begin{array}{lll} 0123456789 & 0123456789 & 3825079164 \\ 1902634875 & 2438910567 & 1743862059 \\ 4276398510 & 6981074235 & 6587431290 \\ 9018573642 & 5602137894 & 8204697531 \\ 3795182064 & 8516703942 & 4690583712 \\ 5647021398 & 3079865421 & 7462918305 \\ 6381245907 & 7264589310 & 2976150843 \\ 7569810423 & 4850392176 & 5031246978 \\ 8450967231 & 9347621058 & 9158304627 \\ 2834709156 & 1795248603 & 0319725486 \end{array} \quad (1)$$

$$\begin{array}{lll} 0123456789 & 0123456789 & 9310475826 \\ 1290378564 & 9567241803 & 3245809671 \\ 4869201357 & 5390784162 & 6508347219 \\ 2971865043 & 1258037694 & 5081623794 \\ 3607589421 & 8719325046 & 7963581042 \\ 9584137206 & 7941560328 & 1497062538 \\ 5732614890 & 4876102935 & 8629754103 \\ 6348720915 & 2085693471 & 2754198360 \\ 8015943672 & 6432879510 & 0172936485 \\ 7456092138 & 3604918257 & 4836210957 \end{array} \quad (2)$$

[2]http://hpc.icc.ru

These triples are based on the 5-th and 16-th pairs of MODLS of order 10 found in the experiment respectively (in each triple the first two diagonal Latin squares are from the corresponding pair of MODLS).

The propositional encoding proposed in [11] made it possible to obtain one more class of results. In this encoding we can set the constraint that the characteristics of the triple is 100, thus formulating the problem of search for a triple of MODLS of order 10. In this case when we augment the SAT instance with the known values of cells of two squares, the problem transforms to the following: for a given pair of MODLS of order 10 to find the third diagonal Latin square that forms a triple of MODLS of order 10, or to prove that there is no such square. Thus we constructed 49 SAT instances with assignments of values corresponding to 49 known pairs. For each instance there was added the constraint specifying that the characteristics of the triple is equal to 100. On average to solve one such SAT instance the PLINGELING CDCL solver [14] took about one second, using 32 processor cores (one computing node of the Academician Matrosov cluster of ISC SB RAS). In all cases it was proven that there is no square satisfying specified limitations (this is a consequence of the fact, that all CNFs were unsatisfiable). We even managed to significantly improve this result. When we reduced the specified value of characteristics, the SAT instances considered were getting progressively harder. At the present moment we proved for 49 known pairs that they cannot be used to construct sets of three partially orthogonal diagonal Latin squares of order 10 with the characteristic equal to 87. The solving of corresponding SAT instances took PLINGELING SAT solver on average about 8 hours per instance on 1 cluster node.

V. Related Works

There are several examples of application of high performance computing to the search for combinatorial designs based on Latin squares. In [15] there was proven that there is no finite projective plane of order 10 via special algorithms based on constructions and results from the theory of error correcting codes. The corresponding experiment took several years. On its final stage there was used quite a powerful (at that moment) computing cluster. More recent example is the proof of hypothesis about the minimal number of clues in Sudoku [16] where special algorithms were used to enumerate and check all possible Sudoku variants. To solve this problem a modern computing cluster had been working for almost a year. The volunteer computing project Sudoku@vtaiwan [17] was used to confirm the solution of this problem.

In [18] there was described the application of SAT solvers to finding systems of orthogonal Latin squares. Among other things, the author of [18] used a specially constructed small desktop grid for more than 10 years in an attempt to find a triple of MODLS of order 10. Unfortunately, the experiment did not yield any success.

Apparently, [19] became the first paper about the use of a desktop grid based on the BOINC platform for solving SAT. It did not evolve into a publicly available volunteer computing project (like SAT@home did).

VI. Conclusion

In the paper we describe the results obtained by applying high performance computing to SAT-based searching for systems of diagonal Latin squares of order 10. Using SAT@home we found 29 pairs of MODLS of order 10 (in addition to 20 previously known pairs). Using a computing cluster we found two triples of partially orthogonal diagonal Latin squares of order 10 with characteristics equal to 73 (in addition to one such triple found earlier). We also proved that based on all 49 known pairs of MODLS of order 10 it is impossible not only to construct a triple of MODLS of order 10, but even to construct a triple of partially orthogonal diagonal Latin squares of order 10 with characteristics equal to 87.

Acknowledgment

The research was funded by Russian Science Foundation (project No. 16-11-10046).

References

[1] C. J. Colbourn and J. H. Dinitz, *The CRC handbook of combinatorial designs*. CRC Pr I Llc, 1996.

[2] A. Biere, M. J. H. Heule, H. van Maaren, and T. Walsh, Eds., *Handbook of Satisfiability*, ser. Frontiers in Artificial Intelligence and Applications. IOS Press, February 2009, vol. 185.

[3] J. Marques-Silva, I. Lynce, and S. Malik, *Conflict-Driven Clause Learning SAT Solvers*, ser. Frontiers in Artificial Intelligence and Applications. IOS Press, February 2009, vol. 185, ch. 4, pp. 131–153.

[4] M. Posypkin, A. Semenov, and O. Zaikin, "Using BOINC desktop grid to solve large scale SAT problems," *Computer Science (AGH)*, vol. 13, no. 1, pp. 25–34, 2012.

[5] D. P. Anderson and G. Fedak, "The computational and storage potential of volunteer computing," in *Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2006), 16-19 May 2006, Singapore*. IEEE Computer Society, 2006, pp. 73–80.

[6] I. Foster, *Designing and Building Parallel Programs: Concepts and Tools for Parallel Software Engineering*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1995.

[7] A. Semenov and O. Zaikin, "Using monte carlo method for searching partitionings of hard variants of boolean satisfiability problem," in *Parallel Computing Technologies - 13th International Conference, PaCT 2015, Petrozavodsk, Russia, August 31 - September 4, 2015, Proceedings*, ser. Lecture Notes in Computer Science, V. Malyshkin, Ed., vol. 9251. Springer, 2015, pp. 222–230.

[8] N. Eén and N. Sörensson, "An extensible sat-solver," in *Theory and Applications of Satisfiability Testing, 6th International Conference, SAT 2003. Santa Margherita Ligure, Italy, May 5-8, 2003 Selected Revised Papers*, ser. Lecture Notes in Computer Science, E. Giunchiglia and A. Tacchella, Eds., vol. 2919. Springer, 2003, pp. 502–518.

[9] O. Zaikin, A. Semenov, and I. Otpuschennikov, "Solving weakened cryptanalysis problems for the Bivium cipher in the volunteer computing project SAT@home," in *Second International Conference BOINC-based High Performance Computing: Fundamental Research and Development (BOINC:FAST 2015), Petrozavodsk, Russia, September 14-18, 2015*, ser. CEUR-WS, vol. 1502, 2015, pp. 22–30.

[10] J. Brown, F. Cherry, L. Most, E. Parker, and W. Wallis, "Completion of the spectrum of orthogonal diagonal latin squares," *Lecture notes in pure and applied mathematics*, vol. 139, pp. 43–49, 1992.

[11] O. Zaikin and S. Kochemazov, "The search for systems of diagonal Latin squares using the SAT@home project," in *Second International Conference BOINC-based High Performance Computing: Fundamental Research and Development (BOINC:FAST 2015), Petrozavodsk, Russia, September 14-18, 2015*, ser. CEUR-WS, vol. 1502, 2015, pp. 52–63.

[12] I. Lynce and J. Ouaknine, "Sudoku as a SAT problem," in *International Symposium on Artificial Intelligence and Mathematics (ISAIM 2006), Fort Lauderdale, Florida, USA, January 4-6, 2006*, 2006.

[13] J. Egan and I. M. Wanless, "Enumeration of MOLS of small order," *Math. Comput.*, vol. 85, no. 298, pp. 799–824, 2016.
[14] A. Biere, "Lingeling essentials, A tutorial on design and implementation aspects of the the SAT solver lingeling," in *POS-14. Fifth Pragmatics of SAT workshop, a workshop of the SAT 2014 conference, part of FLoC 2014 during the Vienna Summer of Logic, July 13, 2014, Vienna, Austria*, ser. EPiC Series, D. L. Berre, Ed., vol. 27. EasyChair, 2014, p. 88.
[15] C. Lam, L. Thiel, and S. Swierz, "The nonexistence of finite projective planes of order 10," *Canad. J. Math.*, vol. 41, pp. 1117–1123, 1989.
[16] G. McGuire, B. Tugemann, and G. Civario, "There is no 16-clue sudoku: Solving the sudoku minimum number of clues problem via hitting set enumeration," *Experimental Mathematics*, vol. 23, no. 2, pp. 190–217, 2014.
[17] H.-H. Lin and I.-C. Wu, "Solving the minimum sudoku problem," in *The 2010 International Conference on Technologies and Applications of Artificial Intelligence*, ser. TAAI '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 456–461.
[18] H. Zhang, *Combinatorial Designs by SAT Solvers*, ser. Frontiers in Artificial Intelligence and Applications. IOS Press, February 2009, vol. 185, pp. 533–568.
[19] M. Black and G. Bard, "SAT over BOINC: an application-independent volunteer grid project," in *12th IEEE/ACM International Conference on Grid Computing, GRID 2011, Lyon, France, September 21-23, 2011*, S. Jha, N. gentschen Felde, R. Buyya, and G. Fedak, Eds. IEEE Computer Society, 2011, pp. 226–227.

Architectural Models for Deploying and Running Virtual Laboratories in the Cloud

E. Afgan[1,2], A. Lonie[3], J. Taylor[1], K. Skala[2], N. Goonasekera[3,*]
[1] Johns Hopkins University, Biology department, Baltimore, MD, USA
[2] Ruder Boskovic Institute (RBI), Centre for Informatics and Computing, Zagreb, Croatia
[3] University of Melbourne, Victorian Life Sciences Computation Initiative, Melbourne, Australia
enis.afgan@jhu.edu, alonie@unimelb.edu.au, jxtx@jhu.edu, skala@irb.hr, ngoonasekera@unimelb.edu.au

Abstract - Running virtual laboratories as software services in cloud computing environments requires numerous technical challenges to be addressed. Domain scientists using those virtual laboratories desire powerful, effective and simple-to-use systems. To meet those requirements, these systems are deployed as sophisticated services that require a high level of autonomy and resilience. In this paper we describe a number of deployment models based on technical solutions and experiences that enabled our users to deploy and use thousands of virtual laboratory instances.

I. INTRODUCTION

The past decade has seen cloud computing go from conception to the *de facto* standard platform for application deployment. Cloud infrastructures are delivering resources for deploying today's applications that are more scalable [1], more cost-effective [2], more robust [3], more easily managed [4], and more economical [5]. Researchers and research groups are no different from the rest of the industry, expecting robust, powerful cloud platforms capable of handling their data analysis needs. However, deploying such platforms still requires a significant amount of effort and technical expertise. In this paper, we build on our experiences from 5+ years of building and managing virtual laboratories that were deployed thousands of times on clouds around the world. We present viable architectural deployment models and extract best practices for others developing or deploying their own versions of robust research platforms.

The theme of this paper revolves around deploying the concept of a virtual laboratory as a platform for performing data analysis [6]. Virtual labs offer access to a gamut of data analysis tools and workflow platforms that are closely linked to commonly used datasets; they offer access to scalable infrastructure that has been appropriately configured, beforehand as well as dynamically at runtime. Once built, the virtual labs are often deployed on demand by the researchers themselves. However, in order to make these platforms available to domain researchers, there is a requirement to build, configure, and provision the necessary components. Depending on the complexity of a virtual lab, this is often a complex task spanning expertise in system administration, platform development, and domain-specific application setup.

In addition to deploying variations of said virtual labs on public clouds, institutions are increasingly setting up academic clouds (e.g., NeCTAR, Chameleon, JetStream). Locality of the infrastructure, restrictions on off-shoring data [7], avoiding vendor lock-in and the no-cost or merit-based allocation of resources are attractive reasons for utilizing those clouds. From a platform deployment standpoint, this brings up additional challenges because the platforms need to be deployed, managed, maintained, and supported on these additional clouds while coping with any differences among the cloud providers. It is hence imperative to design scalable, robust and cloud agnostic models for deploying these systems. Figure 1 captures the core concepts enabling development of such models: (a) a cross-cloud API layer; (b) automation; (c) a configurable and 'composable' set of resources. These concepts, detailed in the remainder of the paper, embody the notion that for successfully building a global virtual lab a common platform rooted in automation is needed.

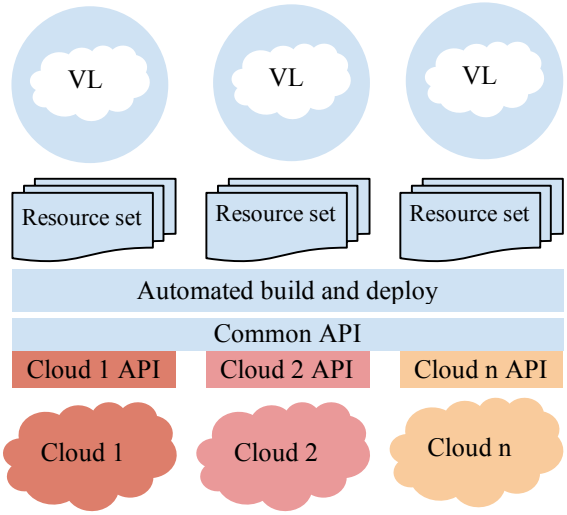


Figure 1. Virtual lab deployment stack unified for multiple clouds.

II. FUNCTIONAL REQUIREMENTS

The choice of a Virtual Lab architecture is driven by a variety of aspects and cross-cutting concerns [8]. While some of these decisions are general architectural decisions applicable to software in general, and some are highly specific to the domain in question, there are some concerns which are applicable to virtual labs in general. In this section, we provide a treatment of such concerns, and list some of the various architectural concerns that must be addressed in designing and developing a virtual lab environment.

For example, a virtual lab would need to determine the level of customisation that is required by a user. If a significant customisation is required, it is often the case that it will impact other users, and therefore, isolated or

* Corresponding author

individualised access to resources is preferable over access to a common pool of shared resources. For example, a user-owned container or virtual machine, as opposed to a pre-deployed web service.

Similarly, small job sizes can typically be catered to by a single, individualised VM, whereas large job sizes may require an architecture that can dynamically scale to accommodate more diverse needs.

The choice of appropriate strategy is dependent on several additional factors, including the purpose of the virtual lab, target cloud(s) capabilities, available people effort commitment and similar. Table 1 supplies a core list of design questions to answer when weighing the available options. Note that there is no single answer to the supplied questions but they are largely dependent on the aims of the virtual lab. Deciding what are the acceptable answers for the particular lab will help guide the myriad of technical choices related to implementation. In the sections that follow, we discuss various compute and data provisioning strategies that can accommodate these decisions.

TABLE I. FUNCTIONAL DESIGN QUESTIONS TO CONSIDER WHEN DESIGNING A VIRTUAL LAB.

Challenges	**Description**
Infrastructure	
Infrastructure maturity	Which cloud to use? How stable/mature is the infrastructure? Will the deployed lab be robust for users?
Infrastructure agnosticism	How easy is it to support multiple infrastructure providers? Is this desirable/necessary to increase accessibility/robustness?
Support	What type of support does the provider offer, for the virtual lab and individual users?
User Management	
Per-user customisation	Can each user customise the virtual lab according to their needs, and have a safe environment in which to learn through failure?
Data management	How is data put into/taken out of the virtual lab?
Quota management	What resource quotas should be enforced for the user? Does the infrastructure provider support that?
Users' Management of VL	
Instance lock-in	Is an upgrade path available so that the user can always use the latest version of the virtual lab?
Replicability	Can the user replicate their experiments - with a guarantee that all software versions remain unchanged?
Reliability	How reliable should the service be? Can losses be tolerated?
Service Management	
Software management	Can a user manage the software on the virtual lab on their own, or do they need system administration skills?
Licensing constraints	Are there specific licensing constraints that limit the use of the software?
Security	
Authentication	Should the virtual lab allow for single-signon with institutional credentials?
Credentials	How are institutional credentials translated into cloud provider credentials?
Authorization	What actions are users allowed to perform within a virtual lab?

III. VIRTUAL LAB INFRASTRUCTURE COMPONENTS

Answering above questions and provisioning a virtual lab requires a marriage of various, complex software components to their required storage and processing resource requirements. Depending on the intended usage for the virtual lab, there are a number of choices regarding the use of appropriate cloud resources. Table 2 provides a snapshot of the available approaches for supplying compute capacity along with the pros and cons for each option.

TABLE II. COMPUTE PROVISIONING STRATEGIES.

Provisioning strategy	Description	Pros	Cons
Machine Image	A pre-built machine image with all required software already installed.	• Quick startup • Excellent reproducibility	• Difficult to upgrade due to monolithic nature • Software packages not self-contained - causing potential version conflicts • Software potentially tied to OS version • Limitations on size • Breached applications may affect entire machine
Container	A pre-built container (such as Docker or LXC), which is deployed on top of a running Machine Image or a cloud container service.	• Extremely quick startup • Excellent reproducibility • Containers mostly independent of underlying machine's operating system and version • Easier updates to individual components • Breaches contained to container	• Must be pre-built • Still very new so quickly changing technology
Runtime	Required software installed at runtime, using automation software such as Ansible, Chef, Puppet, etc.	• Push or Pull update models • Updates easier to make	• Slow deployment/startup times • Less reproducible (transient network errors, software version changes)
Hybrid	A pre-built machine image/container for quick startup, brought up-to-date through runtime deployment/ extensibility.	• Can benefit from the advantage of all of the above models	• More complex to implement

Compute resources need to be matched with suitable storage capacity. Table 3 differentiates among the currently available cloud storage resource types and captures the pros and cons of each. The supplied information examines ways that data can be brought to the compute infrastructure, since this is still the dominant way to work with existing scientific software. We do not consider the reverse model as many virtual labs still struggle to shed the weight of their accumulated legacy software that require a shared, UNIX-based file-system to run, and are not yet in a position to take advantage of such models despite their benefits.

TABLE III. DATA PROVISIONING STRATEGIES FOR GETTING REQUIRED STATIC DATA TO THE COMPUTE.

Storage Model	Description	Pros	Cons
Volumes / Snapshots	A volume or snapshot containing the required data, which is attached to an instance at runtime.	• Quick to create/attach • Suitable for large amounts of data	• Not shareable between clouds (in OpenStack, not shareable between projects) • Not guaranteed to be available (e.g., infrastructure/quota restrictions) • Limited sharing ability between nodes (e.g., volumes only attachable to 1 instance at a time)
Shared POSIX filesystem	A shared filesystem containing the required data (e.g. NFS, Gluster), which is mounted on the target node at runtime.	• One-time setup • Very fast to attach • Updates visible at runtime to all virtual lab instances • Suitable for very large amounts of data	• Must be setup on each supported cloud • Centralised management/single point of failure • Not geographically scalable
Remotely fetched data archive	An http/ftp link containing the required data, which is downloaded and extracted onto local/transient storage.	• Cloud agnostic • Scalable	• Slow - takes a long time to fetch and extract • Not suited to very large amounts of data - to reduce download times/costs
Object-store	Object-storage service provided by the cloud provider (e.g. S3, Swift).	• High scalability	• Not suitable for random access • Not supported by legacy tools

IV. DEPLOYMENT OPTIONS

The resources available and required to compose a virtual lab can be assembled in a variety of configurations. Configurations support different use cases and require varying levels of technical complexity to deploy. Hence, depending on the intended purpose of the virtual lab, it is important to choose an appropriate deployment model. We define the following deployment models and supply a flowchart in Figure 2 to navigate among the models:

- **Centrally managed resource** is a virtual lab which is presented as a public service to the community. Typically available as a web portal, this virtual lab requires little or no setup from the user's side and permits the user to readily utilize resources offered by the virtual lab. Because it is a public resource, the user is likely to experience limited functionality, such as usage quotas, no *ssh* access, no possibility for customisation and other similar constraints typical of public services. While the users do not require any setup for this type of virtual lab, the lab maintainers need to manage and update underlying infrastructure supporting the supplied services. Resource management needs to account for the scaling of the supplied services, upgrades, and reproducibility of user's results. In addition to accessibility, other main drivers for choosing this model for lab deployment are data management restrictions (in case data is too large for feasible sharing) and software licensing constraints. Examples of this type of virtual lab include XSEDE science gateways (https://portal.xsede.org/web/guest/gateways-listing), Characterisation Virtual Lab (https://www.massive.org.au/cvl/), usegalaxy.org portal [9];
- **Standalone image** represents a feature-full version of the virtual lab in a small package. A user is required to have appropriate access to the cloud provider where the image is available and must personally launch an instance of the virtual lab; various launcher applications can make this a straightforward process. Once launched, the user will have full control over the virtual lab services but will also need to manage the services, particularly when upgrades or fixes are necessary. In addition to managing the services, the user is in charge of data management, ensuring that the data is not lost when an instance is terminated. Virtual lab providers need to bundle the virtual lab into an instantiatable image, such as a virtual machine image or a container, and provide periodic upgrades to the image. Examples include CloudBioLinux [10];
- **Persistent short-lived scalable cluster** is a dynamically scalable version of the virtual lab image with additional services to handle infrastructure scaling. These services (i.e., cluster management services) are used to provision a virtual cluster at runtime (e.g., Slurm, SGE, Hadoop) or utilize cloud provider services for scaling (e.g., container engine). The cluster manager software will also supply additional cluster management services, such as cluster persistence, allowing a user to shutdown the cluster when not in use while ensuring the data is preserved. This deployment model requires coordination of several resource types from Section 3 and use of cluster management software, hence implying a significant deployment effort from the virtual lab deployers. Examples include the Genomics Virtual Lab (GVL) [11];
- **Long-lived scalable cluster** has the same characteristics of a short-lived cluster as well as the ability to upgrade running services. The upgrades are typically handled by the cluster management software.

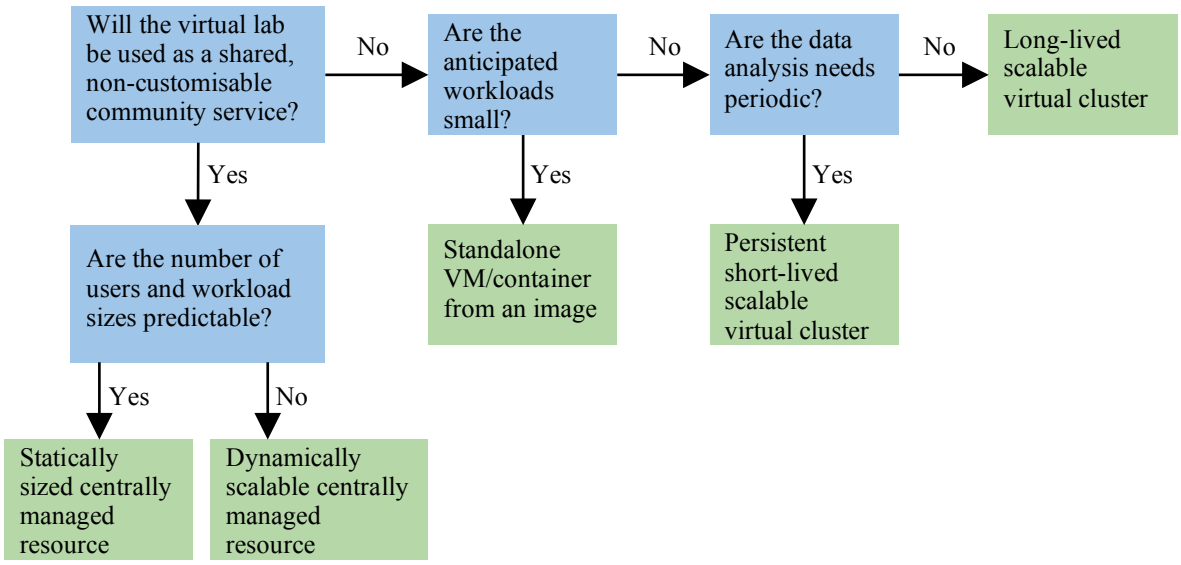


Figure 2. Data provisioning strategies for getting required static data to the compute.

V. DISCUSSION

In addition to the hardware and functional requirements for establishing a virtual lab, there are other important technical and management decisions that affect its deployment. One of the key attractions for using virtual labs is the high-level, software-as-a-service experience delivered to a user. An implication is that the functions offered by the deployed services need to function well,

which drives a need for good testing strategies and quality assurance. However, complex services and frequent releases make this a challenge for the deployers. It is hence advisable to automate the testing procedure and develop a quality control process before each release. Ideally, the testing process is decentralized, testing the individual services for proper functionality and focusing the testing of the virtual lab on the configuration setup. Such testing can be achieved by adopting a user-centric view of the virtual lab and using tools such as Selenium to automate the typical user actions [12].

Further, virtual labs should be complemented with a set of training materials, describing all the steps required to access the services supplied by a virtual lab. Additional training materials for using the services are also beneficial, particularly if they are accompanied with webinars or hand-on workshops.

Besides the technical implementation of the virtual lab, arguably the most challenging piece of a virtual lab is long-term support. Domain researchers will rely on the virtual lab to perform data analyses and publish new knowledge based on the obtained results. For reproducibility purposes, it is hence important to maintain their access to the resources required to use a virtual lab. When using a commercial cloud provider and supplying shared resources, note should be taken to questions of what happens when the project funding runs out or software being used becomes obsolete. Such questions imply that the virtual lab should make appropriate provisions to allow all the data and accompanying methods to be downloaded or transferred off the infrastructure initially used.

Related to the long term support is the notion of upgradeability. For example, a user working with a particular version of a virtual lab, may wish to upgrade to the latest available version. It is generally undesirable to foist an upgrade on users, as this can adversely affect reproducibility when software versions are changed. Therefore, a controlled migration or exit path is often necessary so that users can switch to newer versions of a virtual lab when appropriate for their circumstances.

VI. Summary

With the increased proliferation of cloud computing infrastructures, we believe the concept of a virtual lab - a composite platform capable of performing open-ended data analyses - will become a prevalent platform for utilizing cloud resources by researchers. In this paper we've described the components required to compose a virtual lab. Technical and managerial aspects of the decision making process have been presented shining light on the tradeoffs among viable options.

Looking to the future, it is expected that the concept of a virtual lab will continue to evolve towards a more integrated, quickly deployable system that is instantly accessible by users. Containers, automation solutions, and serverless runtime platforms are likely key technologies that will be adopted to realize this evolution.

Acknowledgments

This project was supported in part through grant VLS402 from National eCollaboration Tools and Resources, grant eRIC07 from Australian National Data Service, grant number HG006620 from the National Human Genome Research Institute, and grant number CA184826 from the National Cancer Institute, National Institutes of Health.

References

[1] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," *J. Internet Serv. Appl.*, vol. 1, no. 1, pp. 7–18, Apr. 2010.

[2] M. Armbrust, A. Fox, R. Griffith, A. Joseph, and RH, "Above the clouds: A Berkeley view of cloud computing," *Univ. California, Berkeley, Tech. Rep. UCB* , pp. 07–013, 2009.

[3] G. Garrison, S. Kim, and R. L. Wakefield, "Success factors for deploying cloud computing," *Commun. ACM*, vol. 55, no. 9, p. 62, Sep. 2012.

[4] G. Garrison, R. L. Wakefield, and S. Kim, "The effects of IT capabilities and delivery model on cloud computing success and firm performance for cloud supported processes and operations," *Int. J. Inf. Manage.*, vol. 35, no. 4, pp. 377–393, Aug. 2015.

[5] S. P. Ahuja, S. Mani, and J. Zambrano, "A Survey of the State of Cloud Computing in Healthcare," *Netw. Commun. Technol.*, vol. 1, no. 2, p. 12, Sep. 2012.

[6] S. D. Burd, X. Luo, and A. F. Seazzu, "Cloud-Based Virtual Computing Laboratories," in *2013 46th Hawaii International Conference on System Sciences*, 2013, pp. 5079–5088.

[7] J. J. M. Seddon and W. L. Currie, "Cloud computing and trans-border health data: Unpacking U.S. and EU healthcare regulation and compliance," *Heal. Policy Technol.*, vol. 2, no. 4, pp. 229–241, Dec. 2013.

[8] A. Garcia, T. Batista, A. Rashid, and C. Sant'Anna, "Driving and managing architectural decisions with aspects," *ACM SIGSOFT Softw. Eng. Notes*, vol. 31, no. 5, p. 6, Sep. 2006.

[9] E. Afgan, J. Goecks, D. Baker, N. Coraor, A. Nekrutenko, and J. Taylor, "Galaxy - a Gateway to Tools in e-Science," in *Guide to e-Science*, X. Yang, L. Wang, and W. Jie, Eds. Springer, 2011, pp. 145–177.

[10] K. Krampis, T. Booth, B. Chapman, B. Tiwari, M. Bicak, D. Field, and K. Nelson, "Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community," *BMC Bioinformatics*, vol. 13, p. 42, 2012.

[11] E. Afgan, C. Sloggett, N. Goonasekera, I. Makunin, D. Benson, M. Crowe, S. Gladman, Y. Kowsar, M. Pheasant, R. Horst, and A. Lonie, "Genomics Virtual Laboratory: A Practical Bioinformatics Workbench for the Cloud.," *PLoS One*, vol. 10, no. 10, p. e0140829, Jan. 2015.

[12] E. Afgan, D. Benson, and N. Goonasekera, "Test-driven Evaluation of Galaxy Scalability on the Cloud," in *Galaxy Community Conference*, 2014.

A CAD Service for Fusion Physics Codes

Marijo Telenta, Leon Kos and EUROFusion MST1 Team*
University of Ljubljana, Mech. Eng., LECAD, Ljubljana, Slovenia
marijo.telenta@lecad.fs.uni-lj.si, leon.kos@lecad.fs.uni-lj.si

***Abstract*—There is an increased need for coupling *machine descriptions* to various fusion physics codes. We present a computer aided design (CAD) service library that interfaces geometrical data requested by fusion physics codes in completely programmatic way for use in scientific workflow engines. Fusion codes can request CAD geometrical data at different levels of details (LOD) and control major assembly parameters. This service can be part of the scientific workflow that delivers meshing of the CAD model and/or variation of the parameters. In this paper we present re-engineering of the ITER tokamak using an open source CAD kernel providing standalone library of services. Modelling of the machine is done with several LOD, starting from the rough one and building/replacing with more detailed models by adding more details and features. Such CAD modelling of the machine with LODs delivers flexibility and data provenance records for the complete CAD to physics codes workflow chain.**

I. Introduction

Several commercial "workbench"-style *workflows* can integrate CAD data for further analysis and parametric optimisation. Usually, the CAD data for use in analysis is prepared or extracted and simplified in such a way that it is then easily processed further on by meshing tools. Control of the CAD data output relies on proprietary plug-ins that allow extraction to some extent for interaction foreseen by the "integrators". Depending on the *workbench*-suite, building a coupled simulation with external codes/solvers is only partially supported in general and does not offer required flexibility to "custom" physics codes used by the scientists. Importance of the "open" approach to data and results is often prerequisite for (re)use in scientific community and that rules out direct coupling to CAD packages within the "scientific" workflow engines such as Kepler [1].

Instead of "extracting" the *machine descriptions* from CAD models directly, most codes use manually prepared machine data for their input and have a "hard time" to verify its correctness with the CAD model. For each new machine to be introduced into the code, one needs to repeat the procedure of *machine description* creation by maintaining *"internal" input format* used by the code. Most of these *custom* input file formats are relatively simple and meant to be read in line with Fortran and other programming languages. Little attention is given to consistency checking of the input format. Most of the fusion-physics codes use "discretised" geometry for input by losing some "precision" in *machine descriptions*. Notable exceptions are *Monte Carlo* codes [2] which do not use *grids* for determining the particle position but rather simple primitives (spheres, bounding blocks, etc.) and to some extent *higher level* primitives bounded by NURBS surfaces commonly used in CAD modelling. Such codes are even faster if the space is not discretised with dense "grids" which can easily increase computational complexity of positional checking. In principle, *input grid density* controls the computational complexity of the code and a proper balance of the grid and available time is always sought by the scientists to achieve acceptable accuracy. It should be noted that *computational grid* usually differs from *input grid* although both should be in "consensus" when introducing important *geometrical details* that are neglected with "rough" grids. Described rationale means that for scientific modelling several *input grids* are needed that correspond to the "latest" CAD model available as the only "true" input source from which grids or other geometrical primitives should be generated for use as a code input. CAD modelling process of the tokamak machine in Fig. 1 was divided into five assemblies: blanket, divertor, magnets, cryostat, and vessel. For every assembly different

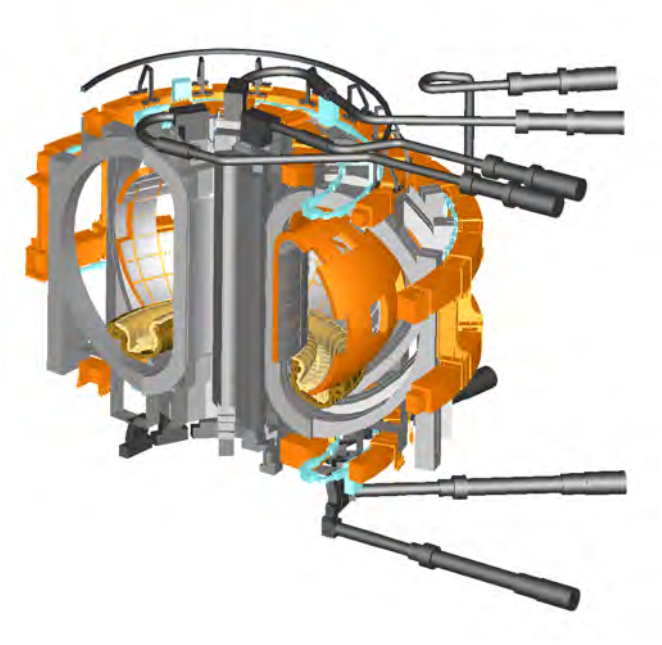


Fig. 1. ITER tokamak complex re-modelled with Open CASCADE kernel that provides diverse programmable geometric *services* from CAD data to fusion codes at several levels of details (LOD). Incremental LOD add features to assemblies while maintaining the physical correctness are one of the parameters of the *CAD service*.

* See http://www.euro-fusionscipub.org/mst1

level of details (LOD), leading geometrical dimensions, and other parameters such as number of sections are defined. The resulting CAD model is quite complex, with many parts, details and *features*. Physics codes need only specific parts of the CAD model prepared in a code-compatible format often called a "machine description". Including the LOD parameter that provides increased number of parts and geometrical *features* to analyses of physics codes accuracy and corresponding computational complexity is usually not performed due to tedious input preparation.

Manual reduction of features done by removing and simplifying unnecessary details in the CAD model is often called *defeaturing*. However, in this paper, opposite approach was taken. The CAD model was reverse engineered, i.e., the CAD model is re-modelled from bottom up, starting from simpler shape with basic features and modelling the geometry by adding features in parametric way. In this process different LOD's are defined where each LOD has pre-determined specific complexity of the model. The complexity of the model is defined by the number of features and repetitive pieces used to create the CAD model.

II. Level Of Details (LOD)

CAD model of the tokamak machine is reverse engineered using the Open CASCADE kernel. Open CASCADE [3] is an open source CAD kernel that offers programmatic flexibility in modelling and meshing of CAD models which mesh can be used as an input for fusion codes. The different levels of details are created in the process of programming the geometry with different CAD model complexity respectively. Basic shapes are modelled with little of geometrical details and with different number of segments as shown without the cryostat assembly in Fig. 2. The first LOD, LOD-0, includes the basic shape and features of the model. This LOD could be used for analysis where details are not as important and make the grid unnecessarily more complex.

Several LODs for different sections are shown with different number of segments. This means that the physics codes could be supplied with varying LODs and segments in a programmatic way. Furthermore, the shapes may differ between LODs to follow physics requirements for the simulation. For example, in Fig. 2(a) the divertor is modelled in a "fused" way resembling basic shapes without any void space inside; Blanket and vessel in Fig. 2(b, c) are having no ports apertures. If holes are important for simulation then one can increase the LOD for that part only. If there is a specific requirement on the shape or number of holes one can "inherit" the code and add or remove features. Numeration of LOD doesn't need to follow agreed levels of detail among assemblies if they are properly developed and documented. Users can consider the code that generates the CAD models as a starting point for specific modifications that can later on become a part of the CAD service. When shape of some part needs modifications due to precision of obeying the code requirements on the complexity, the same principles of LOD modifications can be followed.

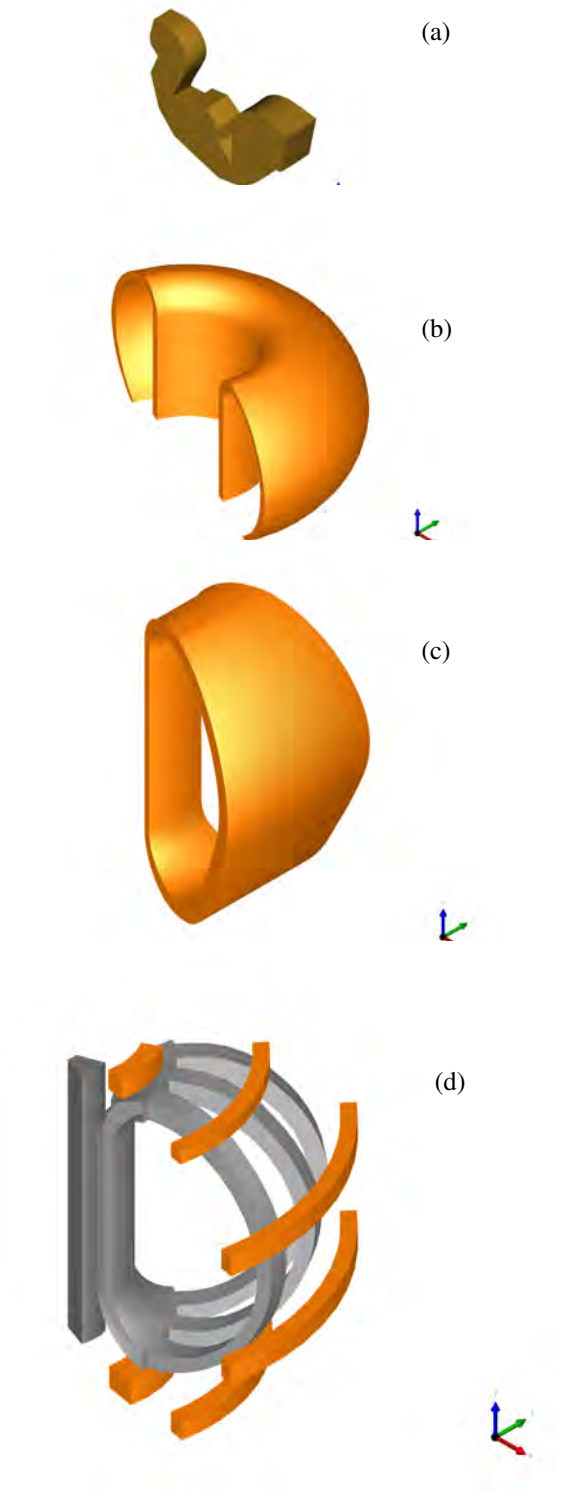


Fig. 2. LOD-0 of the divertor segment (a), half revolution of the blanket (b), quarter revolution of vessel (c), and magnets with 3 segments. The geometry for all parts uses the basic shape without the features that are added or being replaced at a higher LOD.

Drawback of the programmatic approach is clearly non-graphical representation of the models. For easier modifications and comprehension we are using Python "wrapping" of the PythonOCC [4] CAD kernel. The Python code is interpreted and skips the need of recompilation at source code modifications. CAD kernel performance by using Python is not significantly affected, as all kernel operations are written in C++. Repeating lengthy process of model creation with increasing LOD is even more comfortable with programming in Python than within a CAD package done interactively.

Listing 1. Snippet where separate sections of the tokamak are called in programmatic way in Python. In addition, CAD service for cross section is called for divertor.

```
divertor(display, level_of_detail, number_of_bodies,
        divertor_outer_radius, divertor_base_radius,
        divertor_outer_dimension, divertor_inner_dimension)

magnets(display, level_of_detail,
        Magnet_Outer_Radius, Magnet_Inner_Radius,
        Magnet_System_Height, Number_Of_Elements)

vessel(display, level_of_detail, NumberOfSlices,
        VacuumVesselInnerRadious, VacuumVesselOuterRadious,
        VacuumVesselUpperHeight, VacuumVesselBottomHeight)

blanket(display, level_of_detail, revolve_for,
        height_above_ks, height_below_ks, inner_radius,
        outer_radius, top_outer_radius, top_inner_radius,
        shell_width)

cryostat(display, level_of_detail,
         cryostat_base_inner_radius, sections)

Section(display, shape, angle)
```

As one can see from the Python example in Listing 1, each assembly that outputs to *display* has several input parameters, such as LOD, number of segments, and parameters which define the geometry. In this way, desired level of details is set, and the CAD model geometry can be altered by changing the input parameters, for example, vessel's inner and outer radius can be changed and all other dimensions of the vessel accordingly. Instead of outputting to *display*, one can choose to output shapes/assemblies and combine them into desired assembly for further operations on geometry. Operations, such as *Section()* in the last line of Listing 1 are part of geometry services provided by the library.

Creation of the each assembly with PyhonOCC undergoes several steps in programming CAD kernel. Python code for each assembly is also self-consistent. It acts as a module that can be imported independently into other Python codes. Modules have additional geometric parameters such as dimensions and positions of the features that are not exposed but available for modifications by the users.

The second level of details in Fig. 3 adds details and new shapes that were left out at LOD-0. Decision on what is included at each level is guided by the physics code requirements. As one can see, for the second LOD-1 apertures are modelled in the blanket and the vessel. The finest LOD-2 shown in Fig. 4 defines the most detailed CAD geometry where further geometric details are added to the CAD model in reference to the LOD-1, such as fillets, fasteners, etc.

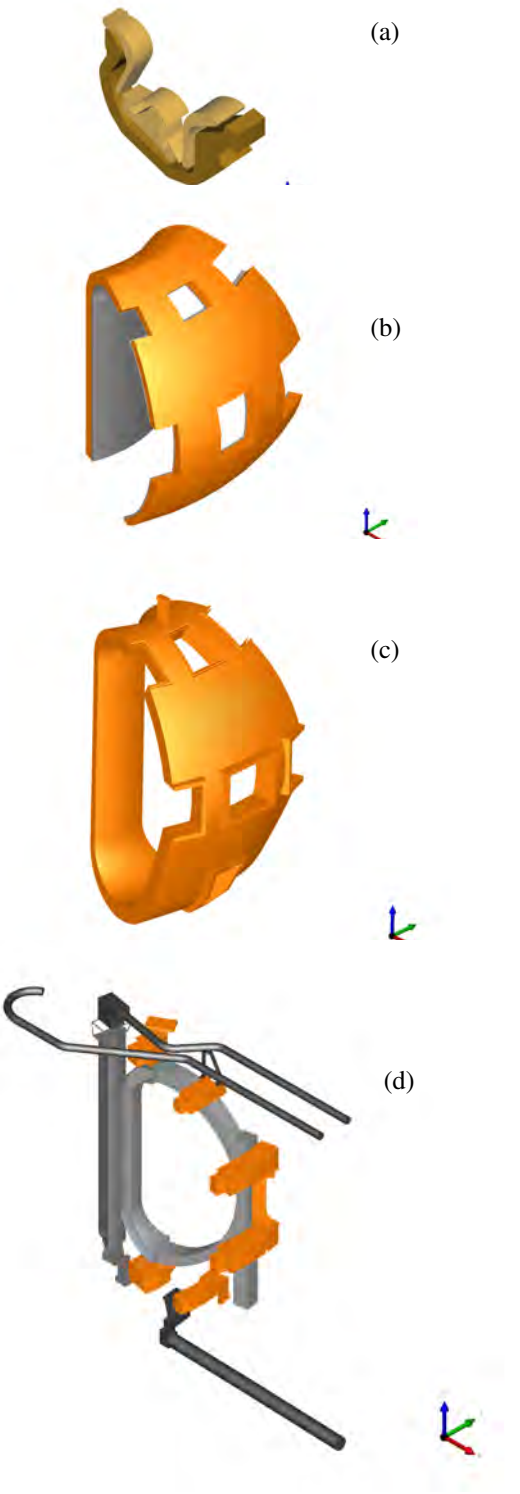


Fig. 3. LOD-1 of the divertor segment (a), quarter revolution of the blanket (b), quarter revolution of the vessel (c), and one segment of the magnets (d). Divertor is now split into two bodies with better description of the shapes. Apertures for diagnostic ports, antennas and piping are added to the blanket and the vessel. Toroidal magnets are more precisely described and cooling piping is added.

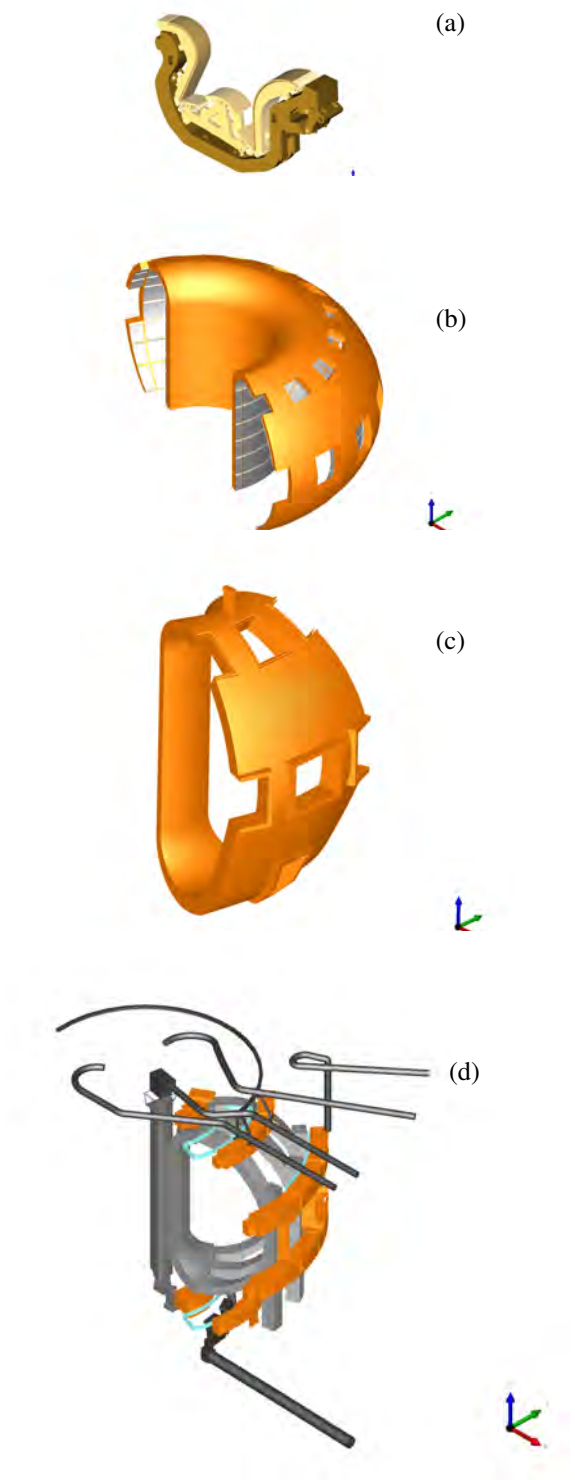


Fig. 4. LOD-2 of the divertor segment (a), half revolution of the blanket (b), quarter revolution of the vessel (c), and 3 segments of the magnets (d). Blankets that were previously modelled as a single shape with apertures are now separated and have gaps in-between. Fillets are added to vessel. Divertor was further detailed with holes. Magnets received details in piping and edge features. With the finest LOD-2 one can study influence of the tiny details to the simulations. The choice of what is included is still left to the user with the modification of the code.

III. A CAD SERVICE

The aim of the presented CAD service is to deliver a common interface for transfer of CAD data into *machine descriptions* for use by physics codes of several facilities (ITER, AUG, JET, MAST-U, TCV and W7-X). The ultimate goal is to control, in programmatic way, the disfeaturing process of the CAD model to the LOD. The product of the disfeaturing process could be than used as an input for different meshing codes and eventually meshed correctly.

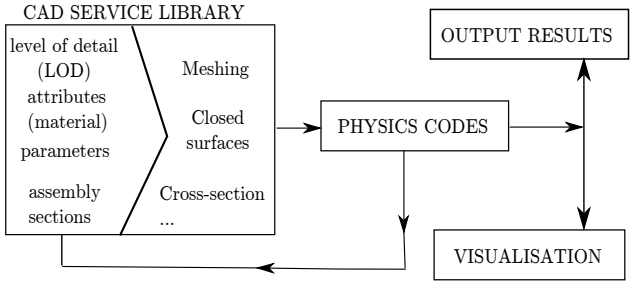


Fig. 5. CAD service library coupling with physics codes that can control generated input.

A CAD service, when used within a scientific workflow, as shown in Fig. 5, delivers an output in a format suitable for reading by general meshing tools. Meshing operation is treated as a black box that outputs the grid which is then used an input for physics codes. The output mesh, as well as, the results of the physics codes, which are calculated using these grids are stored together in a scientific database format. The mesh format from the meshing tools is favoured to be stored in a common fusion modelling grid description. Ideally, it should be possible all results to be stored in general grid description (GGD) format, as this enhances further processing and analysis of the results using visualisation tools that are available within the EUROFusion Integrated Modelling community. IGES and STEP standards are commonly used for the CAD data exchange and can serve as a common data format for input to meshing codes. Otherwise, if a "mesher" is unable to read the standard CAD format then this usually means that custom input needs to be prepared by the CAD service. For example, physics code PFCFLUX [5] calculates the heat flux on the plasma facing components (PFC). Plasma-facing components are a generic term for any part of the tokamak machine where large amount of power deposition is found. The PFCFLUX needs as an input triangular mesh of the 3D PFC surfaces. The usual way in producing required mesh is to manually separate the PFC surfaces from the CAD geometry, fill any holes and gaps not needed on those surfaces, assign different materials to different surfaces, and heal any surfaces if needed. The CAD service library can supply the PFCFLUX code with needed input data in programmatic way which could save time and effort in preparing the mesh. Since PFCFLUX needs only surface triangles, Open CASCADE kernel can be used to build the triangular mesh [6]. Similar approach was

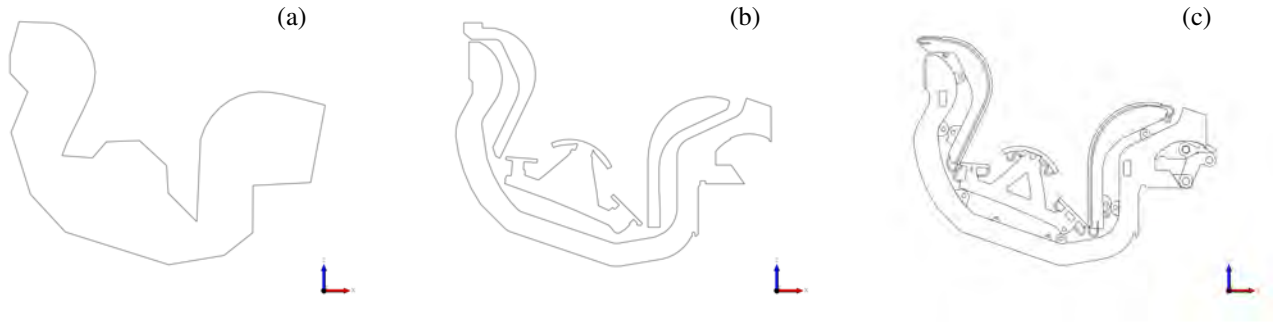


Fig. 6. CAD service for cross-section of divertor with (a) at LOD-0, (b) at LOD-1, and (c) at LOD-2.

taken in SMARDDA software library which offers possibility to perform design-relevant calculations for PFCs [7].

The CAD service is missioned as a library or a module which would be then installed and used as a service for specific physics code. Different CAD service library/module would be created for different tokamak machines. It can provide closed surfaces, needed for analysis, for example, in plasma facing components. In addition, the service could offer, as mentioned above, different LODs which would satisfy the specific needs of the physics code considered. It can give specific attributes of the model such as *material*. Also, basic mesh can be provided or different cross sections of the model. Finally, depending on the need, different assembly sections and parameters can be provided to the mesher and/or physics codes. If provided data to the physics codes are not sufficient or not valid, CAD service library can re-provide requested data in appropriate form. One can see from Fig. 6 the product of the CAD service for cross section of the divertor with different levels of details. This service is used by the physics codes which use 2D geometry for calculations such as SOLPS-ITER [8].

IV. Conclusion

Programming CAD kernel Open CASCADE is used to model CAD geometry of a tokamak machine from simple shape to more complex one. In the process of creating the geometry, different levels of details are defined and programmed. Product is a CAD service library/module which can by request serve physics codes with appropriate data. Variety of the services could be offered depending on the physics code used, such as closed surfaces, cross sections, basic mesh, and material. Numerous physics codes could potentially have benefit using this CAD service. The time spent on preparing the CAD model for meshing will be significantly shorter when using this library. Also, provenance is recorded such that others could repeat the procedure to produce the particular mesh in question for the physics code. Reverse engineering approach for CAD geometry modelling presents opposite approach of CAD model disfeaturing. "Manual" preparation of the CAD models is tedious task that can take several months. With this service many of the preparation steps for simplification are avoided and can be fine-grained and uniformly controlled. Usual simplification steps include geometric simplification, suppression/hide of irrelevant details, and decomposition in LODs of complex parts. Future work will include additional tokamaks, services and actors for Kepler.

Acknowledgement

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 under grant agreement No. 633053, task agreement AWP15-EEG-JSI/Telenta. The views and opinions expressed herein do not necessarily reflect those of the European Commission. The authors would like to acknowledge R. Pitts, S. Pinches and X. Bonnin from ITER for helpful discussions. Thanks goes to many students from the Faculty of Mechanical Engineering at the University of Ljubljana that have contributed to this work.

References

[1] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock, "Kepler: An extensible system for design and execution of scientific workflows," in *SSDBM '04: Proceedings of the 16th International Conference on Scientific and Statistical Database Management*. Washington, DC, USA: IEEE Computer Society, 2004, p. 423.

[2] L. Lu, U. Fischer, Y. Qiu, and P. Pereslavtsev, "The CAD to MC geometry conversion tool McCad: Recent advancements and applications," in *ANS MC2015 - Joint International Conference on Mathematics and Computation (M&C), Supercomputing in Nuclear Applications (SNA) and the Monte Carlo (MC) Method*, 2015.

[3] "OpenCASCADE – Open CASCADE technology," http://opencascade.org, 2016.

[4] "PythonOCC website – Open CASCADE for Python," http://pythonocc.org, 2016.

[5] M. Firdaouss, V. Riccardo, V. Martin, G. Arnoux, C. Reux, and J.-E. Contributors, "Modelling of power deposition on the jet iter like wall using the code pfcflux," *Journal of Nuclear Materials*, vol. 438, Supplement, pp. S536 – S539, 2013, proceedings of the 20th International Conference on Plasma-Surface Interactions in Controlled Fusion Devices. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022311513001190

[6] K. L. Telenta, M, R. Akers, and E.-I. Team, "Interfacing of CAD models to a common fusion modelling grid description," in *Proceedings of the International Conference Nuclear Energy for New Europe*, Portorož, Slovenia, 2015, pp. 707.1 – 707.8.

[7] W. Arter, V. Riccardo, and G. Fishpool, "A cad-based tool for calculating power deposition on tokamak plasma-facing components," *Plasma Science, IEEE Transactions on*, vol. 42, no. 7, pp. 1932 – 1942, sept. 2014.

[8] S. Wiesen, D. Reiter, V. Kotov, M. Baelmans, W. Dekeyser, A. Kukushkin, S. Lisgo, R. Pitts, V. Rozhansky, G. Saibene, I. Veselova, and S. Voskoboynikov, "The new solps-iter code package," *Journal of Nuclear Materials*, vol. 463, pp. 480 – 484, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022311514006965

Correlation between attenuation of 20 GHz satellite communication link and Liquid Water Content in the atmosphere

Maks Kolman* and Gregor Kosec**
*student at Jožef Stefan Insitute, Department of Communication Systems, Ljubljana, Slovenia
**Jožef Stefan Insitute, Department of Communication Systems, Ljubljana, Slovenia
maks.kolman@student.fmf.uni-lj.si, gregor.kosec@ijs.si

Abstract – The effect of Liquid Water Content, i.e. the mass of the water per volume unit of the atmosphere, on the attenuation of a 20 GHz communication link between a ground antenna and communication satellite is tackled in this paper. The wavelength of 20 GHz electromagnetic radiation is comparable to the droplet size, consequently the scattering plays an important role in the attenuation. To better understand this phenomenon a correlation between measured LWC and attenuation is analysed. The LWC is usually estimated from the pluviograph rain rate measurements that captures only spatially localized and ground level information about the LWC. In this paper the LWC is extracted also from the reflectivity measurements provided by a 5.6 GHz weather radar situated in Lisca, Slovenia. The radar measures reflectivity in 3D and therefore a precise spatial dependency of LWC along the communication link is considered. The attenuation is measured with an in-house receiver Ljubljana Station SatProSi 1 that communicates with a geostationary communication satellite ASTRA 3B on the 20 GHz band.

I. Introduction

The increasing demands for higher communication capabilities between terrestrial and/or earth-satellite repeaters requires employment of frequency bands above 10 GHz [1]. Moving to such frequencies the wavelength of electromagnetic radiation (EMR) becomes comparable to the size of water droplets in the atmosphere. Consequently, EMR attenuation due to the scattering on the droplets becomes significant and ultimately dominant factor in the communications quality [2]. During its propagation, the EMR waves encounter different water structures, where it can be absorbed or scattered, causing attenuation [1]. In general, water in all three states is present in the atmosphere, i.e. liquid in form of rain, clouds and fog, solid in form of snow and ice crystals, and water vapour, which makes the air humid [3]. Regardless the state, it causes considerable attenuation that has to be considered in designing of the communication strategy [2]. Therefore, in order to effectively introduce high frequency communications into the operative regimes, an adequate knowledge about atmospheric effects on the attenuation has to be elaborated.

In this paper we deal with the attenuation due to the scattering of EMR on a myriad of droplets in the atmosphere that is characterised by LWC and drop size distribution (DSD). A discussion on the physical background of the DSD can be found in [4], where authors describe basic mechanisms behind distribution of droplets. Despite the efforts to understand the complex interplay between droplets, ultimately the empirical relations [5] are used. The LWC and DSD can be related to the only involved quantity that we can reliable measure, the rain rate [6]. Recently it has been demonstrated that for high rain rates also the site location plays a role in the DSD due to the local climate conditions [7].

In general, raindrops can be considered as dielectric blobs of water that polarize in the presence of an electric field. When introduced to an oscillating electric field, such as electromagnetic waves, a droplet of water acts as an antenna and re-radiates the received energy in arbitrary direction causing a net loss of energy flux towards the receiver. Some part of energy can also be absorbed by the raindrop, which results in heating. Absorption is the main cause of energy loss when dealing with raindrops large compared to the wavelength, whereas scattering is predominant with raindrops smaller than the wavelength [2]. The very first model for atmospheric scattering was introduced by lord Rayleigh [8]. The Rayleigh assumed the constant spatial polarization within the droplet. Such simplifications limits the validity of the model to only relatively small droplets in comparison to the wavelength of the Incident field, i.e. approximately up to 5 GHz when EMR scattering on the rain droplets is considered. A more general model was developed by Mie in 1908 [2], where a spatial dependent polarization is considered within the droplet, extending the validity of the model to higher droplet size/EMR wavelength ratios. Later, a popular empirical model was presented in [9], where attenuation is related only to the rain rate. The model, also referred in as a Marshal-Palmer model, is widely used in evaluation of LWC from reflectivity measured by weather radars [10]. Marhsall-Palmer model simply states the relation between the attenuation and rain rate in terms of a power function. In this paper we seek for correlation between the LWC and attenuation measurements. LWC is extracted from reflectivity measurements provided by a weather radar situated in Lisca and operated by Slovenian Environment Agency [11]. Attenuation is measured by in-house hardware that monitors the signal strength between Ljubljana Station SatProSi 1 and communication satellite ASTRA 3B [12, 13, 14, 15]. The main purpose of this paper is therefore to investigate correlation between precipitation measured in 3D with the meteorological radar and the measured attenuation.

II. Governing models

Before we proceed to measurements some basic relations are discussed.

Attenuation (A) is a quantity measured in [dB] that describes the loss of electromagnetic radiation propagating through a medium. It is defined with starting intensity I_s and the intensity received after propagation I_r as

$$A = 10 \log_{10} \frac{I_s}{I_r}. \tag{1}$$

The specific attenuation ($\alpha = A/L$) measured in [dB/km] as a function of rain rate (R) measured in [mm/h] is commonly modelled as [5]

$$\alpha(R) \sim a\, R^b\ . \tag{2}$$

Coefficients a and b are determined empirically by fitting the model to the experimental data. In general, coefficients depend on the incident wave frequency and polarization, and ambient temperature. Some example values for different frequencies are presented in Table 1.

Drop size distribution (DSD) is a quantity that, unsurprisingly, describes the distribution of droplet sizes. The simplest characterization of rain is through rain rate R, measured in [mm/h]. However, rain rate do not give any information about the type of rain. For example a storm and a shower might have the same rain rate, but different drop size distribution. However, a simple DSD model is presented in [9]

$$N(D) = U \exp(-V\, R^{\delta} D), \tag{3}$$

where D stands for drop diameter measured in [mm], $N(D)$ describes number of droplets of size D to $D + \mathrm{d}D$ in a unit of volume measured in [$\mathrm{mm}^{-1}\mathrm{m}^{-3}$] and R rain rate measured in [mm/h]. The values of equation parameters were set to $U = 8.3 \cdot 10^3$, $V = 4.1$ and $\delta = -0.21$. The DSD was also determined experimentally for different rain rates [5]. The experimental data is presented in Figure 1, where we can see that the typical diameter of drops is in range of mm. There is a discrepancy between the theoretical and experimental data with very small droplets. This can be fixed with a modified DSD, however scattering on such small droplets is negligible so the difference is not relevant.

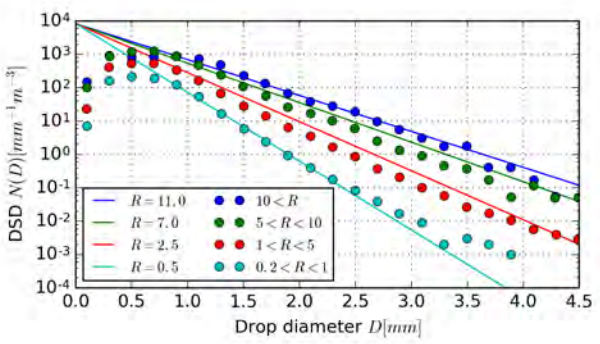


Figure 1. DSD measured in Czech Republic (one year measurement, rain rate R is the parameter of particular sets of points) [5]. Lines represent the theoretical value as determined by (3).

III. Measurements

A. Measurements of signal attenuation

Jožef Stefan Institute (JSI) and European Space Agency (ESA) cooperate in SatProSi-Alpha project that includes measuring attenuation of the communication link between on ground antennas and a satellite, more precisely between the ASTRA 3B satellite and SatProSi 1 station. The ASTRA 3B is a geostationary communication satellite located on the $23.5°E$ longitude over the equator. It broadcasts the signal at 20 GHz, which is received at SatProSi 1 with an in-house receiver, namely 1.2 m parabolic antenna positioned on the top of the JSI main building with a gain of about 47 dB. The SatProSi measures attenuation every 0.15 seconds, resulting in over 500000 daily records, since 1. 10. 2011.

B. Measurements of rainfall rate

Two sources of rain measurements are used in this paper. The first one is a pluviograph installed locally in the proximity of the antenna. The rain rate is measured every five minutes.

Another, much more sophisticated, measurements of rain characteristics are provided by meteorological radars. The basic idea behind such radars is to measure EMR that reflects from water droplets. The measured reflectivity is then related with rain rate with Marhsall-Palmer relation. Radar reflectivity factor Z is formally defined as the sum of sixth powers of drop diameters over all droplets per unit of volume, which can be converted into an integral

$$Z = \int_0^{\infty} N(D) D^6 \mathrm{d}D\ . \tag{4}$$

Note that the form of relation follows the Rayleigh scattering model [16]. Z is usually measured in units $\mathrm{mm}^6\mathrm{m}^{-3}$. When conducting measurements a so-called Equivalent Reflectivity Factor

$$Z_e = \frac{\eta \lambda^4}{0.93 \pi^5} \tag{5}$$

is used, where η means reflectivity, λ is radar wavelength and 0.93 stands for dielectric factor of water. As the name suggests both are equivalent for large wavelengths compared to the drop sizes, as in Rayleigh model [16].

Reflectivity factor and rainfall rate are related through Marshall-Palmer relation as

$$Z_{[\mathrm{mm}^6\mathrm{m}^{-3}]} = \tilde{a} R^{\tilde{b}}_{[\mathrm{mm/h}]}\ , \tag{6}$$

where $Z_{[\mathrm{mm}^6\mathrm{m}^{-3}]}$ is reflectivity factor measured in $\mathrm{mm}^6\mathrm{m}^{-3}$ and $R_{[\mathrm{mm/h}]}$ is rainfall rate measured in mm/h. In general, empirical coefficients $\tilde{a}$ and $\tilde{b}$ vary with location and/or season, however, are independent of rainfall R. Most widely used values are $\tilde{a} = 200$ and $\tilde{b} = 1.6$ [9, 10]. Meteorologists rather use dimensionless logarithmic scale and define

$$dBZ = 10\ \log_{10} \frac{Z}{Z_0} = 10\ \log_{10} Z_{[\mathrm{mm}^6\mathrm{m}^{-3}]}\ , \tag{7}$$

where Z_0 is reflectivity factor equivalent to one droplet of diameter 1 mm per cubic meter.

f[GHz]	10	12	15	20	25	30
a	0.0094	0.0177	0.0350	0.0722	0.1191	0.1789
b	1.273	1.211	1.143	1.083	1.044	1.007

TABLE 1. VALUE OF COEFFICIENTS FOR MARSHAL-PALMER RELATION $\alpha(R)$ AT DIFFERENT FREQUENCIES. [5]

The meteorological radars at Lisca emit short ($1\,\mu s$) electromagnetic pulses with the frequency of $5.62\,GHz$ and measure strength of the reflection from different points in their path. Radar collect roughly 650000 spatial data points per radar per every atmosphere scan, which they do every 10 minutes. They determine the exact location of all their measurements through their direction and the time it takes for the signal to reflect back to the radar.

In addition to reflectivity radars also measure the radial velocity of the reflecting particles by measuring the Doppler shift of the received EMR, but this is a feature we will not be using.

IV. DATA ANALYSIS

The analysis begins with handling approximately 20 GB of radar data for the academic year 2014/15 accompanied with 3 GB of signal attenuation data for the same time period and approximately 5 GB of attenuation and local rain gauge data for years 2012 and 2013.

A. Preprocessing the radar spatial data

Radar data was firstly reduced by eliminating spatial points far away from our point of interest, namely the JSI main building where antenna is located. The geostationary orbit is $35\,786\,km$ above the sea level, therefore the link between the antenna and the satellite has a steep elevation angle $36.3°$. In fact just $20\,km$ south of the antenna the ray rises above $15\,km$, which is the the upper boundary for all weather activities [3]. Knowing this, a smaller area of the map can be safely cropped out, reducing the number of data points from around $650\,000$ to approximately 6500 for each radar scan covering an $40\,km \times 40\,km$ area.

Although we already gravely reduced original data size, we must still reduce thousands of points into something tangible. The positions of both the antenna and the satellite are known at all times, a lovely consequence of them being stationary, therefore the link between them can be easily traced. Roughly 150 points on the ray path are used as a discrete representation of the link, referred to as link points in future discussions. For each link point a median of n closest radar measurements is computed as a representative value. The other way of extracting reflectivity factor was simply to take closest n points to the antenna and select the median value of those. A visualisation of both methods is presented in Figure 2.

Now we are left with multiple scalar quantities as a function of time. Antenna attenuation for every $0.15\,s$, local rain gauge for every $5\,min$ and various extractions of reflectivity factor for every $10\,min$. Note, that radar values are not averaged over 10 minutes, radar simply needs 10 minutes to complete a single scan. In Figure 3 an example of rainfall rate measured with weather radar and the measured attenuation for a three day period is presented. A correlation between quantities is clearly seen on the figure but a closer inspection is needed to reveal more details about the correlation.

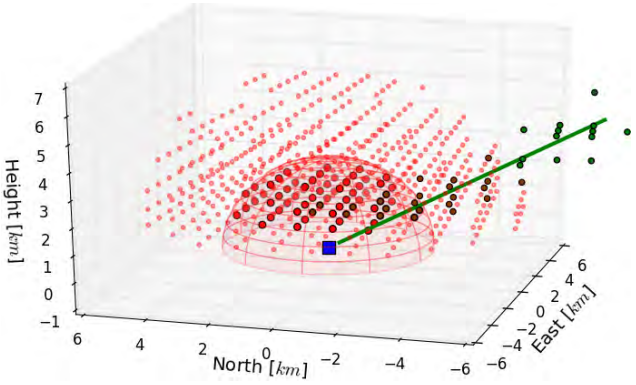


Figure 2. Positions of radar measurements. The blue rectangle is the location of the antenna and the rain gauge. The 64 points closest to the antenna are enclosed in a red sphere and marked as red circles. Red dots mark the remainder of 512 closest points. The green line is the ray path between antenna and satellite with green circles representing corresponding support nodes for support size $n = 4$.

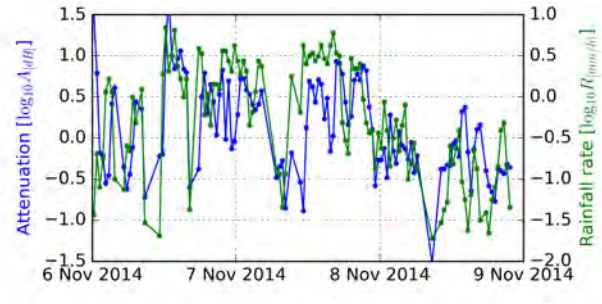


Figure 3. Measured antenna attenuation and rain rate extracted from 64 radar measurements closest to the antenna. Both datasets have been sorted into 30 minute bins.

B. Correlation between rain and attenuation

In order to find a relation between rain rate and electromagnetic attenuation, measurements of both quantities must be paired. There is no obvious way of doing this since both are measured at a vastly different time-scale. We ended up dividing time into bins of duration t_0 and pairing the measurements that fall within the same bin. The maximum values of every quantity was selected as a representative for the given time period.

The correlation coefficient between two variables X and Y can be calculated using

$$corr(X,Y) = \frac{\text{mean}((X - \text{mean}(X)) \cdot (Y - \text{mean}(Y)))}{\text{std}(X)\text{std}(Y)} \tag{8}$$

and is a good quantity for determining linear dependence between X and Y.

According to the Marshall-Palmer power law a linear relation exists between logarithms of rain rate and specific attenuation. Our measurements are of total attenuation A and not of specific attenuation so we must adjust the equation. We assume a typical distance L as a connecting factor between the, which gives us

$$\log_{10} A = \log_{10} La + b \log_{10} R. \tag{9}$$

The exact value of L is not relevant as only the parameter b will interest us. Therefore a slope on a log-log graph, such as on Figure 4, is equal to the model parameter b. We used a least square linear fit on each set of data to get the corresponding values for b.

In addition, correlation between logarithmic values of rain rate and attenuation

$$corr\left(\log_{10} A_{[\mathrm{dB}]}, \log_{10} R_{[\mathrm{mm/h}]}\right) \tag{10}$$

is used as a quality measure of their relation.

V. Results and Discussion

Once we paired attenuation and rainfall data we can scatter the points on a graph. In Figure 4 the attenuation against rain rate at 8 h bin size is presented. For the local radar representation $n = 2^6$ and for integral representation $n = 2^2$ support size is used. The correlation can be clearly seen, however not unified, as one would expect if measurements and rain rate - reflectivity model would be perfect. Since we introduced two free parameters, namely time bin t_0 and spatial support size n for integral and n for local radar representation, a sensitivity analyses regarding those parameters is needed.

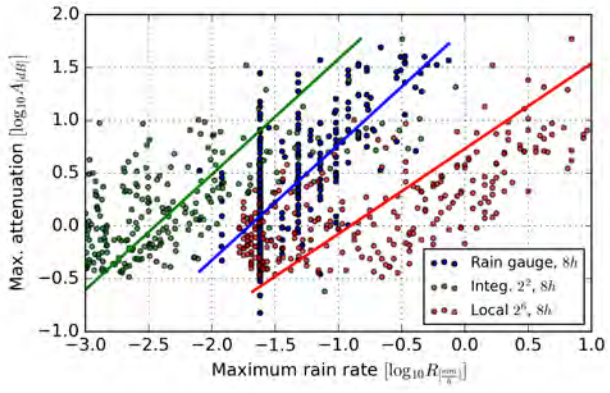


Figure 4. Attenuation dependency on the rain rate measured in three different ways. Local rain gauge (blue), path integration on each step selecting closest 4 points (green) and from 64 points closest to the antenna (red). All measurements have been put into $8h$ bins.

In Figure 5 a correlation with respect to the number of local support nodes and time bin size is presented. The best correlation is obtained with $8h$ time bins and a local $n = 2^6$ support size.

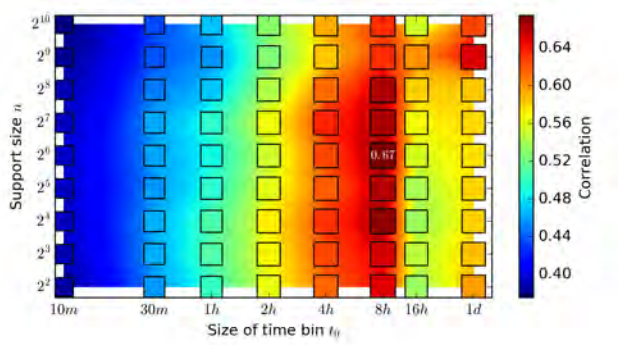


Figure 5. Correlation between rain rate and attenuation with respect to the number of **local** support size n and time bin size t_0.

Similarly a correlation with respect to the number of integral support nodes and time bin size is presented in Figure 6 . Again, the best correlation is obtained with $8h$ time bins, however with the integral model a small integral support, i.e. $n = 2^2$, already suffices to obtain fair correlation. Such a behaviour is expected. In an integral mode we follow the ray and support is moving along, therefore there is no need to capture vast regions for each link point. On the other hand in a local approach only one support is used and therefore that support has to be much bigger to capture enough details about the rain conditions.

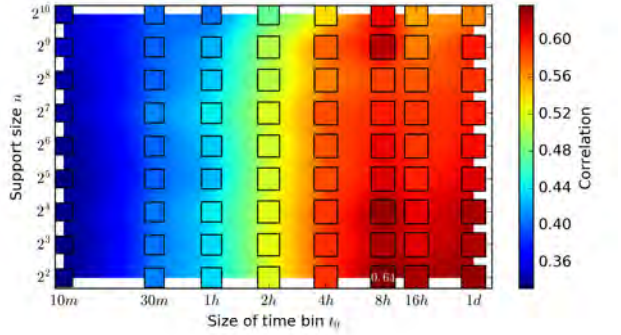


Figure 6. Correlation between rain rate and attenuation with respect to the number of **integral** support size n and time bin size t_0.

To compare measurements acquired with radar and the ones acquired with the local rain gauge a simpler presentation of correlation is shown on Figure 7. One set of data has rain rate extracted from radar using the integral method with support size 4 and two sets using closest either 64 or 512 nodes.

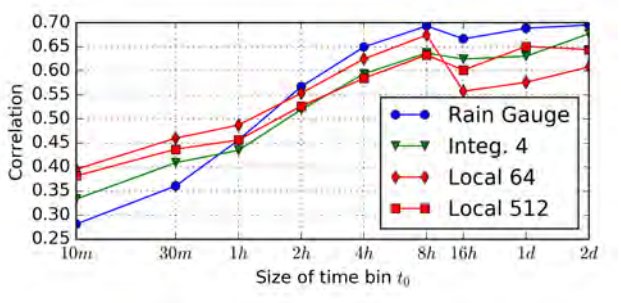


Figure 7. Correlation between rain rate and attenuation as a function of time bin size t_0 for different ways of extracting the rain rate.

In the next step we compare our measurements with a Marshall-Palmer model, specifically the exponent b. According to [9] in 20 GHz the $b_0 = 1.083$ should hold. In Figure 8 differences between our measurements and b_0 with respect to the time bin size are presented for the same sets of data as were used in the correlation analysis of Figure 7. An order of magnitude improvement is visible between local rain gauge and data extracted with radar.

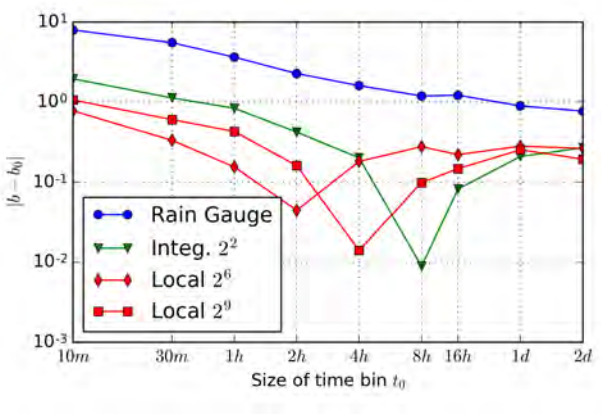


Figure 8. Exponent in attenuation to rainfall relation b compared to value b_0 from Table 1 for 20 GHz as a function of bin duration t_0 for a few ways of extracting rainfall.

VI. Conclusion

This paper deals with the correlation between attenuation of EMR signal due to the scattering on the ASTRA 3B - SatProSi 1 link and measured rain rate. The main objective of the paper is to analyze the related measurements and comparison of results with Marshall-Palmer model. The attenuation is measured directly with an in-house equipment with a relatively high time resolution (0.15 s). The rain characteristics are measured with a rain gauge positioned next to the antenna and a national meteorological radar. The rain gauge measures average rain rate every five minutes at a single position, while the radar provides a full 3D scan of reflectivity every 10 minutes. Although the attenuation depends mainly on the DSD, a rain rate is used as a reference quantity, since it is much more descriptive, as well as easier to measure. The reflectivity measured with the radar is therefore transformed to the rain rate with the Marshall-Palmer relation. More direct approach would be to relate the attenuation with the measured reflectivity directly, however that would not change any of the conclusions, since, on a logarithmic scale, a simple power relation between reflectivity and rain rate reflects only as a linear transformation. The analysis of support size and time bin size showed quite strong dependency on the correlation. It is demonstrated that time bin $8h$ and support sizes of $n = 2^6$ and $n = 2^2$ for local and integral approach, respectively, provide a decent correlation $(0.6 - 0.7)$ between logarithms of measured attenuation and rain rate. Furthermore, the power model has been fitted over measured data and the value of the exponent has been compared to the values reported in the literature. The model shows best agreement with the Marshall-Palmer model, when the rain rate is gathered from the integral along the communication link. somewhat worse agreement is achieved with a local determination of rain rate. Results obtained with the rain gauge are the furthest from the expected, despite the fact that the correlation with the measured attenuation is the highest with the rain gauge measurements. The localized information from the rain gauge simply cannot provide enough information to fully characterize the rain conditions along the link. There are still some open questions to resolve, e.g. what is the reason behind the 8 h time bin giving the best result, how could we improve the correlation, etc. All these topics will be addressed in future work.

Acknowledgment

The authors acknowledge the financial support from the state budget by the Slovenian Research Agency under Grant P2-0095. Attenuation data was collected in the framework of the ESA-PECS project SatProSi-Alpha. Slovenian Environment Agency provided us with the data collected by their weather radars.

References

[1] J. Goldhirsh and F.L. Robison. Attenuation and space diversity statistics calculated from radar reflectivity data of rain. *Antennas and Propagation, IEEE Transactions on*, 23(2):221–227, 1975.

[2] M. Tamošiunaite, S. Tamošiunas, M. Žilinskas, and M. Tamošiuniene. Atmospheric attenuation due to humidity. *Electromagnetic Waves*, pages 157–172, June 2011.

[3] J. Rakovec and T. Vrhovec. *Osnove Meteorologije*. Društvo Matematikov Fizikov in Astronomov Slovenije, 2000.

[4] E. Villermaux and B. Bossa. Single-drop fragmentation determines size distribution of raindrops. *Nature Physics*, 5(9):697–702, September 2009.

[5] O. Fiser. The role of dsd and radio wave scattering in rain attenuation. *Geoscience and Remote Sensing New Achievements*, 2010.

[6] P. Pytlak, P. Musilek, E. Lozowski, and J. Toth. Modelling precipitation cooling of overhead conductors. *Electric Power Systems Research*, 81(12):2147–2154, 2011.

[7] S. Das, A. Maitra, and A.K. Shukla. Rain attenuation modeling in the 10-100 GHz frequency using drop size distributions for different climatic zones in tropical India. *Progress In Electromagnetics Research B*, 25:211–224, 2010.

[8] C.R. Nave. Blue sky and rayleigh scattering. http://hyperphysics.phy-astr.gsu.edu/hbase/atmos/blusky.html. Accessed: 2015-12-5.

[9] J.S. Marshall and W.McK. Palmer. The distribution of raindrops with size. *Journal of meteorology*, pages 165,166, August 1948.

[10] R. Uijlenhoet. Raindrop size distributions and radar reflectivity–rain raterelationships for radar hydrology. *Hydrology and Earth System Sciences*, pages 615–627, August 2001.

[11] Slovenian Environment Agency. (Agencija Republike Slovenije za Okolje).

[12] A. Vilhar, G. Kandus, A. Kelmendi, U. Kuhar, A. Hrovat, and M. Schonhuber. Three-site Ka-band diversity experiment performed in Slovenia and Austria. In *Antennas and Propagation (EuCAP), 2015 9th European Conference on*, pages 1–5. IEEE, 2015.

[13] U. Kuhar, A. Hrovat, G. Kandus, and A. Vilhar. Statistical analysis of 19.7 ghz satellite beacon measurements in ljubljana, slovenia. In *The 8th European Conference on Antennas and Propagation, to be held at the World Forum in The Hague, The Netherlands, on 6-11 APRIL 2014*, EuCAP 2014, pages 944–948, 2014.

[14] A. Vilhar, G. Kandus, A. Kelmendi, U. Kuhar, A. Hrovat, and M. Schönhuber. Three-site ka-band diversity experiment performed in slovenia and austria. In *The 9th European Conference on Antennas and Propagation, Montreal, Lisbon, Portugal, on 12-17 APRIL 2015*, EuCAP 2015, page 5, 2015.

[15] C. Kourogiorgas, A. Kelmendi, A. Panagopoulos, S.N. Livieratos, A. Vilhar, and G.E. Chatzarakis. Rain attenuation time series synthesizer based on copula functions. In *The 9th European Conference on Antennas and Propagation, Montreal, Lisbon, Portugal, on 12-17 APRIL 2015*, EuCAP 2015, page 4, 2015.

[16] Lord F.R.S. Rayleigh. XXXIV. On the transmission of light through an atmosphere containing small particles in suspension, and on the origin of the blue of the sky. *Philosophical Magazine Series 5*, 47(287):375–384, 1899.

Practical Implementation of Private Cloud with traffic optimization

D. G. Grozev, M. P. Shopov, N. R. Kakanakov
Technical University of Sofia / Department of Computer systems and Technologies, Plovdiv, Bulgaria
dgrozev@vmware.com, mshopov@tu-plovdiv.bg, kakanak@tu-plovdiv.bg

Abstract - This paper presents a practical implementation of a private cloud, based on VMware technology, optimized to support CoS and QoS (even when overlay technology like VXLAN is used) in a field of smart metering in electrical power systems and IoT. The use of cloud computing technologies increase reliability and availability of the system. All routing, firewall rules and NATs are configured using NSX. Implementation of CoS and QoS in virtual and physical network will guarantee necessary bandwidth for normal operation among other virtualized services.

I. INTRODUCTION

Cloud services and Internet of things (IoT) are among the most discussed terms in the latest research topics. They represent the two extremes of computer science – from small distributed devices to large computational infrastructure. The both topics are used together as IoT is one of the important sources of data that should be processed in large volumes [1, 2, 3]. But the use of cloud technologies in IoT are not limited to data processing and storage. The infrastructure can be used to deploy virtual devices and ease their management and dissemination of configurations. Deployment of virtual devices in the cloud will lead to new services: sensing-as-a-service or even sensor-as-a-service [4, 5].

A sustainable energy future depends on an efficient, reliable and intelligent electricity distribution and transmission system, i.e., power grid. The smart grid has been defined as an automated electric power system that monitors and controls grid activities, ensuring the two-way flow of electricity and information between power plants and consumers – and all points in between [6].

The smart power grid is a native application of IoT and Cloud technologies as the intelligent meters are highly distributed and they generate huge amount of measurement data. Texas instruments propose a smart grid architecture with intelligent meters and data concentrators. A data concentrator is the core of the energy management framework. It provides the technology to measure and collect energy usage data. The concentrator can also be programmed to analyse and communicate this information to the central utility database. Not only can the utility providers could use this information for billing services, but can improve customer relationships through enhanced consumer services such as real-time energy analysis and communication of usage information. Additional benefits of fault detection and initial diagnosis can also be achieved, further optimizing the operational cost [7].

All data processing and manipulation services as data concentrators could benefit of the elasticity of the cloud platforms and should be built very scalable. Nevertheless, exporting them in online processing clusters could reduce the privacy and increase the cost of communication. The best alternative in such cases is building a private cloud. Nowadays most of companies have their own cloud platform and virtual infrastructure, where they run business critical application and store data. Advantages of the private cloud includes:

- It's reliable and scalable. All resources are virtualized and in case of demand we can add more storage, or computing without downtime or impact.
- Fast provisioning. Using techniques like templates we can deploy thousands machines with few clicks.
- Automation. Pretty much everything can be automated using specific command-line tools or REST API.
- Common user interface: Decoupling the computation infrastructure and the input system, enables multiple user interfaces to exist side by side allowing user-centric customization.

In the paper, an architecture for a private cloud for energy measurements and data processing for smart power grid is presented. This architecture is consisting of a private cloud that hosts data storage, management and presentation services together with network of virtualized sensors. The paper discusses techniques to build and/or tune private cloud to adopt device management software along with other virtualized services.

II. BACKGROUND

A. Traffic optimization using NSX distributed logical router

There are two types of traffic in the cloud: North-South and East-West. North-South traffic is usually ingress/egress. This includes traffic from/to Internet and Intranet. Also it includes traffic between VMs from different port groups and networks that need to be routed in layer 3 physical devices. East-West traffic is VM or Management traffic that bounce between hypervisors (ESXi) [8].

With Distributed Logical Routing (DLR) deployment we move routing functionality to the hypervisor (kernel level) and effectively remove sub-optimal traffic path. Each ESXi host can route between subnets at line rate or nearly line rate speed. This means that if we have two

The presented work is supported by the National Science Fund of Bulgaria project under contract E02/12.

VMs from different networks, but on same host and use DLR communication between them won't leave the host. The idea behind DLR is not new one, but it is unique in virtualization world. All data centre switches have separation between control and data plane. This allows network engineers to restart or update control plane, while data plane is working and there is no down time. DLR has two components. The first one is DLR Control VM that is a virtual machine and second one is the DLR kernel module that runs in all ESXi hypervisor. DLR kernel module is called "route-instance" and has same copy of information in each ESXi host [10].

DLR has three types of interfaces: Uplink, Logical interface (LIF) and Management. Uplink is used by DLR Control VM to connect to upstream routers and it's called "transit" interface between physical and logical space. DLR support OSPF and BGP on its Uplinks, but you cannot run them both at same time. LIFs are interfaces that each ESXi host has in its kernel. They are layer 3 and acts as default gateway for all VM traffic connected to logical switches. Management interface is second interface on DLR control VM and again the idea comes from data centre switches, where management interface is separate physical port. This allow all management traffic to be separated in out of band (OOB) network. Services that should run on management interface are: SSH, Syslog, SNMP.

Propagation of routing information in DLR is a complex one. Initially all routing information is obtained and stored in Control VM. Actually routing daemons runs here. Then it is sent to NSX controller, which via permanent SSL encrypted channel push it in each ESXi host kernel. DLR LIFs have same IP addresses and vMACs in each ESXi host. vMACs are not visible in physical network. Fig. 1 shows all DLR components and relation between them.

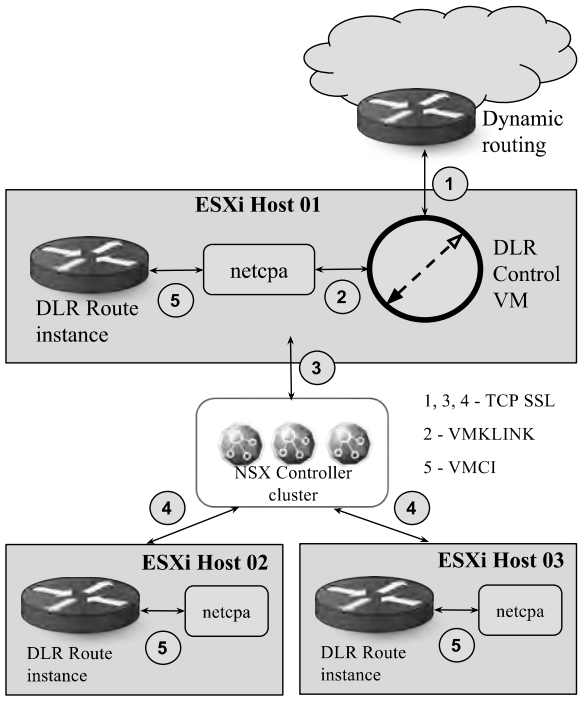


Fig.1. DLR configuration distribution [9]

If we move routing in the cloud using NSX ESRs, physical routers and layer 3 switches will be offloaded, but traffic between hypervisors won't be reduced. Virtual router (ESR) lives on a singles host, which means that VM traffic will reach that hypervisor via layer 2 network, will be routed, and goes back to host where VM is. With DLR *netcpa* process on each host stores routing information and when communication virtual instances are on the same host no network traffic is sent in the physical network.

It's always easy to put all VMs in a single port group and network and get rid of routing at all. This design is possible if there no requirements for security. Separating VMs in different networks will give us full control over exchanged traffic.

B. Traffic optimization using vDS features: shaping, NIOC, reservation and limits

Backbone of each cloud is virtual switch, where all VMs are connected. With distributed virtual switch (*vDS*) we can optimize traffic using following features: Shaping – a shaper could be applied on each port group; Network IO control (NIOC).

NIOC concept revolves around resource pools that are similar in many ways to the ones already existing for CPU and Memory. NIOC classifies traffic into six predefined resource pools as follows: vMotion, iSCSI, Fault tolerant (FT) logging, Management, NFS, Virtual machine traffic. With this feature each traffic type is configured with shares. In case of host physical NIC congestion traffic will receive bandwidth that is equal to available physical bandwidth multiplied by ration of traffic shares and sum of all traffic shares of traffic participating during congestion. There are 3 layers in NIOC (Fig 2): teaming policy, shaper and scheduler [9].

There is a new method of teaming called LBT (Load base teaming). It basically detect how busy those physical network interfaces are, then it will move the flows to different cards. LBT will only move a flow when the mean send or receive utilization on an uplink exceeds 75 percent of capacity over a 30-second period. LBT will not move flows more often than every 30 seconds.

There are two attributes (Shares and Limit) one can control over traffic via Resource Allocation. Resource Allocation is controlling base on vDS and only apply to this vDS. It applies on vDS level not on port group or dvUplink level. Shaper is where limits apply. It limits traffic by the class of traffic. Each vDS has it's own resource pool and resource pool are not shared between vDS.

Shares apply to dvUplink Level and each share rates will be calculated base on traffic of each dvUplink. It controls share value of traffic going through this particular dvUplink and make sure share percentage is correct. This concept allows flexible networking capacity partitioning to help users to deal with over commitment when flows compete aggressively for the same resources.

NIOC ensures traffic isolation so that a given flow will never be allowed to dominate over others, thus preventing drops and undesired jitter. On each traffic type we can configure reservations and limits.

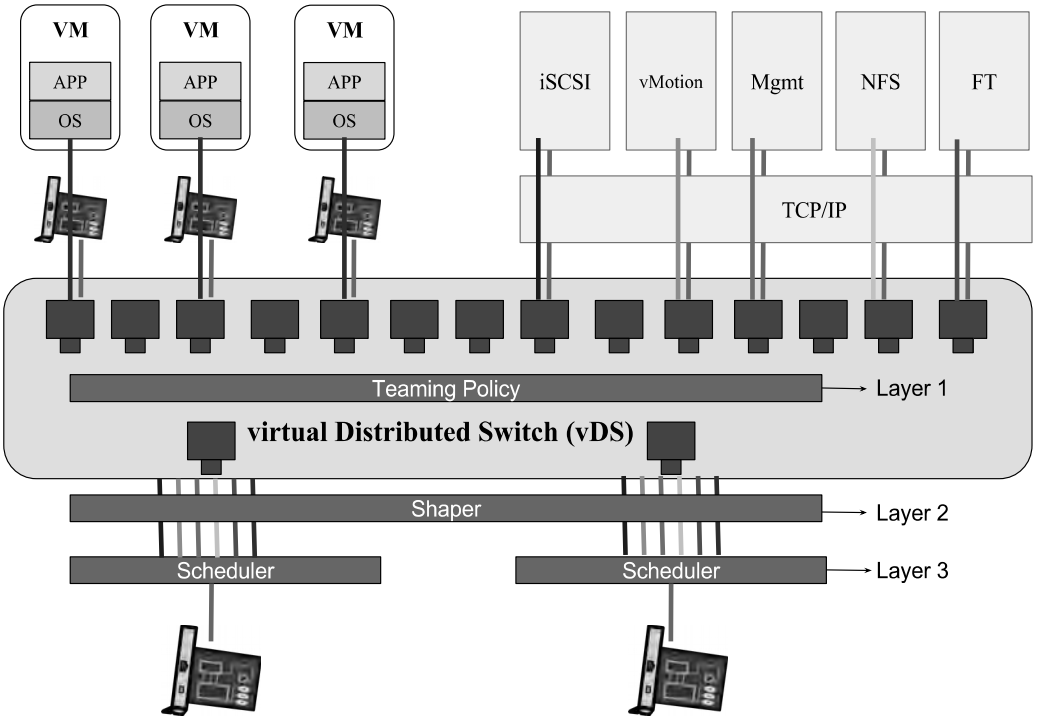


Fig. 2 Network I/O Control [10]

C. Classification and marking, which prepare traffic for further processing in physical network

Usually CoS and QoS are not configured in the cloud. If any, it's applied on physical network devices, where classification, marking and policing happened. Nowadays VMware network virtualization support CoS and QoS. They are relatively new features that come with vSphere 5.5. Enabling CoS and QoS will allow us to mark the traffic on vDS level and offload physical network. We have to configure physical switches to trust CoS and QoS values.

III. USE-CASE SCENARIO: ENERGY MANAGEMENT SYSTEM IN THE CLOUD

A. Architecture of the proposed energy management system in the cloud

The main architecture proposed is consisting of: message broker for supporting communication between sensing elements and application and storage elements; NoSQL data storage (Data Logger) that stores the raw measurements form sensing devices; Application server that provides platform for running the data manipulation and knowledge extraction scripts; relational DB server for storing processed data; Presentations server that runs software for preparing the data visualisation – tables, graphics, maps; and multiple virtual sensor nodes – each one representing a real measurement unit. The virtual sensor nodes should be a replica of the physical sensor – keeping the data extracted from real sensor and providing interface for its configuration. The servers in the infrastructure communicate with virtual sensors for configuration or data extraction and the synchronization between the physical and virtual sensor is done in isolation to the others with specific protocols for the sensor . The logical topology used is shown on Fig. 3.

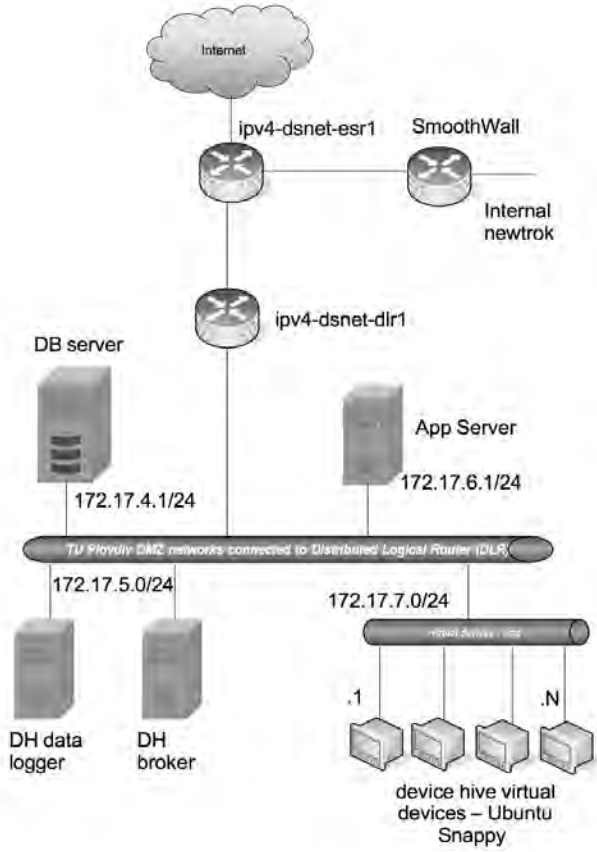


Fig. 3. Logical Topology

B. Private cloud in Technical university Sofia, branch Plovdiv

For the implementation of energy management system a private cloud is built in Technical University – Sofia, branch Plovdiv. It utilizes five hosts with total CPU 83 GHz, total RAM 182GB, total storage >6TB (4 Dell PowerEdge 1950 hosts and one HP ProLiant D80 G7). The hosts are interconnected via two Cisco 3750 Gigabit access switches and one top-of-rack Datacenter switch AS4600-54T-C (with Cumulus Linux), and the they use Fujitsu Eternus DX90 S2 Disk storage system connected using two Brocade Fibre Channel DS-5000B switches. All hosts run ESXi 5.5 operating system, organised in one cluster and are registered in one vCenter. All management services run on separate cluster. Physical topology of our cloud architecture is shown on Fig. 4.

The presented cloud infrastructure can provide different cloud services to energy measurement system. Apart from obvious IaaS where user can deploy its own virtual machines or PaaS where user can use its own preconfigured instances of a message broker, data storage and virtual devices, the following specific cloud services are used:

- Data-as-a-Service – users can extract data from Data Logger or RDB to apply own algorithms for analyses;
- Sensing-as-a-Service – users can extract Raw data from sensor measurements;
- Sensor-as-a-Service – users can use the existing virtual sensor infrastructure to apply monitoring and control functions.

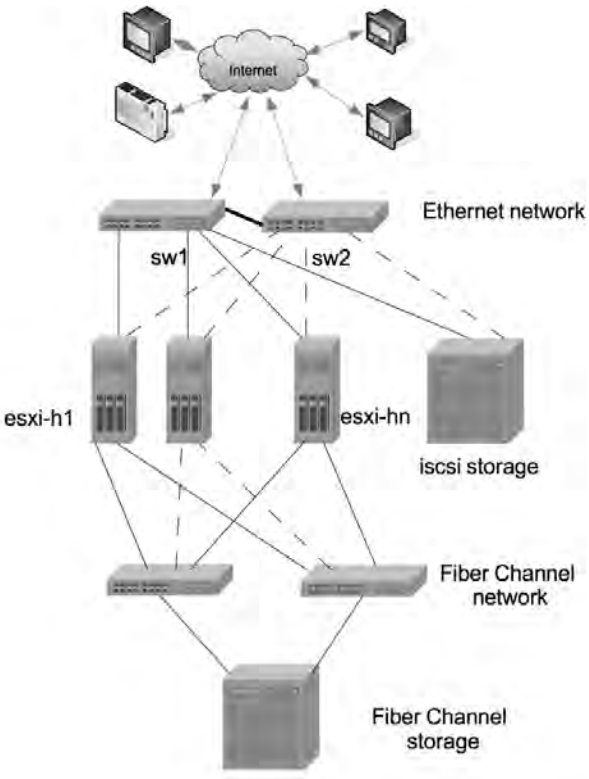


Fig. 4. Physical topology

IV. TRAFFIC OPTIMIZATION IN THE PRIVATE CLOUD

A. Traffic types and classes in the use case scenario

In our private cloud we recognize following types of traffic:

1. Management – this include all traffic generated by virtual infrastructure: virtual machine migration (vMotion), connection to the network storage (iSCSI, NFS), fault tolerance, user interface for management (WebGUI, PowerCLI), etc.

2. Virtual Machine own traffic – this is the traffic generated from all VMs connected to VLAN based port groups in the distributed virtual switches.

3. VXLAN traffic – this include traffic generated from all Virtual Machines connected to virtual logical switches.

Within VM traffic we have following classes:

- Reading sensors;
- Configure sensors;
- Store measurements;
- Extract data series for manipulation;
- Store processed data in RDB;
- Extract data for presentation;
- Export data.

The first two classes are between the message broker node and virtual devices. The third class is between message broker and Data Logger. The forth class is between Data Logger and Application Server. The firth class is between Application server and RDB server. The sixth class is between RDB and presentation server. The seventh class is between data centre and external Internet services.

B. Preliminary results and discussion

The applied optimizations in the private cloud in technical university include:

- NIOC is enabled on vDS that span over hosts running virtual services of the energy management system. We have added custom defined network resource pool for VXLAN and configured it with 100 shares and no limit. This will guarantee that in case of congestion VXLAN traffic will have enough bandwidth. Here we omit any QoS settings, because we want outer VXLAN QoS to match original traffic. All management traffic is placed in system resource pools with default shares – 50 shares per pool.
- DLR is used to route traffic between different DMZ zones. If two heavy communicating VMs are on the same physical host the traffic do not leave the host and physical network is offloaded.
- DRS Affinity Rules are added on vCenter to keep VMs that have to exchange large amount of data between each other on the same host.
- CoS and QoS rules are added on the vDS VXLAN port-group to prepare the traffic for prioritization in the physical networking. We

create exact rules that select traffic between VMs, by source/destination IP and protocol ports. Applying traffic shaping is for entire port-group. Having this in mind we group VMs in appropriate ones.

Some initial tests are done for proving the workability of the system. On the physical cluster, several Virtual Machines with Ubuntu 14.04 server are started. They represent different services, specific to smart metering: NoSQL data store or Data Logger; Application server for data manipulation and knowledge extracting; relational DB store for storing extracted data series and/or data patterns; Presentation server for data visualization. Virtual devices are represented by virtual machines running Ubuntu Snappy. Additional virtual devices can be deployed on emulated ARM processors using QEMU.

These VMs are configured in different subnets according to the logical topology (Fig. 3). Configuration of the DLR on all VMs is the same. The routing table of one VM is shown on Fig. 5. On each VM several LIFs are configured and multicast group is assigned to each:

TABLE I. DESCRIPTION OF A LIF IN VMS

Mode	Routing, Distributed, Internal
ID	Vxlan:100005
IP	172.17.7.254
Connected Dvs	dglabs
VXLAN Multicast IP	239.0.0.5
DHCP Relay	Server List: 172.17.0.17

For testing the optimization traffic generators are started on different VMs and some preliminary results are obtained. The tests gather information about delay and jitter of VXLAN traffic when source and sink run on same ESXi host or on different ones. These values are compared to check if DLR works as expected.

TABLE II. DELAY FOR ONE FLOW BETWEEN TWO VMS*

	min	**max**	**avg**
One host	274.2	283.0	274.3
Separate hosts	836.8	842.5	837.0

* Results are in microseconds (μs)

V. CONLUSION

The paper discusses the traffic optimization techniques that can be used in a cloud infrastructure with virtual networking (VXLAN). These techniques are applied in a use case scenario – an architecture for energy management system in a cloud. The presented architecture is implemented in a private cloud infrastructure in Technical University of Sofia, Plovdiv branch in virtual laboratory for Distributed Systems and Networking (http://dsnet-tu-plovdiv.bg).

Proposed implementation allows building all layer of a energy management system using cloud technologies which ease administration, adds scalability and reliability, and is general enough to be applied in wide range of scenarios.

The preliminary results shows that using DLR for routing between VXLAN reduces the overall traffic in the cluster network and applying NIOC lowers the ratio of dropped and delayed packets due to high network loads.

The results are preliminary and more detailed experiments should be made to check the influence of prioritization of one traffic to others and to obtain detailed numerical values for the reduction of traffic and delay. These results can be verified in different scenarios with different distribution and load of traffic classes.

ACKNOWLEDGMENT

The presented work is supported by the National Science Fund of Bulgaria project “Investigation of methods and tools for application of cloud technologies in the measurement and control in the power system” under contract E02/12 (http://dsnet.tu-plovdiv.bg/energy/).

REFERENCES

[1] Alamri A, Ansari WS, Hassan MM, Hossain MS, Alelaiwi A, Hossain MA, "A survey on sensor-cloud: architecture, applications, and approaches," International Journal of Distributed Sensor Networks, Article ID 917923, Feb 2013, doi:10.1155/2013/917923, 2013.

[2] Beng, L., “Sensor cloud: Towards sensor-enabled cloud services,” Intelligent Systems Center Nanyang Technological University, 2009.

[3] Botta, A.; de Donato, W.; Persico, V.; Pescape, A., "On the Integration of Cloud Computing and Internet of Things," in Future Internet of Things and Cloud (FiCloud), 2014 International Conference on , pp.23-30, 27-29 Aug. 2014.

[4] Zaslavsky, A., C. Perera, D. Georgakopoulos, “Sensing as a service and big data”, Proceedings of the International Conference on Advances in Cloud Computing (ACC), Bangalore, India, July, 2012, Pages 21-29 (8), arXiv preprint (1301.0159).

[5] Rao, B.B.P.; Saluia, P.; Sharma, N.; Mittal, A.; Sharma, S.V., "Cloud computing for Internet of Things & sensing based applications," in Sensing Technology (ICST), 2012 Sixth International Conference on , pp.374-380, 18-21 Dec. 2012.

[6] Wu L., G. Kaiser, C. Rudin, R. Anderson, “Data Quality Assurance and Performance Measurement of Data Mining for Preventive Maintenance of Power Grid”, Columbia University Academic Commons, 2011, http://hdl.handle.net/10022/AC:P:12174.

[7] P. Prakash, “Data concentrators: The core of energy and data management”, Texas Instruments white paper, Copyright © 2013, Texas Instruments Incorporated.

[8] VMware NSX Technical Product Management Team, "VMware® NSX for vSphere Network Virtualization Design Guide ver 3.0", Aug 21, 2014 , online [04.02.2016]:www.vmware.com/files/pdf/products/nsx/vmw-nsx-network-virtualization-design-guide.pdf

[9] B. Hedlund, "Distributed virtual and physical routing in VMware NSX for vSphere", The Network Virtualization Blog, November 25, 2013. online [04.02.2016]: https://blogs.vmware.com/networkvirtualization/2013/11/distributed-virtual-and-physical-routing-in-vmware-nsx-for-vsphere.html

[10] V. Deshpande, "Network I/O Control (NIOC) Architecture – Old and New", VMware vSphere Blog, Posted on January 25, 2013, online [01.02.2016]: https://blogs.vmware.com/vsphere/tag/nioc

GLOSSARY

VXLAN (Virtual Extensible LAN) – A Virtual Network that emulates an Ethernet broadcast domain

NSX - VMware NSX is the network virtualization platform for the Software-Defined Data Center (SDDC);

DRS (Distributed Resource Scheduler) – technology for balancing the computing capacity by cluster to deliver optimized performance for hosts and virtual machines.

VDS – VMware vSphere Distributed Switch (VDS) provides a centralized interface from which you can configure, monitor and administer virtual machine access switching for the entire data center.

DLR – Distributed Logical Router. It separates Control and Data plane. Control plane is a VM, but data plane is part of hypervisor kernel

vMotion – VMware vSphere live migration allows you to move an entire running virtual machine from one physical server to another, without downtime.

```
VDR default+edge-5 Route Table
Legend: [U: Up], [G: Gateway], [C: Connected], [I: Interface]
Legend: [H: Host], [F: Soft Flush] [!: Reject] [E: ECMP]

Destination      GenMask          Gateway          Flags    Ref  Origin   UpTime     Interface
-----------      -------          -------          -----    ---  ------   ------     ---------
0.0.0.0          0.0.0.0          172.17.0.17      UG       1    AUTO     80987      186a000000002
172.17.0.0       255.255.255.240  172.17.0.17      UG       1    AUTO     80987      186a000000002
172.17.0.16      255.255.255.240  0.0.0.0          UCI      1    MANUAL   174713     186a000000002
172.17.4.0       255.255.255.0    0.0.0.0          UCI      1    MANUAL   174543     186a00000000a
172.17.5.0       255.255.255.0    0.0.0.0          UCI      1    MANUAL   174047     186a00000000b
172.17.6.0       255.255.255.0    0.0.0.0          UCI      1    MANUAL   173416     186a00000000c
172.17.7.0       255.255.255.0    0.0.0.0          UCI      1    MANUAL   173290     186a00000000d
194.141.41.224   255.255.255.248  172.17.0.17      UG       1    AUTO     80987      186a000000002
  #
```

Fig. 5. DLR routing table on a host

Improving Data Locality for NUMA-Agnostic Numerical Libraries

P. Zinterhof
University Salzburg/Dept. of Computer Sciences, Salzburg, Austria
peter.zinterhof3@sbg.ac.at

Abstract – Modern multi-socket servers are frequently based on the NUMA paradigm, which offers scalability but also introduces potential challenges. Many numerical libraries have not been designed for such architectures specifically, so their node-level performance relies heavily on the inherent quality of their parallelization (e.g. OpenMP) and the use of highly specialized tools and techniques such as thread-pinning and memory page placement.
In this paper we propose a simple and portable framework for improving performance of NUMA-agnostic numerical libraries by controlling not only the location of threads but also the location of allocated memory. In addition to many related approaches we also apply fine-grained control over memory placement that is implemented by means of the operating system's first-touch principle.

I. INTRODUCTION

We see a steady trend towards ever increasing performance on all levels of computational hardware. It is a well-established fact, that with current silicon-based technology the performance of typical microprocessors cannot be increased any more just by increasing their clock rates. This is mainly owed to the fact that doubling the clock rate roughly equals the quadrupling of power consumption. In such a setting proper cooling quickly becomes either technologically unviable, as the resulting energy density on the surface of the chips would require very advanced and expensive kryo-techniques or would be outlawed by simple terms of economic rationality. The latter especially being the case in the area of supercomputing where power-consumption of the installation is already one of the main limiting factors.

Instead, performance increases are to be achieved by several design factors that are being selectively combined by hardware designers. On the core-level one sees larger cache memories and more powerful instructions that rely on advanced techniques (e.g. speculative execution, instruction re-ordering, etc.) and wide execution units for vector processing. On the chip level we observe ever increasing numbers of parallel cores, located within a single die and connected to each other by caches or special bus systems. This level also marks the beginning of the non-uniform memory architecture (NUMA). On the system-level multiple CPUs are integrated into single systems, or, as in the case of large NUMA-machines, multiple systems seamlessly integrate into a single, potentially very large system that is operated under a single system-wide operating system kernel. Each level offers performance advantages but also potential pitfalls.

In this paper we focus on performance optimization on cache coherent NUMA-systems, as they constitute the main building blocks of many current high-performance computers and compute clusters.

Certainly, there is a wide range of computational workloads on cluster installations, expressed by an even greater range of software applications. It is our conjecture, that many of these applications are based on external numerical libraries, which in some cases might still be amenable to improvements. Despite much work in the field and a growing awareness of the problems in achieving high performance, software development and performance tuning can be still tricky, especially when knowledge, expertise, or proper tools are scarcely available. It is the goal of this paper to demonstrate an easy path to performance improvements in the use of numerical libraries. That is, we take a zero-knowledge approach with regards to the concepts and implementation details of these libraries, but concentrate on the orchestration of working threads and memory allocations, both of which can be controlled to some extent at the level below the actual library functions.

We will base our work on the well-known basic linear algebra subroutines-library (BLAS) that is a tremendously successful numerical library, important to many fields of research as well as application design. The BLAS-library already supports symmetric multi-processors by means of OpenMP, but other than that is agnostic to the special characteristics of NUMA-based systems. Nevertheless, there do exist tools for thread-pinning that allow certain threads of the parallel execution to be pinned to certain cores of the CPUs. Henceforth, the allocation of working memory for BLAS-operations and the distribution of it within a system comprising several NUMA-nodes is either left to the application or to some allocation policy that is defined at the start of the application.

Therefore the proposed framework aims at the proper allocation of memory without requiring intricate knowledge of the inner structure and workings of the BLAS-library and at some more detailed level than offered by most available tools. It seems important to note

that our framework uses BLAS as a test case but is not being limited to this library.

II. Related work

A. Research

Improving performance for applications on parallel computers is an important area both in research and development. The importance is certainly augmented by the increasing number of high-performance machines worldwide and the fact that almost every new system today is based on multi-core CPUs.

One approach to optimize performance is to find optimal placements for multiple threads onto a multi-core CPU. In [1] this placement is guided by hardware performance counters, that have been introduced to modern x86-based CPUs. The proposed autopin-tool[1] is probing several different settings of thread-scheduling until an optimum is declared and being used for scheduling of the application for the remaining runtime.

A similar "on-the-fly"-optimization technique will be shown in our framework, too. It has also been shown [2] that there often is no globally optimal thread placement in real-life applications, as for different parallel execution blocks in the code different – locally adjusted – scheduling will be superior. This local adjustment is also the main path we pursue in our proposed framework, although we concentrate on the placement of memory pages in main memory in conjunction with the standard thread placement that is governed by OpenMP. Another interesting feature of the approach described in [2] is the operation on binary executable files instead of the source code of the application. This certainly allows for an easier deployment and possibly wider adoption among users, as no source code has to be available and no recompilation takes place. As pointed out in [3], improving data-locality is only one – albeit important – factor with respect to performance. There are situations in which an imbalance in the layout of pages will not hurt performance. In special cases, one can even rely on an aggregation von L3 caches in multi-processor environments to even gain super-linear speedups [4]. On the contrary, even in well-balanced layouts performance might be sub-optimal due to congestion of memory links that introduce additional access latencies. There is no single optimum solution for these kinds of problems, since the answer also involves other aspects such as the tradeoff between time-to-solution vs. speedup and the decision between manual tuning and auto-tuning. Depending on the setting different paths might be regarded to be superior to others. There is also an interesting relation of energy efficiency of a computation and its performance as pointed out in [5], where high cache utilization is enforced by binding threads to cores either by the pthread-API but also by means of the likwid toolset as in our own setting.

In general, improvements on data-locality not only support shorter times to solution but also increase energy efficiency and as a direct consequence of the latter even higher performing computer systems become feasible within the same power envelope.

B. Frameworks and Tools

Amongst solutions for thread-pinning and control of memory allocation at the level of libraries there is a range of tools which offer support for these tasks at the system level. This latter approach removes the need of making decisions by the software developer, but enables to exert control over certain aspects of the run-time behavior of an application by the user or system administrator.

- Likwid

Likwid [6] is a set of tools for performance optimization on NUMA-systems. It can display all kinds of hardware-(topology) information, read out hardware performance counters and energy meters. It can also be used to pin threads to cores for the runtime of an application. We will base our experiments (section IV Results) on Likwid.

- Taskset

Taskset is a command usually included in Linux distributions. It enables the user to retrieve or set processor affinities for a parallel program. An application will only be run on a set of cores that can be specified by some mask. For instance, the command 'taskset –c 1,2,8,16 application.exe' will ensure that the 4 parallel threads of applications.exe will not be executed on any other cores that 1,2,8, and 16. It is not entirely clear, whether affinities are fixed during the run time of the job or whether the kernel is still allowed to change affinities of threads within the requested bounds of cores 1,2,8, and 16. If one application is intended to be confined to a single CPU socket or NUMA-node (there is no strict 1:1 mapping between the two terms as there frequently can be found more than one NUMA-node on a single CPU socket) the above question seems irrelevant. Whereas in the case of high performance computing we most likely want to employ all available cores on the machine which would render an affinity set of <1,2,3, .., MaxCore> rather redundant under the assumption that threads would still be allowed to move between cores.

- Numactl

Numactl is comparable to the taskset command in Linux, but this command adds functionality by allowing to bind memory to specific NUMA-nodes. Also, it offers a more relaxed way of specifying the location of threads by using the coarser term of "NUMA-node" instead of core numbers. With the intention of locking threads to closely connected memory in order to increase data locality, it is sufficient to tell the system that a specific set of threads shall be confined to a certain NUMA-node as all cores within a NUMA-node enjoy the same memory access characteristics.

- AutoNUMA

AutoNUMA [7] is both a memory and thread placement tool that has been included in the Linux kernel. By keeping track of an application's memory accesses is determines an online strategy for page migration which is

then executed by a separate kernel thread. Unfortunately, some distributions need to be patches before using AutoNUMA, thus rendering the software less portable.

III. IMPROVING DATA LOCALITY

Our framework aims at performance improvements by locally adjusting memory distribution over different NUMA-nodes by exploiting the well-known first-touch policy in Linux. Under first-touch policy memory pages of some newly allocated block of memory will first be mapped to some Zero-page, such that reading from any position of the block will result in bit patterns 0. Not until the first write operation the operating system will decide which NUMA-node the actual page of memory will be physically placed on. Linux will take that NUMA-node to hold the page that is closest to the thread which is making the write. Hence Linux already provides high data locality for the given thread and block of memory under the rationale that the first-calling thread will continue to make frequent use of a memory page afterwards.

Furthermore, memory distribution is being controlled in a fine-grained fashion by selectively choosing NUMA-nodes for all arrays that will be used by BLAS within a given application program.
In general we find two ways of providing and initializing arrays for use with BLAS.

- Allocation with explicit initialization.
  Explicit initialization of arrays can be the consequence of the applications 'natural' dataflow (e.g. some input data have to be stored in memory before BLAS operations can take place on them) or the personal programming style of the developer.
  A important factor will be whether this initialization is accomplished by a sequential thread or a pool of parallel threads, especially on a system employing first-touch policy. A sequential thread will typically force all pages to be physically located at its local NUMA-node, until this node will be exhausted and follow-up pages will 'spill over' onto neighboring NUMA-nodes. Initialization within a 'parallel'-block in OpenMP will lead to a balance of the distribution of pages over a certain number of NUMA-nodes, as OpenMP always aims to provide a balanced distribution of tasks.

- Allocation without initialization.
  When no explicit initialization precedes the use of some memory block or array by BLAS, the operating system will choose some NUMA-node on the basis of the system-wide placement policy. This may be either the above mentioned first-touch or the round-robin policy, depending on Linux kernel settings.

In a typical application we do not see a single call to some library function but a series of calls, often embedded in a dynamic control flow. Here one can encounter a series of different library functions repeatedly operating on the same block of memory, with each function potentially sporting a different memory placement for its optimal performance, but as the placement of memory is done only once during the first-touch phase this placement might prove inadequate for some of the subsequently invoked library functions. This can affect performance negatively as a high degree of data-locality cannot always be ensured in such a setting.

In order to deal with that problem, our framework not only enables control of memory placement at the level of single pages in a multi-threaded (OpenMP) environment but also allows changes to that placement dynamically, albeit not in an automatic fashion.

Control of placement is based on the definition of the topology of the placement, followed by a function which employs a set of parallel threads to force the physical mapping of pages onto the NUMA-node targets (Fig. 1). This mapping has to be accomplished by way of parallel threads in order to exploit the first-touch policy, which states that a page will be physically stored on that NUMA-node of which the calling thread is part of during the first write operation to that page.

The placement is defined as a vector with each element denoting the identification of the thread which is to execute the first touch on the corresponding memory page. We base our experiments on a Linux kernel with a page size of 4096 bytes.
Function 'touch' (Fig. 1) applies some write operation to array 'memblock' at memory position i*4096, thus forcing ownership of page i to the NUMA-node that corresponds to the calling thread 'TID'.

```
void placePages (placementVector,
memblock)
{
#pragma omp parallel private (i, TID)
{
TID=omp_get_thread_num();
 for (i=0; i < number_of_pages; i++)
    {
     if (TID == placementVector[i])
        touch (memblock, i * 4096);
     }
}
}
```

Figure 1: OpenMP based page placement

By fr : the placement can be changed at will during runtime. This operation could be substituted with the page migration function of the NUMA library, but with maximum portability in mind the above approach has been favored. The operation imposes very little overhead as all memory use in a NUMA-system will have to go through some initialization anyway. The only overhead is given by the

parallel for loop, which seems rather negligible by never exceeding 8 ms wall time in our experiments.

Three methods of deriving proper placement vectors have been employed.

- random placement
- genetic optimization[1]
- Monte-Carlo based random search

The "random placement" method places pages on randomly chosen NUMA-nodes. It works surprisingly well (see section IV. Results). By subjecting the placement of pages to optimization by means of a genetic algorithm we also aim to increase data-locality and overall system performance. The genetic algorithm is based on a simulated population of genetic individuals that encode different entries of the placement vectors. By selectively recombining and permuting individuals under evolutionary pressure the quality of the placement will be improved from generation to generation. This algorithm is based on the assumption that for each two individuals of the population a clear distinction of their respective performance can be made. As it turns out, pinning threads to specific cores still allows for some variance in execution times that effectively counteracts this necessary distinction and thus hampers the applicability of the genetic algorithm.

As a result the genetic algorithm shows similarities to another method which can be called Monte Carlo-based random search.
Our framework therefore offers support for performance-sampling (based on Monte Carlo method) for an arbitrary number of repetitions of calls to a BLAS-function. Over time this random search typically yields placement vectors of increasing quality. The user has then to provide some stopping criteria that concludes the search operation with the optimal placement vector to be employed for all subsequent calls to that BLAS-function. It has to be noted that the search operation can be integrated within the normal operation of the application ('on-the-fly' optimization), so that no serious overhead is incurred. Placement vectors can also be stored and retrieved for later use.

IV. RESULTS

For assessment we applied the framework to three operations (dgemm, dgesv, dsytrd+dorgtr+dsteqr) of the BLAS library in double precision (64 bit). BLAS functions have been applied to randomized input data and linked against the Intel MKL v11.2 library on a quad-socket SuperMicro system consisting of 4 AMD Opteron 6386 SE CPUs (64 cores total) and 512 GB RAM, which is arranged in the form of 8 NUMA nodes. The likwid-pin[6] utility has been built for the employed CentOS 6.3 Linux system and subsequently being used to pin 64 OpenMP threads to the 64 physical cores of the test system. Each experiment has been repeated 10 times and reported numbers denote the average performance or run-time, respectively.

The first experimental run involves the solution of a linear system of equations, where the 'naïve' setting of extending an otherwise sequential code with the parallelized BLAS routine 'dgesv' is compared to the code parallelized with OpenMP and the code parallelized and optimized by the proposed framework.
Also, likwid [6] has been used for pinning threads during the latter two trials, leaving the naïve code under the standard regime of Linux both for scheduling threads and placement of memory pages.

Table 1 displays various execution times and Fig. 2 displays run times for different problem sizes, ranging from 1024 up to 16384 in steps of 512.
Our framework constantly performs better than the other two approaches.

Example two is based on a triple call to BLAS which computes eigenvalues and eigenvectors of a dense matrix. The placement strategies 'OpenMP' and 'framework' perform better than naïve (single core) memory allocation and placement along with the lack of thread pinning up to some 30 %. Nevertheless, execution times of 'OpenMP' and 'framework' (Fig. 3) are very close and probably within the margins of measurement errors, so we cannot declare a clear winning scheme in this realm.

Table 1: performance for BLAS dgesv function, in seconds of wall time

Placement method/ problem size	naive	OpenMP	framework
4096	1.04	0.58	0.57
8192	7.39	4.09	4.05
12800	13.69	8.66	8.58
16384	37.79	24.50	23.64

[1] Experiments have been conducted by means of the Genetic Algorithm Utility Library, http://gaul.sourceforge.net with an initial population of 128 individuals under a Darwinian scheme with crossover/mutation and migration probabilites of 0.03/0.09/0.07.

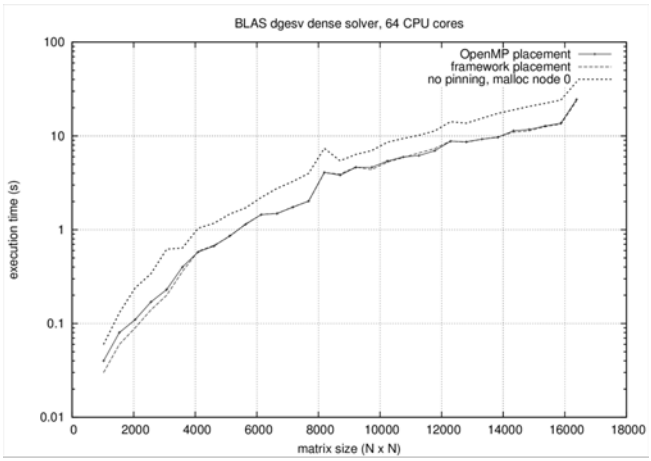


Figure 2: solve system of equations, BLAS dgesv

Example three is based on matrix/matrix-multiplication by means of the dgemm-function of the BLAS. The implementation is based on a blocking algorithm, which aims at high utilization of the three levels of processor caches in modern CPUs. Thus, dgemm already provides near to optimum performance in its standard incarnation and it should be less amenable to optimizing page placements. Fig. 4 displays performance for dgemm with input matrices A, and B being initialized in parallel by the 64 OpenMP-threads, placements being done by the proposed framework and also by using first-touch based placements as a result of the output of a preceding dgemm operation. So in the latter case we effectively ran two (dummy) dgemm matrix multiplications first.

By just employing the first-touch policy within these operations we obtained page placements for the resulting output matrices C1 and C2 that reflect typical placements of output matrices for dgemm operation. Only in the last step we measured performances of the dgemm operation, now being based on C1 and C2 as input matrices. The described setup aims to model the standard data flow for some abstract numerical application which most often bases the input data of a function on the output data of another preceding function.
Our proposed framework will lead to an average increase in performance by 0.5 % over the OpenMP-style

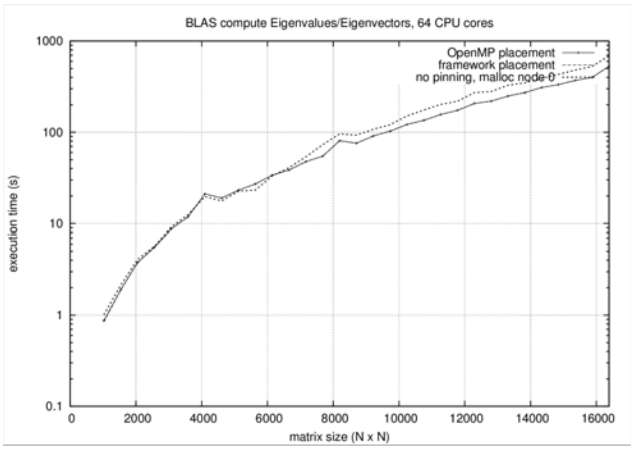


Figure 3: BLAS eigenvector/values computation

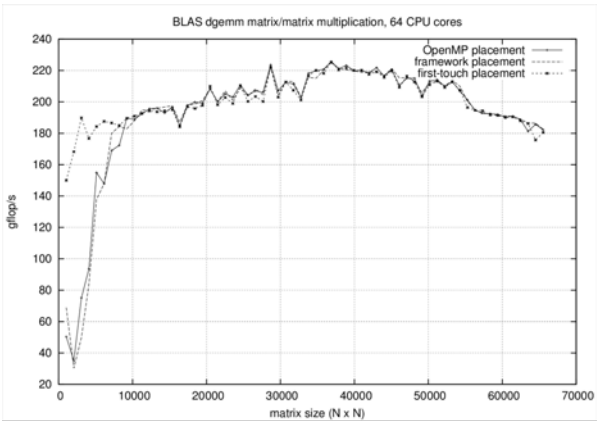


Figure 4: BLAS dgemm performance

initialization and placement while employing first-touch policy yields an average speedup of 0.35 %.

Applying Monte Carlo random search on dgemm for 150 steps yields an additional speed up of 1 % (Fig. 5)
This may sound very modest, but with this optimization mode being easily integrated into existing algorithms it still can be regarded as a meaningful improvement. Also, offline optimization can be used with the resulting placement vector being retained for later use.

V. CONCLUSION AND OUTLOOK

Based on the widely known BLAS library we have demonstrated a simple but nevertheless effective framework for optimizing data-locality within numerical libraries that have not specifically been designed for use on Non Uniform Memory Architecture-based parallel systems. Our approach operates at the level of memory allocation and does only require an operating system that is capable of first-touch page placement policy and thread pinning. Thus we regard the proposed framework to be highly portable.
Also, no prior knowledge of implementation details of the numerical library is required by the user in order to enjoy the benefits the framework.

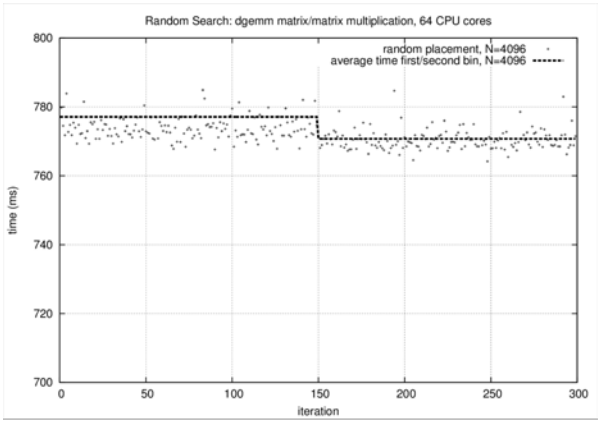


Figure 5: Monte Carlo optimization over first 150 samples. Validation for iterations [151,300]

Future work will explore some fused approach, which integrates optimization of both thread placement and page placement as a single objective. This multi-modal optimization can be expected to deliver further improvements of performance of numerical libraries on modern NUMA-systems.

REFERENCES

[1] T. Klug, M. Ott, J. Weidendorfer, C. Trinitis, „autopin – automated optimization of thread-to-core pinning on multicore systems“, In: P. Stenström (ed.) Transactions on HiPEAC III, LNCS, vol. 6590, pp. 219-235, Springer, Heidelberg, 2011.

[2] A. Mazouz, S. Touati, D. Barthou, “Dynamic Thread Pinning for Phase-Based OpenMP Programs”, Euro-Par 2013 Parallel Processing, pp 53-64 , Lecture Notes in Computer Science, 8097, Springer Berlin Heidelberg, 2014.

[3] F. Gaud et al., “Challenges of Memory Management on Modern NUMA Systems”, Communications of the ACM, pp 59-66, Vol. 58, No.12, 2015.

[4] G. Kosec, M. Depolli, A. Rashkovska, R. Trobec, “Superlinear speedup in a local parallel solution of thermo-fluid problems”, OpenMP parallelization of a local PDE Solver”, J. Computers and Structures, pp 30-38, Vol. 133, Pergamon Press, Elmsford, NY, USA, 2014.

[5] D. Davidovic, M. Depolli, T. Lipic, K. Skala, R. Trobec, “Energy Efficiency of Parallel Multicore Programs”, Scalable Computing: Practice and Experience, Volume 16, Number 4, pp. 437–448. http://www.scpe.org, 2015.

[6] J. Treibig, G. Hager, G. Wellein, “LIKWID: A lightweight performance-oriented tool suite for x86 multicore environments,” http://arxiv.org/abs/1004.4431

[7] J. Corbet, “AutoNUMA: The other approach to NUMA scheduling”, LWN.net, http://lwn.net/Articles/488709/, 2012.

Use Case Diagram Based Scenarios Design for a Biomedical Time-Series Analysis Web Platform

Alan Jovic*, Davor Kukolja*, Kresimir Jozic** and Mario Cifrek*
* University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** INA - industrija nafte, d.d., Zagreb, Croatia
Corresponding author: alan.jovic@fer.hr

Abstract - Biomedical time-series analysis deals with detection, classification and prediction of subjects' states and disorders. In this paper, we present requirements and scenarios of use for a novel web platform designed to analyze multivariate heterogeneous biomedical time-series. The scenarios of use are described with the corresponding UML Use Case Diagrams. We also discuss some architectural and technological issues, including parallelization, visualization and feature extraction from biomedical time-series. The goal of this paper is to present what we currently consider as the best approach for design of such a system, which may also be beneficial for similar biomedical software systems. The paper is focused on design and architectural considerations only, as implementation of the complex system has only just begun.

I. INTRODUCTION

Web and mobile applications development for biomedical services is continuously growing in the healthcare community [1,2]. While the developed software aims either at general population users [2] or at medical professionals for continuous monitoring of patients [3], there is very little interest involved in developing an integrative web platform that could benefit medical professionals, researchers and experienced general users in modeling subjects' states and disorders.

The goal of the currently running Croatian Science Foundation research project HRZZ-MULTISAB[1] is to develop a high quality system for an all-encompassing analysis of multivariate heterogeneous biomedical time-series. It is conceived that this system would be used by all interested users, including medical doctors, biomedical engineers, computer scientists, and others, depending on their goal. As the system will be developed as a web platform, the users will be able to access it from afar, which differentiates it from local or hospital-specific software solutions.

The problem that the project tackles at its core is the hard problem of efficient biomedical time-series features identification. There have been many efforts put forth to discover and describe both domain-specific and general time-series relevant features [4-6]. The platform that we are currently in the process of developing will offer significant advancements in this respect, as we will consider the use of time-series features there were proven to be effective in a wide range of sciences [6], as well as in specific biomedical signal domains [7].

Once the platform's implementation is completed, we plan to investigate and compare the best feature combinations advocated by domain experts with the best general time-series features and with no-features engineering approach using deep learning approaches [8]. The goal is to achieve as accurate as possible models of subject's states, including various medical disorders.

Presently, in this paper, the goal is to describe system requirements for such a web platform and to elucidate the way in which the platform will be built. We focus on the requirements, with a brief overview of the system's architecture and the technologies involved.

II. SYSTEM REQUIREMENTS

System requirements, which were agreed upon by the entire project working group were the following:

- An integrative software solution for the analysis of multivariate heterogeneous biomedical time-series.
- Software solution implemented in the form of a web platform, as thin client - fat server.
- Software logic layer on the server written in Java programming language.
- Interface towards the user implemented with a set of contemporary web development technologies (HTML5, CSS3, TypeScript...).
- Multiple input file formats: European data format (EDF) and EDF+, textual format for signals and annotations, images formats. Files that contain meta-data (anamnesis, disorder annotations, etc.)
- Visualization of signals in 2D (records inspection) and specific body disorders in 3D using graphical hardware.
- Biomedical time-series preprocessing, such as signal filtering, R peak detection in ECG series, data transformations in: time, frequency, and time-frequency domains.
- Biomedical time-series feature extraction - features would be chosen by: 1) a medical expert system implemented in the platform, which would be based on current medical knowledge from guidelines and

[1] This work has been fully supported by the Croatian Science Foundation under the project number UIP-2014-09-6889.

relevant scientific papers, 2) an expert user, manually; a large number of features need to be supported by the platform, both general and domain-specific biomedical time-series features.

- Feature selection methods, classification, regression, and prediction machine learning algorithms should be used to construct accurate subject state models.
- Results reporting in contemporary formats.

The platform will necessary support the analysis of the following biomedical time-series:

- ECG – electrocardiogram – cardiovascular activity
- HRV – heart rate variability – a series of RR-intervals whose variability is analyzed
- EEG – electroencephalogram – brain activity
- EMG – electromyogram – muscle activity

Additionally, optionally, the platform will support:

- electrooculogram (EOG), electroglottogram (EGG), cardiotocogram (CTG), electrocorticogram (ECoG), photoplethysmogram (PPG), pressure (arterial blood pressure - ABP, pulmonary artery pressure - PAP, central venous pressure - CVP), respiration (impedance), oxygen saturation (SpO2), CO2, gait rhythm, galvanic skin resistance, etc.

The platform will support multivariate heterogeneous analysis (e.g. ECG + HRV + SpO2).

In the implementation phase, the focus will be primarily to demonstrate the working capabilities of the platform on a smaller set of signals and medical disorders.

III. UML Use Case Diagrams for System Requirements Visualization

System requirements have been detailed in the form of UML Use Case Diagrams, v.1.4+ [9], drawn in Astah Community edition tool [10], which present how a user would interact with the system. These diagrams provide a behavioral description of a system's use, without insight into a detailed temporal sequence in which the interaction with the system is performed. Regardless of the absence of a temporal view on the interaction process, all the scenarios of use can be easily identified. This facilitates web platform development.

There were eight UML Use Case Diagrams that we identified, which describe the phases of the analysis:

1. Analysis type selection
2. Scenario selection
3. Input data selection
4. Records inspection
5. Records preprocessing
6. Feature extraction
7. Model construction
8. Reporting

Some of these phases may be skipped entirely, depending on the user.

Additionally, the 9. diagram, User account, depicts registration, login and profile management. Finally, the 10. diagram, Platform administration, shows the administrative access to the web platform.

In the subsequent subsections, we present the diagrams and their detailed explanation.

A. *Analysis Type Selection*

Analysis type selection, Fig. 1, is conducted in a way that a user chooses the goal for an analysis that he wants to perform: detection, classification, prediction, or inspection and visualization. He also chooses the data type on which the analysis will be based. The data type may include exclusively biomedical time-series or biomedical time-series with additional domain data (e.g. subject anamnesis, metadata which describes record type, etc.).

Biomedical time-series may be univariate by type, which means that there is only a single measured data array available in the record, e.g. the times of heart beats, or multivariate, which means that there are several measured data arrays present in the record. Additionally, multivariate data may he homogeneous, which means that they come from the same source measurement device, e.g. EEG data, or heterogeneous, which means that there are more than one measurement device involved at the same time, e.g. ECG + HRV + systolic ABP. An example of analysis type selection would be classification based on EEG data.

B. *Scenario Selection*

Scenario selection, Fig. 2, enables the user to select a predefined scenario, based on previously selected analysis type. It also allows the construction of a completely new scenario through full customization. A list of predefined analysis scenarios with all of the corresponding aspects (used components) will be provided to the user. The user will be able to confirm a scenario or change some of its aspects, such as preprocessing methods, feature extraction methods or model construction methods. When defining a

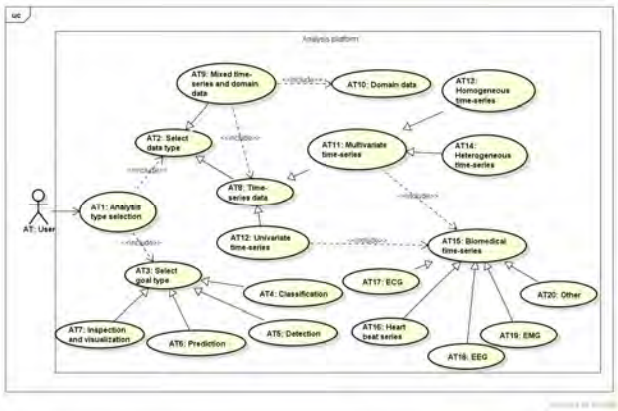


Figure 1. Analysis type selection.

new custom scenario, or changing an existing scenario, the user will be able to select any of the methods for accomplishing his goals. The predefined scenarios will contain features that are to be extracted from signals, which will be selected by the medical expert system specifically designed for this purpose. The medical expert system will be implemented based on current relevant medical guidelines, standards, and relevant medical and biomedical engineering literature. An example of a scenario would be to detect epilepsy in EEG, using wavelet transform modulus maxima and correlation dimension features, and a C4.5 decision tree.

C. *Input Data Selection*

Input data selection, Fig. 3, includes the choice of file extensions (or file formats) that will be uploaded for the analysis. These formats include EDF, EDF+ (with or without annotations), separate textual annotation files (e.g. heart beat annotations), textual raw signal files (without annotations) and image files (e.g. JPEG, TIFF). The user will have to select the appropriate file format in order for file upload to work without errors. Aside from selecting the format, the user also specifies the files that will be uploaded or selects the folder from which all the required files can be found. If there are some specific details related to records loading into platform (e.g. only partial file loading, limitations for some image formats, etc.), these can also be specified during this phase.

D. *Records Inspection*

After the records are uploaded into the platform, it will be possible to inspect and visualize (Fig. 4) each loaded record through the platform's graphical user interface. Here, the user will be able to: select specific record segment, select temporal and amplitude scaling of the signals, display some or all of the signal trails (arrays), inspect record headings and domain data (if such information exists). 3D visualization will be enabled for certain specific states for some biomedical signals (e.g. visualization of myocardial infarction location based on ECG). Additionally, the user will be able to annotate some of the available signals or change the existing annotation files and save these changes.

E. *Records Preprocessing*

Records preprocessing, Fig. 5, assumes various procedures for signal filtering and data transformations. Signal filtering includes noise filtering (e.g. notch filters for 50 / 60 Hz), baseline wandering corrections (e.g. for ECG because of breathing, body movements or electrodes impedance changes), records segmentation into windows of certain width. Window width may be chosen by the user or it may be implicit, determined by the very features that are to be extracted (e.g. a single RR interval for detailed ECG analysis). Other filtering methods include various linear and nonlinear filtering procedures with the goal to obtain higher quality signals.

Data transformations will be performed after the filtering procedures and will include various time (e.g. principal component analysis - PCA), frequency (e.g. fast Fourier transform, Hilbert Huang transform), time-frequency (e.g. wavelet transform) and other types of data

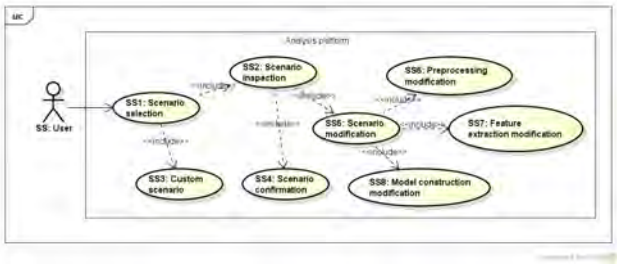


Figure 2. Scenario selection.

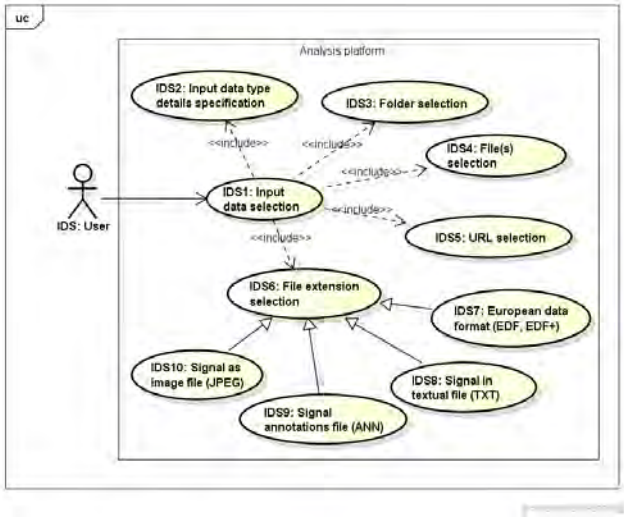


Figure 3. Input data selection.

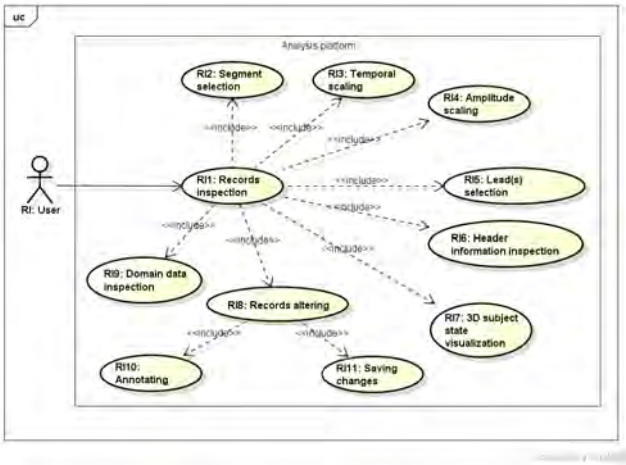


Figure 4. Records inspection.

transformations. The goal of data transformations is to obtain the data in a form from which it is easier to calculate precise domain or general features for describing specific subjects' states.

F. *Feature Extraction*

Feature extraction, Fig. 6, is the central step in the analysis of biomedical time-series. The user will already be provided with a list of features in the scenario selection phase. Regardless of the scenario, in the feature extraction phase, it will be possible so select additional features and

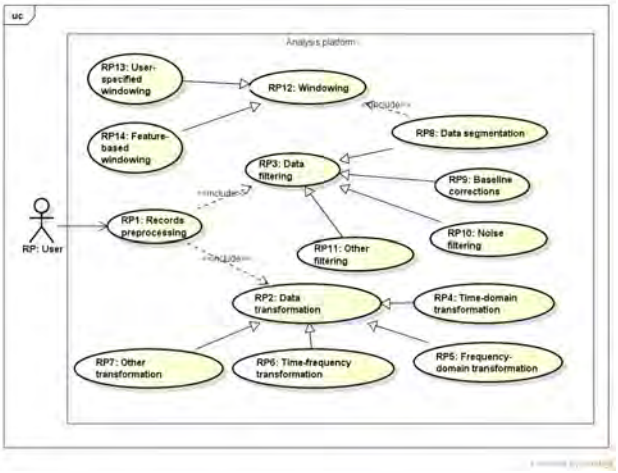


Figure 5. Records preprocessing.

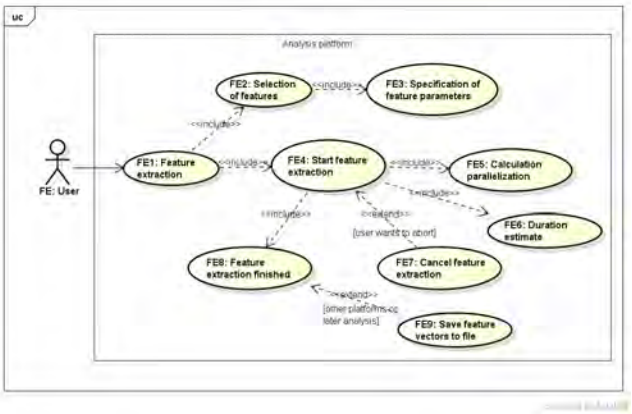


Figure 6. Feature extraction.

specify parameters for calculation of a particular feature, for those features that are parametric. When starting feature extraction, the user will be able to inspect the information about the estimation of the analysis duration and will be able to modify the type of calculation parallelization that will be performed. It will be possible to cancel feature extraction execution at any time. The user will be allowed to save the extracted feature vectors in a file for future analysis, if he wants to have the intermediate results recorded.

G. Model Construction

Model construction, Fig. 7, is a complex step in the analysis of biomedical time-series that takes place after feature extraction and includes various machine learning algorithms for the analysis of the extracted feature vectors. In this step, the user will be able to load the file with feature vectors that were calculated and saved earlier and resume the analysis or continue the analysis from the recently obtained and unrecorded extracted feature vectors. Model construction phase starts with the option to select feature vectors' preprocessing methods. Herein, one can use feature vectors' dimensionality reduction methods such as various transformations (e.g. PCA) or feature selection methods (filters, wrappers, embedded, and other methods [11]).

Dimensionality reduction is necessary in the case where a large number of features were extracted (either by the choice of the expert system or by the user's selection). The goal of dimensionality reduction is to keep only relevant and non-redundant features in order to improve speed and accuracy of model construction. Aside from feature space dimensionality reduction, it will be possible to use some methods to alter the feature vectors themselves (e.g. re-sampling, missing values replacement, etc.).

After the feature vectors preprocessing step, the user will be able to select a model construction method. During this step, first a method is selected and then, the method parameters are defined. Method parameters selection will be made possible if the modeling method is parametric (e.g. C4.5 decision tree, random forest).

Before starting model construction, it is necessary to specify the learning method that will be used: holdout

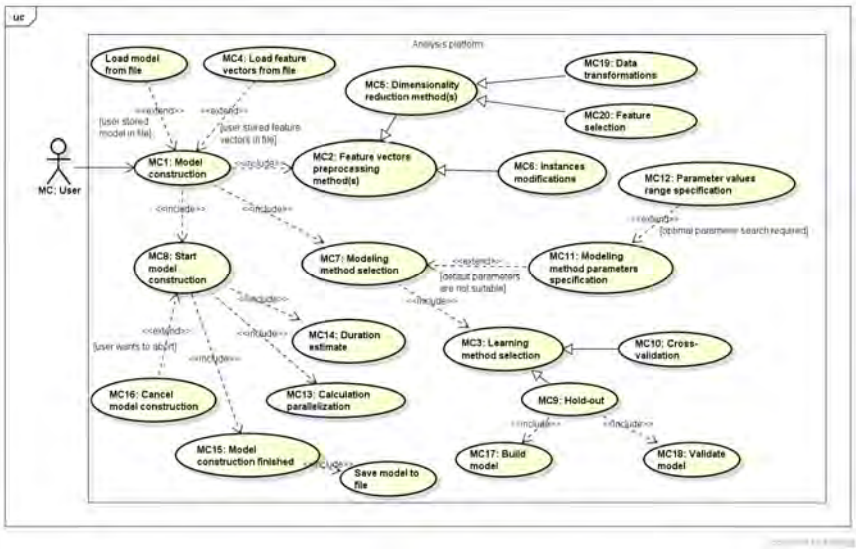


Figure 7. Model construction.

procedure, where a part of dataset is used training and the other part for testing, or cross-validation, where testing is performed on dataset segments so that all the dataset is eventually used for testing. It will also be possible to perform only model testing on a set of new samples, if there is a model present in the system that was trained earlier on the same feature set. When starting model construction, the user will be able to inspect the information regarding the estimation of the construction duration and will be able to modify the type of computing parallelization that will be performed. The construction procedure could be cancelled at any time. After model construction, it will be made possible to save the trained model into a file for later use.

H. Reporting

Reporting, depicted in Fig. 8, is the step in which a user chooses the way in which the results of the analysis are displayed. Thereafter, the report is displayed in the selected form (on a web page only, in PDF, MS Excel, Word or ODF form). It will be possible to save the copy of the report onto the client's computer, whereas its copy will be stored on the server. Additionally, the user will be able to grant access to the report to other registered users, at his convenience.

I. User Account

An account for a new user will be opened by registering the user onto the platform, which will include entering the data about the user (necessary data: user name, password and e-mail address; additional data: first and last name, title, affiliation, affiliation's address, telephone), Fig. 9. Logging into the system will be allowed after confirmation that the registration was accepted. Profile management will be enabled after a successful login. It will include editing personal information, accessing generated (finished) reports that are linked to the user or resuming the last analysis conducted by the user.

J. Platform Administration

Platform administrator will have to login into the platform in the same way as the user, by entering his user name and password, Fig. 10. The administrator will have at his disposal a number of options. First, he will be able to edit his personal profile. He will also be able to manage user accounts for all the platform's users. This will include opening a new user account (as if a user has registered), confirming the registration of a new user, and removing an existing user account. For each account, administrator will be able to inspect the user's personal information (except password), without the possibility of modification.

The administrator will be able to inspect and search all of the reports generated in the platform. Also, he will be allowed access to all of the other files stored in the system, including all the saved input files, all the feature vectors files, and all the constructed models file. He will be able to delete them at his convenience. Moreover, the administrator will have access to all platform's logs, which will be sorted by date. The logs will contain relevant information about the users' actions during the use of the

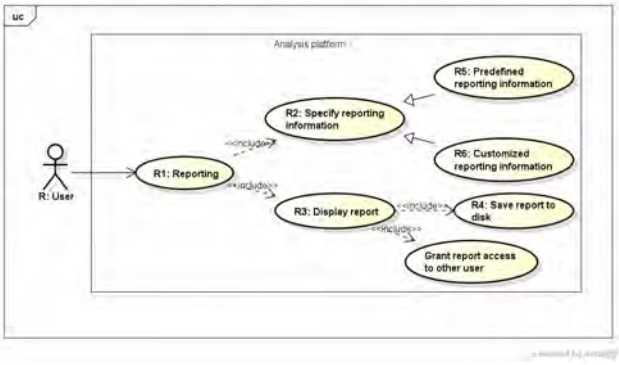


Figure 8. Reporting.

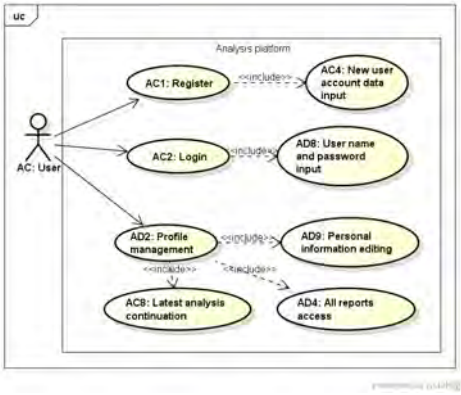


Figure 9. User account.

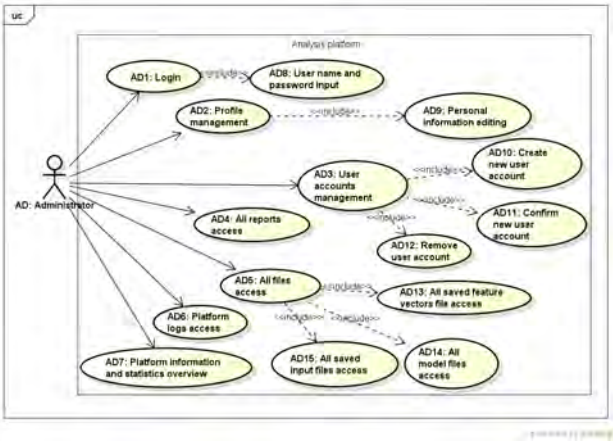


Figure 10. Platform administration.

platform. Additional information about the platform as well as its usage statistics will also be at his disposal.

IV. PLATFORM ARCHITECTURE

A. Web Technologies

System architecture was envisioned as a web portal, which will be accessed by the users through their browsers. This type of architecture was selected, as it is necessary to enable access to the platform from any remote location. With this, the range of potential users is widened, which is important for platform recognition, acceptance, and breadth of applications.

On the client hand side, during platform development, HTML5, CSS3, and Typescript languages will be used for designing web pages. As aids in developing web pages, Angular 2 framework will be used for programming and Bootstrap framework will be used for better visual experience and platform responsiveness. Also, WebGL technology will be used for 3D visualization of human body sections that are of interest to the user.

On the server hand side, Java programming language will be used. This language was chosen because of a large number of existing signal processing libraries, data parsing, implemented machine learning algorithms, ease of web development, and efficient parallelization support. Server side will communicate with the client side through the RESTful protocol. In order to lower the requirements on the server's resources, we will test the use of 1) a standard Java application server with a minimum set of options, and 2) a stand-alone framework that supports the required functionality (e.g. Spring Boot). A minimum set of options will include a user authentication library and an object relational mapping library (e.g. Hibernate).

For reporting, JasperReports® Library, an open-source Java library that supports all the required formats (including HTML) will be used.

B. Architectural Details

The developed architecture should be both modular and scalable. Its modularity should enable the development of various add-ons, while its scalability should support the concurrent work of many users. The platform should also support two important aspects of contemporary software: 1) parallelization and 2) dynamic module loading.

Parallelization will be used to accelerate calculations, which is especially efficient for the algorithms that can be divided into independent execution sections, without the need for synchronization. Parallelization will be enabled in two phases of the analysis: for feature extraction and for model construction, because these phase are the most computationally demanding. We may also consider employing it for records preprocessing, if deemed necessary. In any case, the goal will be to maximize the resource capacity on the server with respect to the number of available CPU cores, as well as employing general purpose GPU parallelization. GPU parallelization will be achieved using JCuda or jocl frameworks, depending on the available hardware. The platform will evaluate the available hardware and the number of running processes in order to enable maximum support for a user's demands.

Dynamic module loading will be a useful feature of the platform, which is significant from the perspective of server resources' optimization. If, for example, a user wants to classify ECGs using C4.5 decision trees, without any additional algorithms, then there will be no need to load all the other platform modules into memory. The modules will be implemented so that they will have maximum cohesion of services within the module, while retaining minimum coupling between other similar modules [12]. This will be enabled through an early definition of analysis scenario (see section III. *B*). The exact sequence of module execution to achieve the required goal could be determined. Auxiliary packages would contain only signatures (name, parameters) of the available components for a specific analysis phase.

The platform will contain a large number of biomedical time-series features, both general and domain-specific, i.e. related to a specific type of time-series (e.g. EEG). In the platform, we plan to use the features from HRVFrame [5], EEGFrame [13] and Comp-Engine [6] frameworks, with appropriate modifications, language translations and testing. Additionally, other domain-specific frameworks, such as the one for ECG analysis will have to be implemented mostly from scratch.

V. CONCLUSION

We have shown the requirements and architecture for an integrative biomedical time-series analysis web platform. Future work will involve implementing the requirements and reporting on the achieved modeling results of the subjects' states.

REFERENCES

[1] J. Oster, J. Behar, R. Colloca, Q. Li, Q. Li, and G. D. Clifford, "Open source Java-based ECG analysis software and Android app for atrial fibrillation screening," in: Computing in Cardiology Conference (CinC) 2013, IEEE, pp. 731–734, 2013.

[2] A. Szczepanski and K. Saeed, "A Mobile Device System for Early Warning of ECG Anomalies," Sensors, vol. 14, pp. 11031–11044, 2014.

[3] G. Paliwal, A. W. Kiwelekar, "A Product Line Architecture for Mobile Patient Monitoring System," in: Mobile Health, Vol. 5 of Springer Series in Bio-/Neuroinformatics, pp. 489–511, 2015, doi: 10.1007/978-3-319-12817-7_22.

[4] A. Bravi, A. Longtin, and A. J. E. Seely, "Review and classification of variability analysis techniques with clinical applications," BioMed. Eng. OnLine, vol. 10, p. 90, Oct. 2011.

[5] A. Jovic, N. Bogunovic, and M. Cupic, "Extension and Detailed Overview of the HRVFrame Framework for Heart Rate Variability Analysis," in: Proceedings of the Eurocon 2013 Conference, I. Kuzle, T. Capuder, and H. Pandzic, Eds. Zagreb: IEEE Press, pp. 1757–1763, 2013.

[6] B. D. Fulcher, M. A. Little, and N. S. Jones, "Highly comparative time-series analysis: the empirical structure of time series and their methods," J. Roy. Soc. Interface, vol. 10, p. 20130048, April 2013.

[7] A. Jovic and N. Bogunovic, "Evaluating and Comparing Performance of Feature Combinations of Heart Rate Variability Measures for Cardiac Rhythm Classification," Biomed. Signal Process. Control, vol. 7 no. 3, pp. 245–255, May 2012.

[8] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, Book in preparation for MIT Press, 2016, url: http://goodfeli.github.io/dlbook/.

[9] Object Management Group, "Unified Modeling Language® (UML®) Resource Page", http://www.uml.org/, last accessed on: 2016-02-21.

[10] Change Vision Inc., "Astah, Community Edition," http://astah.net/editions/community, last accessed on: 2016-02-21.

[11] A. Jovic, K. Brkic, N. Bogunovic, "A review of feature selection methods with applications," in: Proceedings of the MIPRO 2015 Conference, P. Biljanovic, Ed., Rijeka: MIPRO Croatian Society, 2015, Rijeka: MIPRO Croatian Society, pp. 1447–1452, 2015.

[12] K. Praditwong, M. Harman, X. Yao, "Software Module Clustering as a Multi-Objective Search Problem," IEEE Trans. Software Eng., vol. 37, no. 2, pp. 264–282, April 2011.

[13] A. Jovic, L. Suc, and N. Bogunovic, "Feature extraction from electroencephalographic records using EEGFrame framework," in: Proceedings of the MIPRO 2013 Conference, P. Biljanović, Ed., Rijeka: MIPRO Croatian Society, pp. 1237–1242, 2013.

Augmented Reality for Substation Automation by Utilizing IEC 61850 Communication

M. Antonijević*, S. Sučić* and H. Keserica *
* Končar-Power Plant and Electric Traction Engineering Inc, Zagreb, Croatia
{miro.antonijevic; stjepan.sucic; hrvoje.keserica}@koncar-ket.hr

Abstract - IEC 61850 standard represents the most commonly used communication technology for new substation automation projects. Despite the fact that IEC 61850 provides a semantic data model and a standardized configuration description, these facts are underutilized in substation automation management today. This is specifically illustrated in the data visualization domain where new technologies such as virtual and augmented reality have reached significant maturity levels and have not been used for IEC 61850 system visualization so far.

In this paper IEC 61850 features have been combined with augmented reality technologies for providing added value visualization capabilities in substation automation domain. The developed prototype demonstrates proof-of-concept solution for regular substation automation checks and maintenance activities.

I. INTRODUCTION

Smart Grid automation has introduced significant novelties in different Smart Grid subsystems including, Distributed Energy Resources (DERs), distribution automation (DA) and substation automation systems (SAS). One of the elementary prerequisites for successful automation is unified communication mechanisms that facilitate subsystem remote monitoring and control. International standard IEC 61850 represents one the of the Smart Grid automation pillars by introducing standardized communication principles and semantic description of controlled systems. Existence of standards-compliant semantics that describe power system process data allow new possibilities for developing unforeseen applications used in SAS domain. An example of these application categories are Augmented Reality (AR) applications that can utilize meta-data provided by IEC 61850 standard and directly access process related information in order to provide added value features for SAS management. This paper analyses possibility of using IEC 61850 standard for developing AR applications used for SAS maintenance activities.

The paper is organized as follows. The following chapter provides an overview of AR technology usage in industrial automation systems. The third chapter describes main features of IEC 61850 standard and AR technologies in order to give an overview of possible use in SAS environment. The fourth chapter describes main functionalities of proposed AR application, its architecture and implementation approach. The fifth chapter provides analysis results of current developed prototype while the last chapter gives a conclusion.

II. AUGMENTED REALITY IN INDUSTRIAL SYSTEMS

Industrial machines have become advanced tools where automation and advanced feedback is made possible through dedicated control computers. The computers allow us to fully or partially automate complex procedures and can help make manual control more secure and precise. Real-time data from the process is available and many parameters can be interactively controlled through the computer interface. Automation can also allow an operator to monitor multiple machines simultaneously, reducing the number of required personnel for a machine pool. Many critical procedures exist, however, that cannot be completely automated. In such cases, the operator might need to be able to visually follow and interactively control parts of the current operation, while simultaneously monitoring numerous rapidly changing parameters. The maturity of AR technologies has allowed its usage in different industrial automation environments.

AR technologies have been successfully applied in manufacturing [1], smart building management [2], automotive and aerospace industries [3]. The AR application in SAS environment have already been used in [4] and [5]. Hoverer, in both articles AR is used for simulation purposes and operator training. Neither of aforementioned articles deals with process related SCADA used for SAS maintenance purposes as proposed in this paper.

III. IEC 61850 AND AUGMENTED REALITY

IEC 61850 [6] is often regarded as just another remote control protocol for electric utilities, but this international standard is more than a set of rules and encodings for data retrieval from field devices. IEC 61850 defines automation architecture requirements for utility subsystems in order to enable communication and semantic interoperability among multi-vendor equipment. Despite being primarily developed for substations [7], IEC 61850 is now extended for wind power plant domain [8], Distributed Energy Resources (DERs) [9] and hydropower plants [10].

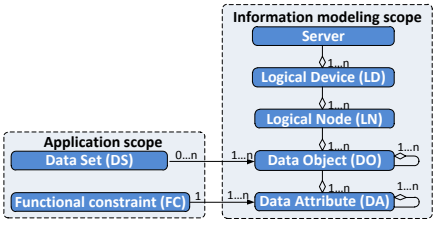


Figure 1 IEC 61850 data class model

A. IEC 61850 basics

1) Data model

Data semantics provided by IEC 61850 are closely related to the functionalities of devices in utility subsystems such as DERs [9]. IEC 61850 data models [11] are based on object-oriented modelling of data relevant to process automation.

Figure 1 shows relationships between IEC 61850 data model classes. The top parent class is the *Server* which represents a physical device, i.e., a device controller.

The *Server* consists of one or more *Logical Devices* (LDs), i.e., virtual representations of devices intended for supervision, protection or control of automated system. LDs are created by combining several *Logical Nodes* (LNs) which represent various device functionalities. LNs are a crucial part of IEC 61850 data semantics.

2) Data exchange

The ACSI is a novel paradigm, introduced by IEC 61850, for describing data exchange procedures in utility subsystems such as substations [11]. ACSI model classes define abstracted information services used for vertical and horizontal communication among IEC 61850 devices. ACSI is not a protocol but a method to tie IEC 61850 abstract services to application layer protocols such as MMS [12].

These ACSI model classes can be used as standardized information interfaces for devices which are realized as IEC 61850 servers. Thus, any IEC 61850 enabled client software can take full advantage of their remote control.

3) Managing and engineering IEC 61850 systems

The engineering process for IEC 61850 systems is based on exchange of XML documents which are formatted according to the *System Configuration description Language* (SCL) [13]. There are several SCL document types depending if they describe device or the integrated system itself. The engineering process based on SCL document exchange is relatively static and most commonly used for substation automation systems where communication system and electric network topology are predefined.

B. Augmented reality – main features

AR is a computer technology that augments the real environment through visually represented information. Thus, to find the opportunities for AR applications in industrial systems it is necessary to review the main AR features to suggest the solutions based on AR. The main AR features are as follows:

- AR can follow the user's viewpoint by means of a tracking system;
- AR can superimpose virtual objects onto the user's view of a real world scene;
- AR can render the combined image of virtual objects and a real world scene in real time;
- AR can locate virtual objects in a real world scene in correct scale, location and orientation.

The control computer often has access to a large amount of process information and we can expect this to increase as the systems become more sophisticated and complex. It is thus important that the data is clearly presented and easily accessible, in order to avoid unnecessarily attention-demanding interfaces and information overload for the user.

Today's control computers typically present their data on a traditional computer display and often use a keyboard/mouse or a touch-screen as input devices. The operator observes the process through the machine's safety glass, while using a computer to the side for control and feedback. This setup results in divided attention if the operator has the need to both follow the procedure visually and simultaneously monitor important values on the computer display. While this problem can in part be addressed through display placement, it inherently separates the process data from the process itself.

Goal of this paper is to provide solution of integrating the process data with the workspace behind the safety glass using AR technology. AR allows combining interactive computer graphics with real objects in a physical environment, such as the workspace of an industrial machine. It is particularly suitable for today's increasingly complex industrial machine processes as it enables intuitive representation and real-time visualization of relevant information in the right place. There are several situations where it might be advantageous to have the capability of annotating the real process with virtual information. An operator might for example want to indicate or emphasize locations inside the workspace behind the safety glass to a co-worker or a student. The type, dimensions and state of the tool currently in operation may be indicated by a virtual label. Process simulation in the real machine using virtual tools and virtual materials could increase safety through virtual previews of the procedure, and provide implicit visual warnings from unintentional geometrical inconsistencies.

1) The ideal augmented reality interface integrates computer graphics with a real environment seamlessly and without encumbering technology. The long tradition of mobile AR systems has required systems based on head mounted displays (HMD) that involve complex tracking and equipment to be worn by the user such as the Oculus Rift. The popularity of video see-through systems can be attributed to the relative ease of implementation and rapid prototyping possibilities provided through various software libraries, such as ARToolkit [14]. Even spatial AR systems like the ASTOR system [15] have been in research and have provided very good results. The majority of AR applications today are marker-based requiring specific markers for 3D tracking and positioning. In this paper QR (Quick Response) [16] has been used as a marker.

2) Augmented reality markers

There are a lot of different types of markers used in AR and although all are applicable in certain situations, one of them, the QR code, is being used a lot more than any of its competitors. Its name is an abbreviation of Quick Response Code and it is a 2D matrix barcode developed by Desno Wave Corporation in 1994 [16], [17].

A QR Code is capable of handling many types of data, such as numeric and alphabetic characters. One pattern can encode up to 7,089 numeric characters or 4,296 alphanumeric characters. Despite a large chunk of data that can be stored in one QR code it can be decoded by a very lightweight smartphone application or a lightweight PC application with camera access, unlike other 2D barcodes which usually require a specific scanner to help decode them.

ARToolkit, the library that is used in this project has its

own markers, but due to QR's superiority it has been decided to use QR codes instead of AR's traditional markers. A QR Code is quite similar to an ARToolkit marker in appearance (Figures 2 and 3), but can encode more information than ARToolkit markers.

The goal is to use QR codes as superior markers and combine them with AR techniques by using QR Codes as traditional ARToolkit markers. Doing the registration process of traditional markers can be omitted therefore simplifying the procedure. Also encoding large amounts of data enables developing more complex and interesting applications.



Figure 2 Fiducial markers



Figure 3 . IEC 61850 QR code

Figure 3 shows an example of a QR code and highlights its parts. A QR code's most important sections are three large square patterns (each contains a small black square with a white border) which are positioned in three corners of the code. They are used to determine the position of the code. Version 2 (and above) expands on this by adding an align square used to align the code after detection. The rest of the code is used to draw a large number of small blocks that encode the data in the code. An additional benefit of this type of encoding is that the code can be reconstructed even if a part of it was damaged or is missing. This is important when scanning the code with a camera because the camera is never going to be positioned perfectly and parts of the code will often be omitted.

There are a couple of correction levels of a QR code which can be seen on Table 1. Higher correction levels allow more of the code to be missing with the drawback of increasing the code size. Superiority of a QR code over other fiducial markers can be seen in Table 2 [18]. These are the arguments for using QR codes in SAS AR applications.

TABLE 1 ERROR CORRECTION LEVEL OF A QR CODE [18]

Level	Percentage of codewords that can be restored (Approximation)
L	7 %
M	15 %
H	25 %
Q	30 %

TABLE II COMPARISON OF QR CODE WITH FIDUCIAL MARKERS [18]

Feature	QR Code	Fiducial Marker
Need to pre-register	No	Yes
Model storing	Internet	Local
Limited number of markers	Larger	Smaller
Universality	Universal barcode	Standalone

IV. IEC 61850 AR APPLICATION

Developing IEC 61850 AR application is based on utilizing IEC 61850 vertical communication, data engineering trough SCL and QR codes as data markers for identifying parts of the SAS equipment according to the IEC 61850 data semantics.

The SCL files provide information about semantically annotated process data. This information is correlated with QR markers assigned to SAS interior (transformers, circuit barkers, switches, feeders, etc.). The QR image recognition recognizes markers, the AG projector draws additional information on the maintenance engineer smartphone screen. The process information overlay SAS equipment directly connecting marked equipment with process information in real time available through SCADA system. By integrating AR the SAS maintenance activities can be significantly simplified allowing unambiguous equipment identification, providing real time measurements and eased SAS monitoring.

A. Application architecture

Application architecture includes several components. The smartphone camera captures the QR code and processes the process data identifier based on IEC 61850. The recognized path represents semantically annotated process data according to IEC 61850 communication standard. This data is correlated with the SCADA process data that retrieves the data from the internal SCADA memory database. The SCADA system gathers data from Intelligent Electronic Device (IED) via vertical IEC 61850 communication. The application architecture is shown in Figure 5.

B. Implementation approach

All substation primary equipment elements have been marked with a QR code which describes the path to the uniquely identified process data according to IEC 61850 standards. Figure 3 shows an example of the QR code for process data representing transformer voltage measurement in phase A (myLD/MMXU1.PhV.phsA.cVal.mag).

The QR code library used for generating QR codes is *qrcode 5.2.2* [19] which is a simple Python library that can generate a QR code with data that you want the code to have. The data is used for identification information for the entity that the QR code is put on (for example the transformer mentioned above). Using a device with a camera, like a smartphone, QR code is detect first the with ARToolkit libraries and process the code's contents. Once it is known

what IEC 61850 process data needs to be displayed the application is connected to a SCADA data exchange interface and ask for the data. The SCADA is then responsible for fetching the data and sending it back to AR application. Once the data has arrived the information is overlaid over the camera feed and the type of overlay will usually depend on the type of data shown. This is also done using ARToolkit libraries for overlaying information. In most cases it will be a simple value and name being overlaid over the marker. After the initial overlay the SCADA is asked to provide real time updates of the data in question as long as the marker is in the camera's vision. This real-time data is provided by the commercial SCADA system PROZA NET. The updates are then used to update the overlaid image and the result is having real-time data being shown on the camera feed as long as it has the marker in its line of sight.

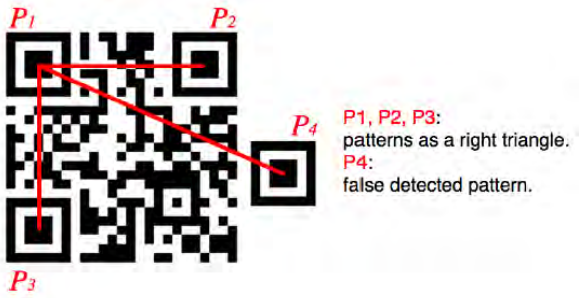


Figure 4 Pattern detection

ARToolkit can calculate the percentage of confidence with which the pattern was recognized. If it is above 50% the pattern is treated as a good candidate for detection. If three candidates are detected in a single frame their special relationship is calculated and those that don't follow the right triangle rule set by QR codes are discarded. Figure 4 shows the special relationships that need to be satisfied.

After determining the three position detection patterns ARToolkit's capabilities can be used to find the fourth corner. Locating the code is a big step in understanding the code and after it has been located in a 3D world it is necessary to decode the information contained in it. A lot of programs used for decoding markers require the user to place the marker exactly in front of the camera to correctly decode the information. Since QR codes are used as ARToolkit markers it cannot be demanded from the user to position the camera in such a way paying attention to the position and the orientation of the camera in regards to the code. Because of that fact QR code must be aligned to facilitate information extraction of the QR code. The code is always projected from one plane on the camera image plane so we can restore its orientation using perspective transformation which we can be written as:

$$\begin{bmatrix} u_i \\ v_i \\ w_i \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}$$

where (x_i, y_i) are original coordinates of the QR code and (u_i, v_i, w_i) are aligned homogenous coordinates of the same code. The rest of the variables can easily be estimated using the coordinates of the four corners of the code.

This perspective transformation can now be used to determine the proper orientation of the code. Now that the QR code is readable by our application it is quite simple to use the information to ask a SCADA system for the data and overlay the data over the QR code using ARToolkit.

V. PROTOTYPE EVALUATION

Evaluation is based on utilizing the developed application prototype that is applied in a real world substation. The preliminary results show that QR codes represent a good selection as markers for IEC 61850 data semantics. Figure 6 shows the application in action.

So far the application is in its infancy, as you can see from the screenshot in Figure 6, but the results are promising enough for us to be confident that it will be highly applicable in real world situations and that the real-time data update will be attractive to end users.

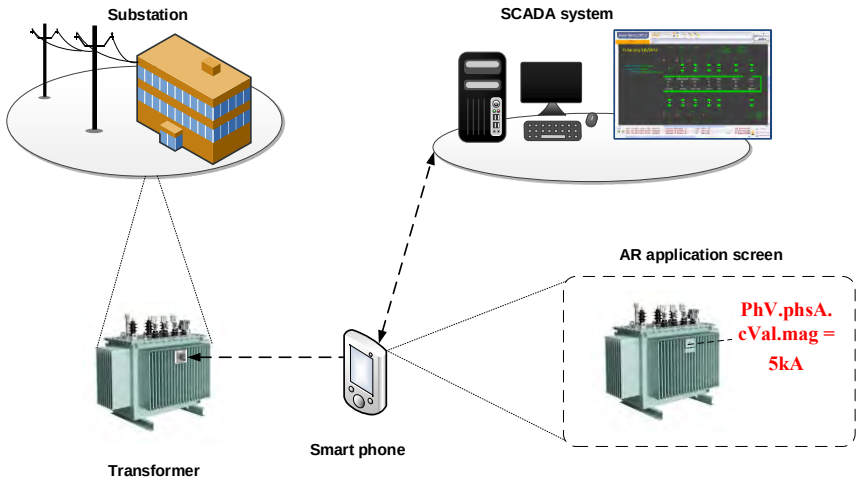


Figure 5 Application architecture



Figure 6 Screenshot of the application in a real world situation

VI. CONCLUSION

In this paper it has been shown how AR technologies can be utilized in SAS systems. For the demonstration purpose, an AR application has been developed based on utilizing QR markers in order to correlate semantically annotated process data with real time SCADA data.

It has been shown that combining standards-based communication and AR technology can provide value-added features for substation maintenance and regular inspection and therefore, reducing the amount of required time and effort for utility maintenance engineers. This paper presents solution applied in SAS environments. However, since IEC 61850 semantics covers several new automation domains it can be easily applied for subsystems such as DER, wond power plants, etc.

REFERENCES

[1] S. K. Ong and A. Y. C. Nee, *Virtual and augmented reality applications in manufacturing*. Springer Science & Business Media, 2013.

[2] L. Romualdo Suzuki, K. Brown, S. Pipes, and J. Ibbotson, Smart building management through augmented reality, in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, 2014, pp. 105–110.

[3] H. Regenbrecht, G. Baratoff, and W. Wilke, Augmented reality projects in the automotive and aerospace industries, *Comput. Graph. Appl. IEEE*, vol. 25, no. 6, pp. 48–56, 2005.

[4] T. R. Ribeiro, P. R. J. dos Reis, G. B. Júnior, A. C. de Paiva, A. C. Silva, I. M. O. Maia, and A. S. Araújo, Agito: Virtual reality environment for power systems substations operators training, in *Augmented and Virtual Reality*, Springer, 2014, pp. 113–123.

[5] P. R. J. dos Reis, D. L. G. Junior, A. S. de Araújo, G. B. Júnior, A. C. Silva, and A. C. de Paiva, Visualization of Power Systems Based on Panoramic Augmented Environments, in *Augmented and Virtual Reality*, Springer, 2014, pp. 175–184.

[6] IEC, Communication Networks and Systems in Substations - ALL PARTS, Int. Std. IEC 61850-SER ed1.0, 2011.

[7] IEC, Communication networks and systems for power utility automation - Part 7-4: Basic communication structure - Compatible logical node classes and data object classes, Int. Std. IEC 61850-7-4, ed2.0, 2010.

[8] IEC, Wind turbines – Part 25-2: Communications for monitoring and control of wind power plants – Information models, Int. Std. IEC 61400-25-2 ed1.0, 2006.

[9] IEC, Communication networks and systems for power utility automation - Part 7-420: Basic communication structure - Distributed energy resources logical nodes, Int. Std. IEC 61850-7-420 ed1.0, 2009.

[10] IEC, Communication Networks and Systems for Power Utility Automation - Part 7-410: Hydroelectric Power Plants - Communication for Monitoring and Control, Int. Std. IEC 61850-7-410 ed1.0, 2007.

[11] IEC, Communication networks and systems for power utility automation - Part 7-2: Basic information and communication structure - Abstract communication service interface (ACSI), Int. Std. IEC 61850-7-2 ed2.0, 2010.

[12] IEC, Communication networks and systems in substations – Part 8-1: Specific Communication Service Mapping (SCSM) – Mappings to MMS (ISO 9506-1 and ISO 9506-2) and to ISO/IEC 8802-3,IEC Std. IEC 61850-8-1 ed1.0, 2004.

[13] IEC, Communication networks and systems for power utility automation - Part 6: Configuration description language for communication in electrical substations related to IEDs, Int. Std. 61850-6, ed2.0, 2009.

[14] Open Source Augmented Reality SDK | ARToolKit.org. [Online]. Available: http://artoolkit.org/.

[15] A. Olwal, J. Gustafsson, and C. Lindfors, Spatial augmented reality on industrial CNC-machines, in *Electronic Imaging 2008*, 2008, pp. 680409–680409–9.

[16] QR code, *Wikipedia, the free encyclopedia*. 21-Feb-2016.

[17] K. Ruan and H. Jeong, An augmented reality system using Qr code as marker in android smartphone, in *Engineering and Technology (S-CET), 2012 Spring Congress on*, 2012, pp. 1–3.

[18] T.-W. Kan, C.-H. Teng, and W.-S. Chou, Applying QR code in augmented reality applications, in *Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry*, 2009, pp. 253–257.

[19] qrcode 5.2.2 : Python Package Index. [Online]. Available: https://pypi.python.org/pypi/qrcode.

Innovation of the Campbell Vision Stimulator with the Use of Tablets

J. Brozek*, M. Jakes ** and V. Svoboda**
* Department of information technologies, University of Pardubice, Pardubice, Czech republic
** Department of software technologies, University of Pardubice, Pardubice, Czech republic
mail@jobro.cz, jakes@asote.cz, svoboda@asote.cz

Abstract - The article covers three fundamental themes: a) performance solutions using gaming to treat multiple eye defects; in particular - Amblyopia; b) an explanation of the issue and design of the software (including games) which is intended for therapeutic or health purposes; and c) highlighting the modern solutions and the power of software products for the needs of the health sector, in particular in the fields of diagnostics and rehabilitation.

The reader can learn basic information about eye diseases and the principles of their treatment, and become acquainted with the reasons why video games are appropriate for rehabilitation.

Very important and beneficial for the reader is the section which focuses on a) the differences in the standard software and the healthcare software, b) the high risks associated with defects of software, or even the risk of side effects with the so-called „perfect software", c) the fact that a major part of software development does not comply with all of the standards.

The article also discusses the advantages of the software solution over other methods of rehabilitation. Most of the paradigms are generally applicable.

Familiarity with the principles of this application can thus be interesting even for developers in the relevant areas.

I. INTRODUCTION

We need to introduce the issue from several different perspectives, such as historical context, scope of the diseases and how to cure them.

A. Context of the issue

There are 500 million people suffering with one of many eye defects at the present time. The current level of human knowledge shows, that 80% of these defects can be treated. Some can be treated surgically, others using conservative methods.

Currently however, the option of conservative treatment (rehabilitation) is dramatically changing. In the past, a rehabilitation option have been severely limited and has had unverifiable results. Modern times, however, brings an ideal device for rehabilitation of eye defects – the tablet. The combination of the intuitive touch input, directly on the visualization unit appears to be ideal; but the level of effectiveness, however, depends on appropriate software.

However, the problem occurs right at the time of the design and implementation of the software. Solutions for health care cannot be arranged as a standard software, because the risks are too high, and the consequences could be disastrous. Therefore, the design of the software (including games) must be executed exceptionally well and according to valid standards.

B. Amblyopia

Our solution primarily targets Amblyopia, also known as the lazy eye disease. According to various sources, 3-4% of the people suffer from this disease. The disease almost exclusively affects only one eye. It can occur without symptoms, or with some visible symptoms, such as strabismus, closing the affected eye, or fear to close the healthy eye. For more information see [1]; [2], [3] or [4].



Figure 1: Health and Amblyopic vision

The disease occurs in early childhood (during the time of the development of vision). As a result of a physical defect, the brain never learns how to properly use signals from the eye. Over time, the brain can completely suppress signals from the affected eye. So, amblyopia is a mixed disorder of eye and brain. According to the current level of knowledge, treatment options are possible only until about 12 years of age, in a period when the brain is still able to get basic sensory perception. However, it should be noted that some studies suggest that the disease can be partially corrected in later age. Treatment principles for Amblyopia are explained in the one of the following chapters.

C. Binocular cooperation

Binocular stereoscopic vision (or cooperation), is absolutely necessary for people in the modern world. After the composition of picture from both eyes, the brain is able to project a third dimension. It means, that when only using one eye, it is not possible to define a distance (warning; trying to emulate this state by short time covering one eye does not work due to the experience of the past, and previous knowledge of sizes and distances).

Fair binocular cooperation is essential not only for activities, such as cycling, or driving the vehicle, but also for simple hand-eye coordination.

The problem with the lack of binocular cooperation arises generally because of: 1) cooperation never being learned, or 2) gradually going blind in one eye. It is possible to say that an age limit for healing Amblyopia is set primarily by the period, when it is possible to teach the brain binocular cooperation of both eyes. Rehabilitation will focus on efforts to practice cooperation; for example hand-eye coordination for which stereoscopic vision is necessary. (E.g. threading beads on a thread.)

For more information see [5] or [6].

II. CURENT THREATMENT METHODS

Amblyopia is a combined (the eye-brain) disease, so then the treatment must respect that. As a first step it is necessary to perform a check of the eye and, if appropriate, remove or correct its physical inadequacy. This correction rarely needs intervention, and most patients need only glasses. The second phase of treatment lies in the practical application of conservative methods (sometimes referred as rehabilitation). All conservative methods have a common basis. Before therapy starts, the patient's healthy eye is prevented from seeing, by using a special shield – occluder. This measure, and on the one hand increases the discomfort of the patient, on the other hand, it is a way how to force the use of the affected eye. The brain cannot ignore the signals from the affected eye for long, and moreover must learn to work with the eye. Gradually this leads to sharpened vision.

The later phases of treatment require the alternate involvement of both eyes (to avoid its degeneration). In the third phase of the deployment comes the treatment of binocular cooperation.

The treatment itself can be accommodated in a variety of ways. During the processing of the work we used primarily [7], [8], [9], [10].

A. Conservative

It is completely conservative principle of treatment. It is based in the fact that, in addition to the blinding of a healthy eye no special measures are taken. This is the oldest and least effective type of treatment. It can only be effective in cases of slight damage. Its big advantage is price. The level of discomfort and treatment duration are major disadvantages.

B. Search in text

One of the old, but still effective methods is considered to be a combination of the blinding of a healthy eye and search in a text (e.g. in the book). The exercise takes place so that the patient simply looks for some particular character. The pace of the workout is determined by the patient. It practices substantial stimulation of vision and the need to select information. The slight drawbacks are: 1) quick eye fatigue, 2) it is not training for various distances, and 3) the exercise requires a high degree of concentration.

C. A conservative with multimedia, or computer using

Modern times allow with a TV, computer or tablet to treat patients by a conservative method, but without the personal risk of being in the physical world. Poor vision can no longer become a cause of accidents and stumbles. Using an occluder and viewing the television or computer reduces risks. On the other hand, narrowing the rehabilitation method restricts the effect of vision stimulation, and loses overlap to motoric and the broader context of eye use (compared with the classic conservative method). In addition, it eliminated the benefit of stimulating the eye for multiple distances.

The biggest problem, however, is the very questionable quality applications (and games) and some phenomena while watching TV. The method seems modern and efficient, but already has a number of disadvantages. It is further extended by the inability to adapt the pace of the patient. There is thus a real risk that patient does not improve, or conversely, that there is a deterioration in the condition of the patient.

D. CAM

Campbell view stimulator, briefly CAM, was developed in the Soviet union round the year 1960 and subsequently was upgraded several times. Actually, it was the first solution that was designed directly by doctors. Because the custom solution presented in this paper is based on the CAM, it will be described in more detail.

CAM is a physical (the most times) electromechanical device that consists of three plates. Far away from the patient is a rotating plate with black and white squares or spirals. Closer to the patient is a transparent plate with the image. The closest part is the glass or foil, on which the patient can draw. The principle of treatment can be likened to a smarter coloring book. These coloring pages are on the moving background. During tracing or rendering the eye is strongly stimulated (even more than is stimulated by the view in the real world). An advantage of the solution is also an implicit connection between treatment of Amblyopia and strengthening the link between vision and motorics. CAM (figure 2) is an elderly solution and current technologies allow its significant evolution.



Figure 2: Campbell vision stimulator

III. OUR SOLUTION

Custom solutions have a historical context and a link to one of the current core developers. The history of evolution illustrated, that the solution is proven. During 10 years there were many steps to the current version and this is first non-medical paper, which explains principles to the IT community.

The development started when the brother of one of the developers began to be treated for Amblyopia. Although the CAM secured the fastest guarantee for success, the old electromechanical CAM was available only in the regional hospital, about 25 km away. The treatment has been complicated by distance, or by the need to keep the patient in the hospital for several weeks (the patient was 5 years old). For this reason the first efforts to create a physical copy of a CAM device started. Unfortunately, the requirements in the area of health care in the combination of electronics and mechanics could not be met. The first generation fails.

The second generation solution came in 2008. The principle of the devices used a standard LCD monitor, through which the glass was installed. On the glass, it was possible to draw and trace. It was a mere evolution of an electromechanical CAM, which in principle was nothing new, it just ensured greater availability of treatment.

The turning point came in the year 2011, i.e. in the period, when tablets in the Czech Republic began to be affordable. The third generation solution has been implemented on a Tablet PC. Dimensions touch control, and availability for individual households changed the face of treatment. The solution was gradually modified and finally in 2013 under the name ANNA (Czech female first name) the technique started to be used in the Czech Republic and neighboring countries. Pardubice Regional Hospital, which has co-worked on the development, still uses this solution. This generation extended the principles with the after-action review system, and "Heal & Play". This version is very popular among child patients.

The last version was released in the year 2015 under the name Anna II, with a version that in addition to the requirements of doctors and patients, also reflects the requirements of the market (special chapter focusses on market needs). The solution is shown at figure 3.



Figure 3: Software solution on tablet

The development of a custom solution is fixed, but for a long time without publishing output. The solution won a few prestigious awards, illustrating that the solution really works. The application itself has gained 1st prize in the Business ideas shows consecutively in the years 2014 and 2015. In 2013, it was awarded a special prize by the Dean (head) of faculty. In 2014 the solution received a special prize from the Rector (head of the University) for outstanding achievement in research and innovation. In 2015, our solution was nominated for the forthcoming research prize given by the National Government of the Czech Republic.

A. The issue of programming for the needs of the health care

Each developer knows or feels intuitively, that the different types of software are subject to various claims; not only in terms of functionality, but also, for example, in testing methodologies, in development methodologies, or even in the restrictions for some of the used algorithms. A demonstration example, made for the need of teaching, or other simple web applications, does not require complex systems design, project management, or structural control. Often enough, the solution works. However, there are also a set of issues, on which human life or health depends. In applications for health care purposes, very strict rules of programming apply. The defect or unexpected behavior of such a program should have disastrous consequences.

It is possible to produce relatively simple programs, which should be developed in a few weeks. However, can you afford to take such a risk? A poorly designed program could endanger the health and lives of users. Such a gamble would be on your conscience if you think only of the financial situation, and utilize incorrect software. Solution without standard is not possible to insure. As the result of the major risks are great demands on the software. During the development, a series of standards must be followed, and the correct development procedures respected (for example, a very often applicable solution is similar to the methodology of the unified process). All must be properly documented and for virtually every step you need to create additional documents. This leads to a

situation, where the (programing) code has thousands of rows and project documentation has a length of ten thousand rows. Additional ten thousand rows then requires the creation of tests and their results.

All of the forms and procedures are standardized. For the design of the software you need to use ISO 90003 [11], for creating health solution ISO13485 [12], ISO testing 17025 [13] and 29119 [14-16], for the proper team management and documentation ISO 21500 [17], 12207 [18] and 9000 [19] and you need to know many others of them.

Software development is so immensely more difficult. Although the two applications may look the same, but the cost of solution witch respects all the standards is ten times more, in comparison to a solution that is not limited by the standards. In addition, it should be noted, that standards may limit the methods of programming, or even restrict the allowed algorithms.

For example, a pragmatic approach teaches the programmers to use algorithms and data structures from the character of their average (often amortized) complexity.

The Splay tree is a data structure with advanced heuristics, the "move to front" operation improves it performances in many classes of application, unlike other forms of search trees. The Quicksort sorting algorithm is often used for its simplicity and speed.

Unfortunately, the standards do not allow the solution, which has a very good average complexity, but the scope focuses on its worst-case complexity (and it is many times not so good in advanced data structures). Therefore, we look for a solution in the selection of algorithms according to the ratio of their performance, and taking into account the worst-case situation.

Methodology for software testing is described very comprehensively in standards. All testing must be conducted against the protocols, and many characteristics of the device, so that the error was virtually eliminated. Each test must be carried out at several levels and must always fulfil the standard conditions of the request. At the same time, it is necessary to test the substandard conditions, influenced by the external environment, and of course the hazards. The software developer must realize (and document) the following types of tests (minimally): unit tests, integration tests, acceptance tests, factory tests, user side tests, integrity tests, tests the stability of the environment. After completion of the test, records of the tests much be archived for a minimum of a program lifetime (standards usually stored 10 years after the closing of the official aid, and the aid must be provided normally at least two years after end of app distribution).

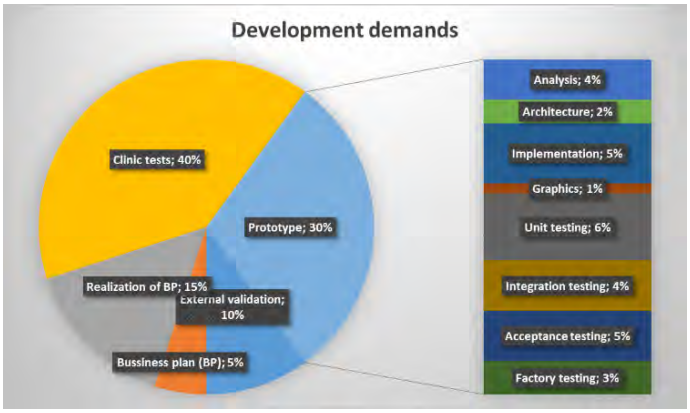


Figure 4: Development demands

Successful completion of effective phases of software tests, from the developer's side, finalizes the development part. The issue does not end there. In the next stage, you should test the software from many other aspects. For example, our solutions have to go through a whole range of display tests. A test of maximum dynamic image changes (high value of dynamic changes of the image would lead to risk of launch epilepsy bout), normal contrast test, specific contrast test (contrast during non-standard lighting) and many others must be measured.

After all of these tests, it is possible, to declare the prototype as ready for first testing on patients. After a further lengthy time, the product is also ready for clinical tests.

B. Base principle

As well as the CAM, our solution is based on active workplace with homogeneously changing background. Every background is changing at the same rate (usually a rotation around a constant Center of rotation), but with different grain size (resolution) of the background. Background with smaller grain sizes (DPI-the underlying shape) creates a simpler solution for the patient. Background with a higher grit is very challenging. Patients with higher degrees of disabilities are not able to work on the environment with a higher grit.

The foreground application varies according to the selected mode. It is possible to choose the mode of the coloring book (thus running mode, which emulates the CAM and brings into the advanced features), or run one of the many games.

The principle is based on the actual painting book. Firstly, patient traces the strokes outline with his finger and thereby draw by selected color. After drawing around the outlines the patient can select the color and with finger holding should fill image with the colors. The maximum therapeutic efficiency is achieved during tracing process. The rendering is only a complementary part of the solution to the patient's picture finished. After he finish, he can Save the picture for future seeing of his progress (of his accuracy) or for printing.

The coloring pages are equipped with several interesting solutions, such as snapping to lines with a scalable tolerance (which brings a higher success rate and

higher motivation), after action review system, and techniques for changing the difficulty of painting.

C. Heal & Play

Heal & Play is a designation used in the Department of software technology of University of Pardubice, where the development is for the health sector using the games. The name is self-explanatory. It is in the process of registration, but there is no further protection. This does not serves as a standard, but for the description of program purpose.

All games are usually very simple and use the same principle with dynamic background as CAM. The aim of the games is not provide a comprehensive gaming experience, but to make the treatment has become a pleasant and popular activities, at least enough to it is possible.

The benefits of the games for the whole solution are clear; the treatment is becoming, to be popular way of time spending. Thanks to a greater spectrum of the games is to ensure that the player does not start to be bored with them. The determination of the appropriate level of difficulty, length and scaling of variables that affect the effectiveness of the treatment, are currently the most effective solution for the rehabilitation equipment in this category.

IV. Discussion

Successive experiments, interviews with physicians in the development of cooperation, or from simple feedback from users, it is possible to characterize some of the profiling features of our solution.

A. The availability of the solution

Free interdisciplinary synthesis of proven principles of the CAM with the programming techniques for Tablet, delivers the benefits of fantastic availability. The tablets are currently expanding strongly in most households, where they are relatively cheap to buy. In addition, the advantage of the use of tablets is that they have a very intuitive, and easily controlled by most people.

B. Mobility

Unlike the original CAM, a great advantage of our solution is united mobility. Treatment can take place by default; in a treatment facility, or in the comfort of your home, or even while traveling. Thanks to this property there has been a dramatic increase in the satisfaction of the treatment. Especially for small children, it is important that the principles of comfort and safety at home are adhered to.

C. The risk of the solution

Each tablet is an electronic bench-mounted unit, which is certified for its operation, so it is safe to use in hospitals and homes.

The risks associated with software are eliminated thanks to the creation of a solution in accordance with all the standards, thanks to the very high quality of testing.

D. Snapping

The software solution enables you to deploy the routines that significantly facilitate the use of the software, or allow you to scale the difficulty of the individual treatments. An example might be: when a child is asked to trace an outline. The software allows you to set, for example, a 0.5 cm tolerance such that the traced line will perfectly to snap onto the intended target. In a paper book, or CAM, such implementation is not possible. This approach has many advantages. The patient feels a greater degree of progress, the overall appearance of the coloring book is better and has easier Diagnostics. Tolerance, however, should be used with caution, so that the error is progressively reduced, therefore, with a view to enhancing the enforcement of claims on the patient.

E. Fun

Implementation of games into a certified therapeutic device has produced fantastic results. The children, on which the software is targeted itself, began take a treatment as fun.

They themselves repeatedly asked for healing cycle. It removes the problem of kids refusing treatment, because it is fun. It is another benefits of our solution, because it could combine combine the high efficiency of treatment with great fun ratio.

Most of the games with this principle are very simple, to do not disturb the heal principle. But they are set, to increase their difficulty. Patient can very easy not succeed when playing games. Just make the games challenging is more important than their high-quality graphic design (graphics design with full textures could very easily degrade the therapeutic effect of the treatment).

F. The secondary effects of treatment

Using the tablet as an I/O device and display unit, creates certain benefits for the user. Unlike other solutions (e.g. almost all conservative with the use of information technology) already used during the treatment of amblyopia, occurs spontaneously with the strengthening of ties between the vision and the motoric abilities. The binding of the eye and the hand is created in such a pleasant manner that may lead to skiping of some specialized rehabilitation treatments.

G. Facilitate the evaluation of the impact of treatment

The after action review system is very convenient for patients and doctors. By storing statistics can be very closely monitored the development of treatment for every individual patients. Thanks to this you can set them the appropriate therapeutic schemes, taking into account the needs of each individual patient. The system is also running on separate devices. These are used with long-term medical supervision. In this case, there are automatic processes, which recommend optimal settings according to patient degree of disability.

V. Conclusion

The article introduced the principles that are used to treat Amblyopia. The paper, however, tried to capture enough of the software development process for health care, so that it could be most beneficial for the reader. The

reader should obtain information on the issue of the development of solutions for the health sector, the demanded steps and also get some inspiration for their own solution.

The entire article is trying to show that through the games on the Tablet you can achieve fantastic results and improve the lives of thousands of people. However, the development of the application, which appears to be relatively simple, is extremely challenging, because it is not easy to meet all the standards. The complexity, however, would not deter developers; since the standards are built so as to protect users and in fact even the developers.

ACKNOWLEDGMENT

The paper submission is particularly financed from Students grant prize at university of Pardubice.

REFERENCES

[1] A. Fielder, Amblyopia: a multidisciplinary approach. Oxford: Butterworth-Heinemann, 2001.

[2] K. J. Ciuffreda, D. M. Levi and A. Selenow, Amblyopia: basic and clinical aspects. Boston: Butterworth-Heinemann, 1991.

[3] J. Osman, Amblyopia. Bolinas, CA: Avenue B, 1993.

[4] M. Schapero, Amblyopia, 1st ed., Philadelphia: Chilton Book Co, 1971.

[5] D. J. Getz, Strabismus and amblyopia, CA: Optometric Extension Program, 1990.

[6] P. H. Spigel, Handbook of Pediatric Strabismus and Amblyopia. New York: Springer Science Business Media, Inc, 2006.

[7] S. W. Ambler and L. L. Constantine, The Unified process transition and production phase: best practices in implementing the UP. Lawrence, Kan.: CMP Books, Masters collection (Lawrence, Kan.), 2002.

[8] R. Cimler, J. Matyska and V. Sobeslav, Cloud based solution for mobile healthcare application. In: Proceedings of the 18thInternational Database Engineering. New York, New York, USA: ACM Press, 2014.

[9] G. Divisova, Strabismus. 2.,. Praha: Avicenum, 1990.

[10] P. Kroll and P. Kruchten, The rational unified process made easy: a practitioner's guide to the RUP. Boston: Addison-Wesley, c2003.

[11] ISO/IEC 90003:2014, Software engineering -- Guidelines for the application of ISO 9001:2008 to computer software

[12] ISO 13485:2003, Medical devices -- Quality management systems -- Requirements for regulatory purposes

[13] ISO/IEC 17025:2005, General requirements for the competence of testing and calibration laboratories

[14] ISO/IEC/IEEE 29119-1:2013, Software and systems engineering – Software testing -- Part 1: Concepts and definitions

[15] ISO/IEC/IEEE 29119-2:2013, Software and systems engineering – Software testing -- Part 2: Test processes

[16] ISO/IEC/IEEE 29119-3:2013, Software and systems engineering – Software testing -- Part 3: Test documentation

[17] ISO 21500:2012, Guidance on project management

[18] ISO/IEC 12207:2008, Systems and software engineering -- Software life cycle processes

[19] ISO 9000:2005, Quality management systems -- Fundamentals and vocabulary

.

Classification of Scientific Workflows Based on Reproducibility Analysis

A. Bánáti[1], P. Kacsuk[2,3] and M. Kozlovszky[1,2]
[1] Óbuda University, John von Neumann Faculty of Informatics, Biotech Lab
Bécsi str. 96/b., H-1034, Budapest, Hungary
[2] MTA SZTAKI, LPDS, Kende str. 13-17, H-1111, Budapest, Hungary
[3] University of Westminster, 115 New Cavendish Street, London W1W 6UW
{banati.anna, kozlovszky.miklos}@nik.uni-obuda.hu,
kacsuk@sztaki.mta.hu

Abstract - In the scientist's community one of the most vital challenges is the reproducibility of a workflow execution. The necessary parameters of the execution (we call them descriptors) can be external which depend on for example the computing infrastructure (grids, clusters and clouds), on third party resources or it can be internal which belong to the code of the workflow such as variables. Consequently, during the process of re-execution these parameters may change or become unavailable and finally they can prevent to reproduce the workflow. However in most cases the lack of the original parameters can be compensated by replacing, evaluating or simulating the value of the descriptors with some extra cost in order to make it reproducible. Our goal in this paper is to classify the scientific workflows based on the method and cost how they can become reproducible.

I. INTRODUCTION

In large computational challenges scientific workflows have emerged as a widely accepted solution for performing in-silico experiments. In general these in-silico experiments consist of series of particularly data and compute intensive jobs, and in most cases their executions require parallel and distributed infrastructure (super/hypercomputers, grids, clusters, clouds).

The successive steps of an experiment are chained to a so called workflow, which can be represented by a directed acyclic graph (DAG). The nodes are so called jobs, which includes the experimental computations based on the input data accessed through their input ports. In addition, these jobs can product output data, which can be forwarded through their output ports to the input port of the next job. The edges of a DAG represent the dataflow between the jobs (Figure 1.).

An essential part of the scientific method is to repeat and reproduce the experiments of other scientist and test the outcomes themselves even in a different execution environment. A scientific workflow is reproducible, if it can be re-executed without failures and gives the same result as the first time. In this approach the failures do not mean the failures of the Scientific Workflow Management System (SWfMS) but the correctness and the availability of the inputs, libraries, variables etc. Different users for different purposes may be interested in reproducing the workflow, for example the

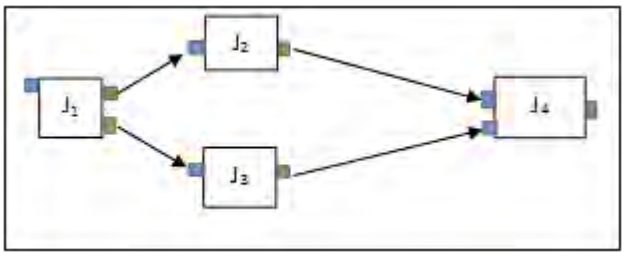


Figure 1. Workflow example with four jobs (J1, J2, J3, J4)

authors of the workflow(s) in order to prove their results, readers or other scientists in order to reuse the results or reviewers in order to verify the correctness of the results [1]. Additionally, nowadays scientific workflow repositories are already available and in this way the scientists can share their results with each other and even they can reuse the existing workflows to create new ones.

The two most significant obstacles of reproducing a workflow are the dependencies of workflow execution and the rich collection of provenance data. The former can be perceived as the necessary and the latter one as the satisfactory requirements of the reproducibility. The dependencies of the execution mean those resources which require external (out of the scientific workflow management system, SWfMS) services or resources such as third party services, special hardwares/softwares or random value generator [2]. Elimination of these dependencies in most cases is not possible, so they have to be handled in some other way: different methods should be set up to make the workflows reproducible.

To achieve our goal we have defined the descriptor space and the decay-parameters of the jobs that give us the possibility to analyze the workflow from a reproducibility perspective. The descriptor space contains all the parameters (call descriptors), which are necessary to reproduce the workflow. There are descriptors, which are constant and do not change in time. Other descriptors are continuously changing (for example a database which continuously get more and more data from sensor networks). Also descriptors based on external services (such as third party services) may exist which can be unavailable after a few years. Finally there are descriptors which are unknown and its behavior is unpredictable. In

this case the workflow is non-reproducible. The decay-parameter describes the type and the measure of the change of the descriptor. With the help of the decay-parameter we have determined five categories of the workflows: reproducible, reproducible with extra cost, approximately reproducible, reproducible with probability *P* and non-reproducible.

The goal of our investigation to find different methods to make reproducible the workflow in the different categories even if it requires extra costs or compromises. In certain cases this goal is implementable but often the result of the workflow is only evaluable with the help of simulations. If there is no method to make the workflows reproducible, our goal is to provide the scientists with useful information about the conditions and probability of the reproducibility of his workflows.

The rest of the paper is organized as follows: In the next section we provide a short background and overview about works related to our research. Section 3 presents the mathematical model of our reproducibility analysis. In section 4 we give the classification of the scientific workflows based on our analysis. In section 5 based on our model we define the general measures of the reproducibility analysis. Finally we summarize our conclusions and reveal the potential future research directions.

II. STATE OF THE ART

Currently the reproducibility of scientific workflows is a burning question which the scientist community has to face with and has to solve. Accordingly in the latter one-two years many researchers investigate this issue. One part of the literature analyzes the requirements of reproducibility and the other part deals with the implementation of such tools or frameworks.

The first group agree on the importance of the careful design [3], [4], [5], [6], [7] which on one hand means the increased robustness of the scientific code, for example with a modular design and detailed description of the workflow, and of the input and output data examples, and consequent annotations [8]. On the other the careful design includes the careful usage of volatile third party or special local services. In these cases two solutions exist, but reproducibility is uninsurable: 1. taking a digital copy of the entire environment using a system virtual machine/hardware virtualization approach capturing and storing metadata about the code and environment that allows it to be recreated later [8].

Zhao et al. [9] in their paper investigate the cause of the so called *workflow decay*, which means that year by year the ability and success of the re-execution of any workflow significantly reduces. They examined 92 Taverna workflows submitted in the period from 2007 to 2012 and found four major causes: 1. Missing volatile third party resources 2. Missing example data 3. Missing execution environment (requirement of special local services) and 4. Insufficient descriptions about workflows. Hettne et al. [10] in their paper list ten best practice to prevent the workflow decay. Grothe et al. [11] analyze the characteristic of applications used by workflows and list the requirements in order to enable the reproducibility of results and determination of provenance. To the former mentioned requirements they assumed the deterministic feature of applications in order to perform appropriate provenance collection.

There exist available tools, VisTrail, ReproZip or PROB [12], [13], [14], which allow the researcher and scientist to create reproducible workflow. With help of VisTrail [12], [15] reproducible paper can be created, which includes not only the description of scientific experiment, but all the links for input data, applications and visualized output which always harmonizes with the actually applied input data, filter or other parameters. ReproZip [13] is another tool, which stitches together the detailed provenance information and the environmental parameters into a self-contained reproducible package.

The Research Object (RO) approach [16], [17] is a new direction in this research field. RO defines an extendable model, which aggregates a number of resources in a core or unit. Namely a workflow template; workflow runs obtained by enacting the workflow template; other artifacts which can be of different kinds; annotations describing the aforementioned elements and their relationships. Accordingly to the RO, the authors in [18] also investigate the requirements of the reproducibility and the required information necessary to achieve it. They created ontologies, which help to uniform these data. These ontologies can help our work and give us a basis to perform our reproducibility analysis and make the workflows reproducible despite their dependencies.

Piccolo et al [19] collected the tools and techniques and proposed six strategies which can help the scientist to create reproducible scientific workflows.

Santana-Perez et al [20] proposed an alternative approach to reproduce scientific workflows which focused on the equipment of a computational experiment. They have developed an infrastructure-aware approach for computational execution environment conservation and reproducibility based on documenting the components of the infrastructure.

To sum up the results mentioned above, we can conclude that the general approach is that the scientist has to create reproducible workflows with careful design, appropriate tools and strategies. But none of them intended to solve the problem related to the dependencies rather they suggested to bypass them. Moreover, they did not deal with the following question: How an existing workflow can be made reproducible?

III. THE MODEL

In our approach a scientific workflow consisted of N jobs can be written as a function of its job:

$$SWf(J_1, J_2, \dots, J_N) = \mathbf{R} \qquad (1)$$

where **R** is the vector of results.

In our investigation we assume, that a given workflow is executed at least one time and the provenance database of the workflow execution is available. In this case we can assign a so called descriptor space to every job of the given workflow.

$$D_{J_i} = \{d_{i1}, d_{i2}, \dots, d_{iK_i}\} \quad (2)$$

The elements of this descriptor space are called descriptors and they give all the necessary parameters to reproduce the job. These parameters can be for example variables of the infrastructure, variables of the code, parameters of system calls, inputs, outputs and partial data or access paths of external resources etc [21]. Every descriptor has a name and a value. In addition we also assign them a so called decay-parameter which describes the type and the measure of the change of the given value. The decay-parameter can be zero, which means that the value of this descriptor is not changing in time, in other word the availability of this descriptor (and its value) can be insured in one, two, ten or any years. In this case this descriptor does not cause dependency and the reproducibility of the job does not depend on this descriptor. The decay parameter can be infinite, if the descriptor's value is unknown. For example in case of random generated values. The value of the decay-parameter can be a distribution function *F(t)* if the availability of the given resource varies in time according to this *F(t)*. The fourth option is that the value of the decay parameter is a function – vary(t, x) – depending on time, which determines the variation of the descriptor's value.

Formally:

$$decay(v_i) = \begin{cases} \mathbf{0}, & \text{if the value of the descriptor is not changing in time} \\ \infty, & \text{if the value of the descriptor is unknown} \\ \boldsymbol{F_i(t)}, & \text{distribution function of the availability of the given value} \\ \boldsymbol{Vary_i(t,\ v_i)}, & \text{if the value of the descriptor is changing in time} \end{cases} \quad (3)$$

The descriptors and its decay parameters can originate from three different sources: from the users, from the provenance database and it can be automatically generated by the SWfMS. [21]

With the help of these expressions we can define the reproducibility in the following way:

Definition: The J_i job is reproducible, if

$$JOB_i\left(t_0, v_{i1}(d_{i1}), v_{i2}(d_{i2}), \dots, v_{iK_i}(d_{iK_i})\right) = \\ = JOB_i\left(t_0 + \Delta t, v_{i1}(d_{i1}), v_{i2}(d_{i2}), \dots, v_{iK_i}(d_{iK_i})\right) = \boldsymbol{R_i} \quad (4)$$

for every Δt.

In addition if a scientific workflow contains N jobs and the jobs are reproducible, the scientific workflow is also reproducible:

$$SWF(t_0, J_1, J_2, \dots, J_N) = SWF(t_0 + \Delta t, J_1, J_2, \dots, J_N) = \mathbf{R} \quad (5)$$

for every Δt.

Also we can assign a cost to the descriptors. This gives the measurement of the "work" or cost which is necessary to make the job reproducible. For example, when the value of the descriptor is a large amount of data which cannot be stored even on extra storage. We can assign a cost to this extra storage. Or another example if the descriptor is changing in time and its decay-parameter is a so called "vary function". In this case to reproduce this workflow we can apply simulation tools based on the sample set which also result an extra cost (see section IV.A).

TABLE 1. THE DESCRIPTOR SPACE OF A JOB AND ITS MEASURES

Descriptor's name	Descriptor's value	Decay-parameter	Cost
d_1	$v_1(d_1)$	$decay(v_1)$	$\boldsymbol{c_1}$
d_2	$v_2(d_2)$	$decay(v_2)$	$\boldsymbol{c_2}$
…	…	…	…
d_K	$v_K(d_K)$	$decay(v_K)$	$\boldsymbol{c_K}$

IV. THE CLASSIFICATION

Analyzing the decay parameters of the descriptors we can classify the scientific workflows. First, we can separate the workflows which decay-parameters for all the jobs are zero. These workflows are reproducible at any time and any circumstance since they do not have dependencies. Than we can determine those ones which can influence the reproducibility of the workflow in other words which also have non-zero decay parameter(s). Four groups have been created:

1. At least one decay-parameter of the descriptor is infinite, but with the help of additional resources or tools this dependency of execution can be eliminated. In this case the cost of this descriptor indicates that there are possibility to reproduce the job with some extra cost.
2. At least one decay-parameter of the descriptor is infinite and the cost of this descriptor is also infinite. In this case the dependency of the workflow can not be eliminated and the workflow is non-reproducible.
3. At least one decay-parameter of the descriptor is a probability distribution function and the other ones are zero.
4. At least one decay-parameter of the descriptor is a vary function and the other ones are zero. (Table 2.)

A. Reproducible workflows

The first group represents the reproducible workflows. In this case all the decay-parameters of all the jobs belonged to a workflow are zero. These workflows are reproducible and they can be executed and re-executed at any time and any circumstance since they are not influenced by dependencies.

B. Reproducible workflow with extra cost

There are workflows, which have dependencies and infinite decay-parameters, but the appropriate cost is not infinite. In this case with the help of additional resources or tools these dependencies can be eliminated. For example, if

a computation is based on random generated value, this descriptor's value is unknown (infinite). In this case with the help of an extra, operation system level tool we can capture the return value of the system call and we can save it in the provenance database [22]. The third example is when a virtualization tool, such as a virtual machine have to be applied to reproduce the workflow.

C. Approximetly reproducible workflows

In certain cases the workflow execution may depend on some continuously changing resource. For example there are continuously growing databases which get the data from sensor networks without intermission. If the computation of a workflow use some statistical parameters of this database, the statistical values never will be the same. In this case the appropriate descriptor's value of the given job may change on occasion of every re-execution, consequently the reproducibility of this workflow could be failed.

If the workflow was executed $\boldsymbol{S}$ times and the provenance database is available, we can create a sample set which contains the $\boldsymbol{S}$ different values of the changing descriptors and the $\boldsymbol{S}$ results of the workflow. In this case we can analyze the change of the descriptor's value, we can write its function and even, we can determine a general evaluating method of the result. On occasion of a later re-execution, if reproducing is not possible, this evaluating method can be applied and an evaluated result can be done with a given probability [22].

D. Reproducible workflows with a given probability

Many investigations revealed the problem caused by volatile third party resources […], when the reproducibility of workflows became uncertain. The third party services or any external resources can be unavailable during the years. If we know this decay of the resources and if we can determine its probability distribution function we can predict the behavior of the workflow on occasion of a re-execution at a later time. Sometime the users may have to know the chance of the reproducibility of their workflow. Assuming that the probability distribution of the third party service is known or assumable we can inform the users about the expected probability of the reproducibility.

To formalize the problem, first, we have separated the M_i descriptors of a given job J_i which depend on external or third party resources and its decay-parameter, which is a probability distribution function given as follows: $F_{i1}(t), F_{i2}(t), \dots, F_{iM_i}(t)$. The rest of the descriptors have zero decay-parameter. In this case, at time t_0, a given descriptor's value $v_{ij}(d_{ij})$ is available with a given probability (for the sake of the easier comprehensibility hereafter we omitted the *i* index referred to the *i*th job of a given scientific workflow):

$$F_1(t_0) = p_1^{(t_0)}, F_2(t_0) = p_2^{(t_0)}, \dots, F_M(t_0) = p_M^{(t_0)} \quad (6)$$

Let us assign to the job J_i a state vector $\mathbf{y_i} = (y_{i1}, y_{i2}, \dots, y_{iM_i}) \in \{0,1\}^{M_i}$, in which the $y_{ij} = 1$, if the *j*th descriptor of the job J_i is unavailable. In this way the probability of a given $\mathbf{y_i}$ state vector can be computed as follows:

$$p(y) = \prod_{j=1}^{M} p_j^{y_j} \left(1 - p_j\right)^{1-y_j} \quad (7)$$

TABLE 2. CLASSIFICATION OF SCIENTIFIC WORKFLOWS

decay-parameter	cost	category
decay(v)=0	*cost = 0*	reproducible
decay(v) = ∞	*cost* = ∞	non-reproducible
decay(v) = ∞	*cost* = C_1	reproducible with extra cost
decay(v) = *F(t)*	*cost* = C_2	reproducible with probability *P*
decay(v) = vary(t,v)	*cost* = C_3	approximately reproducible

In addition a time interval can be given during which the descriptor is available with a given probability *P*.

Since we assume the independency of the descriptors the cumulative distribution function of the job J_i can be written as follows:

$$\mathbf{F}_i(t) = \prod_{j=1}^{M} F_{ij}(t) \quad (8)$$

E. Non-reproducible workflows

There is no method to make the workflow reproducible. In this case the scientific workflow probably contains non-deterministic job or jobs.

V. REPRODUCIBILITY ANALYSIS

It may be important to inform the user about the reproducibility of his workflow or even the cost of the reproducibility. Based on our mathematical model we can determine two measures according to the expected cost: the average cost and the reproducibility probability.

1. Average Cost (AC) expressed as

$$E\big(g(\mathbf{y})\big) = \sum_{y \in Y} g(\mathbf{y}) p(\mathbf{y}) \quad (9)$$

where $g(\mathbf{y}) = \sum_{i=1}^{K} c_i$.

2. Reproducibility Probability (RP)

$$P(g(\mathbf{y}) > C) = \sum_{Y: g(\mathbf{y}) > C} p(\mathbf{y}) \quad (10)$$

where C is a given level of the reproducibility cost.

VI. CONLUSION

In this paper we investigated the possible types of the scientific workflows from a reproducibility perspective. The basis of our analysis is the decay-parameter which describes the type and the measure of the change of the descriptor's values. According to this parameter we determined a cost function which means the "work" required to reproduce the given job or workflow. In this way we could classify the scientific workflows, how they can be reproduced at a later time. In the different categories we set up methods to make the workflows reproducible or we gave the probability and the extra cost of the reproducibility. Finally we gave two general measure to evaluate the expected cost of the reproducibility.

The goal of our research is to support the scientists with methods to make their experiment reproducible and to provide information about the possibility to reproduce their workflows.

REFERENCES

[1] D. Koop, E. Santos, P. Mates, T.Vo Huy, P Bonnet, B. Bauer, M.Troyer, D.N. Williams, J.E. Tohline, J. Freire, C.T. Silva, „A Provenance-Based Infrastructure to Support the Life Cycle of Executable Papers", Internatioonal Conference on Computational Science, ICCS 2011. [Online]. Available: http://www.sciencedirect.com.

[2] A. Banati, P. Kacsuk, M. Kozlovszky, M. Four level provenance support to achieve portable reproducibility of scientific workflows. In Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on (pp. 241-244). IEEE.

[3] P. Missier, S. Woodman, H. Hiden, és P. Watson, „Provenance and data differencing for workflow reproducibility analysis", Concurrency and Computation: Practice and Experience, 2013

[4] R. D. Peng, „Reproducible Research in Computational Science", Science, köt. 334, sz. 6060, o. 1226–1227, dec. 2011

[5] J. P. Mesirov, „Accessible Reproducible Research", *Science*, köt. 327, sz. 5964, o. 415–416, jan. 2010.

[6] D. De Roure, K. Belhajjame, P. Missier, J. M. Gómez-Pérez, R. Palma, J. E. Ruiz, K. Hettne, M. Roos, G. Klyne, C. Goble, és others, „Towards the preservation of scientific workflows", in Procs. of the 8th International Conference on Preservation of Digital Objects (iPRES 2011). ACM, 2011.

[7] S. Woodman, H. Hiden, P. Watson, és P. Missier, „Achieving reproducibility by combining provenance with service and workflow versioning", in Proceedings of the 6th workshop on Workflows in support of large-scale science, 2011, o. 127–136.

[8] A. Davison, „Automated Capture of Experiment Context for Easier Reproducibility in Computational Research", Computing in Science & Engineering, köt. 14, sz. 4, o. 48–56, júl. 2012.

[9] J. Zhao, J. M. Gomez-Perez, K. Belhajjame, G. Klyne, E. Garcia-Cuesta, A. Garrido, K. Hettne, M. Roos, D. De Roure, és C. Goble, „Why workflows break—Understanding and combating decay in Taverna workflows", in E-Science (e-Science), 2012 IEEE 8th International Conference on, 2012, o. 1–9.

[10] K. M. Hettne, K. Wolstencroft, K. Belhajjame, C. A. Goble, E. Mina, H. Dharuri, D. De Roure, L. Verdes-Montenegro, J. Garrido, és M. Roos, „Best Practices for Workflow Design: How to Prevent Workflow Decay.", in SWAT4LS, 2012

[11] P. Groth, E. Deelman, G. Juve, G. Mehta, és B. Berriman, „Pipeline-centric provenance model", in Proceedings of the 4th Workshop on Workflows in Support of Large-Scale Science, 2009, o. 4.

[12] J. Freire, D. Koop, F. S. Chirigati, és C. T. Silva, „Reproducibility Using VisTrails", 2014. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download doi:10.1.1.369.9566

[13] F. S. Chirigati, D. Shasha, és J. Freire, „ReproZip: Using Provenance to Support Computational Reproducibility.", in *TaPP*, 2013.

[14] V. Korolev, A. Joshi, V. Korolev, M. A. Grasso, A. Joshi, M. A. Grasso, D. Dalvi, S. Das, V. Korolev, Y. Yesha, és others, „PROB: A tool for Tracking Provenance and Reproducibility of Big Data Experiments.", *Reproduce'14. HPCA 2014*, köt. 11, o. 264–286, 2014.

[15] D. Koop, J. Freire, és C. T. Silva, „Enabling Reproducible Science with VisTrails", *arXiv preprint arXiv:1309.1784*, 2013.

[16] O. Belhajjame, K. Corcho, D. Garijo, J. Zhao, P. Missier, D. R. Newman, R. Palma, S. Bechhofer, G. C. Esteban, J. M. Gomez-Perez, G. Klyne, K. Page, M. Roos, J. E. Ruiz, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, and C. Goble. Workflow-centric research objects: First class citizens in scholarly discourse. In Proceedings of the ESWC2012 Workshop on the Future of Scholarly Communication in the Semantic Web, 2012

[17] Bechhofer S., De Roure D., Gamble M., Goble C., Buchan I.: Research objects: Towards exchange and reuse of digital knowledge. In: he Future of the Web for Collaborative Science, 2010.

[18] Belhajjame K., Zhao J., Garijo D., Gamble M., Hettne K., Palma R., Goble C.: Using a suite of ontologies for preserving work ow-centric research objects. In: Web Semantics: Science, Services and Agents on the World Wide Web, 2015.

[19] Piccolo S.R., Lee A.B., Frampton M.B.: Tools and techniques for computational reproducibility. In: bioRxiv, vol. 022707, 2015.

[20] Santana-Perez I., Prez-Hernndez M.S.: Towards Reproducibility in Scientific Workflows: An Infrastructure-Based Approach. In: Scientific Programming, vol. 2015, p. 11, 2015.

[21] Banati A., Kacsuk P., Kozlovszky M.: Minimal sufficient information about the scientific workflows to create reproducible experiment. In: IEEE 19th International Conference on Intelligent Engineering systems (INES), Slovakia, 2015.

[22] Banati A., Kacsuk P., Kozlovszky M.: Reproducibility analysis of scientific workflows; Acta Politechnica Hungarica, unpublished

Dynamic Execution of Scientific Workflows in Cloud

E. Kail[1], J. Kovács [2], M. Kozlovszky[1,2] and P. Kacsuk[2,3]
[1] Óbuda University, John von Neumann Faculty of Informatics, Biotech Lab
Bécsi str. 96/b., H-1034, Budapest, Hungary
[2] MTA SZTAKI, LPDS, Kende str. 13-17, H-1111, Budapest, Hungary
[3] University of Westminster, 115 New Cavendish Street, London W1W 6UW
{kail.eszter, kozlovszky.miklos}@nik.uni-obuda.hu,
{jozsef.kovacs, kacsuk}@sztaki.mta.hu

Abstract - Scientific workflows have emerged in the past decade as a new solution for representing complex scientific experiments. Generally, they are data and compute intensive applications and may need high performance computing infrastructures (clusters, grids and cloud) to be executed. Recently, cloud services have gained widespread availability and popularity since their rapid elasticity and resource pooling, which is well suited to the nature of scientific applications that may experience variable demand and eventually spikes in resource. In this paper we investigate dynamic execution capabilities, focused on fault tolerance behavior in the Occopus framework which was developed by SZTAKI and was targeted to provide automatic features for configuring and orchestrating distributed applications (so called virtual infrastructures) on single or multi cloud systems.

I. INTRODUCTION

Over the last few years, cloud computing has emerged as a new model of distributed computing by offering hardware and software resources as virtualization-enabled services. Cloud providers give application owners the option to deploy their application over a network with a virtually infinite resource pool with modest operating and practically no investment costs. Today, cloud computing systems follow a service-driven, layered software architecture model, with Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). In this paper we are primarily focusing on IaaS cloud services. In an IaaS environment the CPU, storage, and network resources are supplied by a collection of data centers installed with hundreds to thousands of physical resources such as cloud servers, storage repositories, and network backbone. It is the task of the cloud orchestrator to select the appropriate resource for an initiated application or service executed in the cloud.

Due to their rapid elasticity and almost infinite resource pooling capabilities cloud services have also gained widespread popularity for enacting scientific experiments. Scientific experiments are widely used in most scientific domains such as bioinformatics, earthquake science, astronomy, etc. In general they consist of multiple computing tasks that can be executed on distributed and parallel infrastructures.

Scientific workflows are used to model these scientific experiments at a high level abstraction. They are graphically represented by Directed Acyclic Graphs (DAGs), where the nodes are the computing tasks and the edges between them represent the data or control flow. Since these experiments mostly require compute and data intensive tasks the execution of scientific workflows may last for even weeks or months, and may manipulate even terabytes of data. Thus scientific workflows should be executed in a dynamic manner in order to save energy and time.

Dynamic execution has three main aspects: fault tolerance, intervention and optimization techniques. Fault tolerance means to continue the execution with the required SLA even in the presence of failures, or to adapt to new situations and actual needs during runtime. Since scientific workflows are mainly explorative by nature scientists often need to monitor the execution, to get feedback about the status of the execution and to interfere with it. Intervention by the scientist, workflow developer or the administrator may also be needed in a planned or in an ad-hoc manner. The third aspect of dynamic execution concerns with optimization mechanisms, such as performance, budget, time or power optimization techniques.

In this paper we are investigating the possibilities of executing scientific workflows dynamically in Occopus, we examine the required extensions to provide a reliable service for workflow orchestration and propose a fault tolerant mechanism which is based on the workflow structure and replication technique.

Occopus [7] is a newly introduced framework, developed by the Hungarian SZTAKI and was targeted to provide automatic features for configuring and orchestrating distributed applications on single or multi cloud systems.

Our paper is structured as follows: in the next section we give a brief overview about the related work on communication middleware used in distributed systems and on fault tolerance in cloud. In section III we introduce Occopus framework in a more detailed fashion. In section

IV we analyze the possibility of executing workflows in Occopus and section V introduces our solution in detail. Finally in the last section we give a brief insight into our fault tolerant proposal and the conclusion closes our work.

II. Related Work

A. Communication middleware in the cloud

As distributed applications transcending geographical and organizational boundaries the demands placed upon their communication infrastructures will increase exponentially. Modern systems operate in complex environments with multiple programming languages, hardware platforms, operating systems and the requirement for dynamic deployments, and reliability while maintaining a high Quality-of-Service (QoS).

Cloud orchestration in general means to build up and manage interconnections and interactions between distributed services on single or multi cloud systems. At first the orchestrator allocates the most appropriate resource for a job from a resource pool, than monitors the functioning of the resource with a so called heartbeat mechanism. However, this mechanism only gives feedback about the physical status of the resources (CPU, memory usage, etc.) it cannot provide reliability in communication and data sharing.

Concerning scientific workflows the main challenge is to provide high availability, reliable communication, fault tolerance and SLA based service. In most scientific workflow management system a special middleware is responsible to maintain the connection, and data movement between the distributed services and to schedule the tasks according to available resources and predefined constraints (data flow model).

To provide a reliable, flexible and scalable communication between the services a suitable communication middleware is needed.

Communication middlewares can be categorized as Remote Procedure Call (RPC) oriented Middleware, Transaction-Oriented Middleware (TOM), Object-Oriented/Component middleware (OOCM) and Message-Oriented Middleware (MOM) [6].

RPC oriented Middleware is based on a client-server architecture and provides remote procedure calls through APIs. This kind of communication is synchronous to the user, since it waits until the server returns a response thus it does not enable a scalable and fault tolerant solution for workflows [5].

A Transaction-Oriented Middleware (TOM) is used to ensure the correctness of transaction operations in a distributed environment. It is primarily used in architectures built around database applications [14]. TOM supports synchronous and asynchronous communication among heterogeneous hosts, but due to its redundancies and control information attached to the pure data for ensuring high reliability, it results in low scalability in both the data volume that can be handled, and in the number of interacting actors.

An Object-Oriented/Component Middleware (OOCM) is based on object-oriented programming models and supports distributed object request. OOCM is an extension of Remote Procedure Calls (RPC), and it adds several features that emerge from object-oriented programming languages, such as object references, inheritance and exceptions. These added features make OOCM flexible, however this solution enables still limited scalability.

A Message-Oriented Middleware (MOM) allows message passing across applications on distributed systems. A MOM provides several features such as:

- asynchronous and synchronous communication mechanisms
- data format transformation (i.e. a MOM can change the format of the data contained in the messages to fit the receiving application [16])
- loose coupling among applications
- parallel processing of messages
- support for several levels of priority.

Message passing is the ancestor of the distributed interactions and one of the realizations of MOM. The producer and the consumer are communicating via sending messages. The producer and the consumer are coupled both in time and space; they must both be active at the same time. The consumer receives messages by listening synchronously on a channel and the recipient of a message is known to the sender.

Message queues are newer solutions for MOM, where messages are concurrently pulled by consumers, as well as a subscription based exchange solution, allowing groups of consumers to subscribe to groups of publishers, resulting in a communication network or platform, or a message bus. Message queues provide an asynchronous communication protocol. Its widespread popularity lies in not only its asynchronous feature but in the fact that it provides persistence, reliability and scalability enabling both time and space decoupling of the so called publishers and consumers.

Advanced Message Queuing Protocol (AMQP) [1] is an open standard application layer protocol for message-oriented middleware. RabbitMQ [3] is an open source message broker software (sometimes called message-oriented middleware) that implements the AMQP and can be easily used on almost all major operating systems.

B. Fault tolerance in cloud

Although cloud computing has been widely adopted by the industry, still there are many research issues to be fully addressed like fault tolerance, workflow scheduling, workflow management, security, etc [8]. Fault tolerance is one of the key issues amongst all. It is a complex challenge to deliver the quality, robustness, and reliability in the cloud that is needed for widespread acceptance as tools for the scientists' community.

To deal with this problem many research has been already done in fault tolerance. Fault tolerance policy can be proactive and reactive. While the aim of proactive

techniques is to avoid situations caused by failures by predicting them and taking the necessary actions, reactive fault tolerance policies reduce the effect of failures on application execution when the failure effectively occurs. Different fault tolerance challenges and techniques (resubmission, checkpointing, self-healing, job migration, preemptive migration) have been implemented using various tools (HAProxy, Hadoop, SGuard,) in the cloud. Also there are a lot of methods created for providing fault tolerant execution of scientific workflows in the cloud. Mostly they heavily rely on sophisticated and complex models of the failure behavior specific to the targeted computing environment. In our investigations we are targeting a solution that is mostly based on the workflow structure and data about the actual execution timings retrieved from provenance database.

III. OCCOPUS ARCHITECTURE

Occopus [7] (Fig. 1) has five main components: enactor issues virtual machine management requests towards the infrastructure processor; infrastructure processor, which is the internal representation of a virtual infrastructure (enabling the grouping of VMs serving a common aim); cloud handler enables federated and interoperable cloud use by abstracting basic IaaS functionalities like VM creation service composer, which ensures that VMs meet their expected functionalities by utilizing configuration management tools and finally the information broker that decouples the information producer and consumer roles with a unified interface throughout the architecture.

After receiving the required infrastructure description the enactor immediately compiles it to an internal representation. It is the role of the enactor to forward and upgrade the node requests to the infrastructure processor and to monitor the state of the infrastructure continuously during the setup and the existence of the infrastructure. This monitoring function is achieved by the help of the info broker. Among others this component is responsible for tracking the information flow between the nodes. If it notices a failure of a node or a connection it notifies the enactor. The enactor then upgrades the infrastructure description and forwards it to the infrastructure processor.

The infrastructure processor receives node creation and node destruction requests from the enactor. During creation infrastructure processor sends a contextualized VM requests to the cloud handler. Within the contextualization information the processor places a reference to some of the previously created attributes of VMs.Node destruction requests are directly forwarded to the cloud handler component.

The cloud handler as its basic functionality, provides an abstraction over IaaS functionalities and allow the creation, monitoring and destruction of virtual machines. For these functionalities, it offers a plugin architecture that can be implemented with several IaaS interfaces (currently Occopus supports EC2, nova, cloudbroker, docker and OCCI interfaces).

The main functionality of the Service Composer is the management of deployed software and its configuration on the node level. This functionality is well developed and even commercial tools are available to the public. Occopus Service Composer component therefore offers interfaces to these widely available tools (e.g., Chef, Puppet, Docker, Ansible).

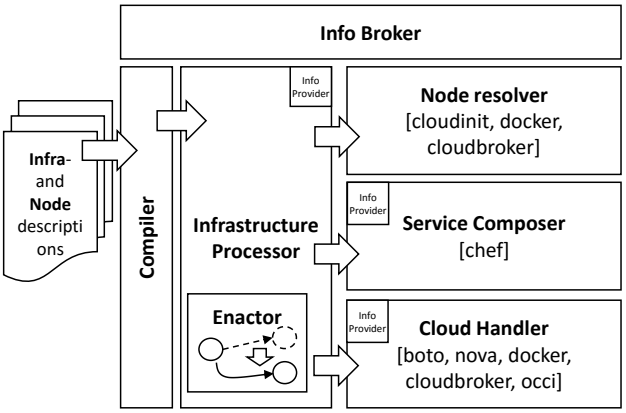


Figure 1. Occopus architecture

IV. SCIENTIFIC WORKFLOWS IN OCCOPUS

A. Scientific workflows

In Occopus framework a virtual infrastructure can be built upon a directed acyclic graph representing some complex scientific experiment which consists of numerous computational steps. The connection between these computational steps represent the data dependency, in other words the dataflow during the experiment. With Occopus the infrastructure descriptor would contain the needed resource requirements for each task and also the SLA for the tasks or for the whole workflow. An SLA requirement could be time or budget constraint or a need for green execution. In such a scenario the execution could be seen as data flowing across the VMs starting from the entry task executed on the first VM, terminating by the exit task executed on the last VM. In a scenario like this every computational step is mapped to an individual VM. After submitting the virtual infrastructure descriptor based on the workflow model Occopus would support the creation of an infrastructure like this. This type of workflow creation and execution is called Service Choreography in related works.

1) Advantages

Concerning the execution of scientific workflows in Occopus has several advantages:

The resources are available continuously, it means that task execution are not forced to wait for free resource. The infrastructure is built up easily without expertise knowledge of the individual cloud providers. The monitoring is also provided by the Occopus framework. There is no need for scheduling and resource allocation is done by Occopus based on the virtual infrastructure descriptors.

2) Problems with scientific workflows executed in Occopus

Concerning scientific workflows there might arise a lot of issues:

Scientific experiments are being data and compute intensive which may last for even weeks or month and may use or produce even terabytes of data. Due to the long execution time many failure could arise. Types of faults that can arise during execution and need to be handled in order to provide a fault tolerant execution can mainly be categorized to the following categories: VM faults, programming faults, network issues, authentication problems, file staging errors, and data movement issues. In order not to lose the already calculated work, fault tolerance must be provided.

• When a node fails, the computation which was done by this node is lost. As described in the previous section Occopus monitors the nodes of the virtual infrastructure and when the enactor notices a failed node, it is deleted from the virtual infrastructure list and a new one is created. The execution could be restarted on it. But what should happen with the data that was consumed by this failed node?

• Let us focus on only one aspect of SLA-s (Service Level Agreement), namely the time constraints. Scientific workflows are often constrained by soft or hard deadlines. While soft deadline means that the proposed deadline should be met with a probability *p*, hard deadline means that the results are useless after the deadline. When there is a failure upon recovery the makespan of the whole workflow is increased and maybe deadlines cannot be met. How can it be ensured that SLAs are met?

• Fault tolerance technique should also be concerned when executing scientific workflows in the cloud. The most frequent fault tolerant techniques in the cloud are using resubmission and replicas. How can it be ensured that more than one successors of the same type (replica) is able to receive the results of the predecessor(s) and how can it be provided that the number of replicas can change dynamically in time?

In the next section we are looking for the best solution that address the issues described above.

V. SOLUTIONS

There are two main widespread used alternatives that can give solutions for the above mentioned problems and are supported by open source softwares. One of them is based on service discovery feature, while the other uses a message queueing system.

A. Service Registry

Service discovery is a key component of most distributed systems and service oriented architectures deploying more services. Service locations can change quite frequently due to host failure or replacement, similarly to a scientific workflow execution. A node must discover somehow the IP address and the port number of the peer application. One solution is to use a dedicated, centralized service registry node.

A service registry, is a database of services, their instances and their locations. The main task of it is to register hosts, ports and authentication credentials, etc. Service instances are registered with the service registry on startup and deregistered on shutdown. Clients of the service query the service registry to find the available instances of a service. If a node fails the service registry database is upgraded with the new node.

Concerning workflow execution if a node fails then the service registry updates its database with the new client, but the computation that was already done is lost. Also the consumed data is lost with the failed node. If a computation has successfully terminated on a VM then the results of this computation task can be (should be) stored in provenance database, but because of the nature and size of the provenance database, it must be located on a permanent storage. To retrieve these data from this storage may have high latency due to geographic location which can be far from the cloud provider.

The flexibility of the solution is also not so good. In this case the nodes know where to send data so they use the synchronous remote procedure call or the asynchronous message passing middleware. As it was mentioned already in the related work the RPC model does not support large volume of data and does not support reliable transport of data. Also with message passing middleware the communication abstract is the channel and a connection must be set up between producer and consumer and consumer listens for the channel synchronously.

Using this solution a special agent would be needed to orchestrate the execution of the workflow itself. Without this agent this solution would work only for small workflows that use does not move high volume of data does not need long time to be executed and the reliability of the resources are high.

B. Message Queuing

Using message queues would simplify almost all of the above mentioned problems. Advanced Message Queuing (AMQ) Protocol is an open standard message oriented middleware. In this approach the message producer does not send the message directly to a specific consumer instead characterize messages into classes without knowledge of which consumer there may be. Similarly, consumers only receive messages that are of interest, without knowledge of existing producers. AMQP operates over an underlying reliable transport layer protocol such as Transmission Control Protocol (TCP). The basic idea is that consumers and producers use a special node to accomplish message passing, which serves as a rendezvous point between senders and receivers of messages. They are the queues which are buffers that temporary or permanently store the messages. The middleware server has two main functionalities: one of them is buffering the messages in memory or on disk when the recipient cannot accept it fast enough and the other one is to route the messages to the appropriate

queue. When a message arrives in a queue the server attempts to deliver it to the consumer immediately. If this is not possible the message is stored until the consumer is ready. If it is possible the message is deleted from the buffer immediately or after the consumer has acknowledged it. The reliability lies in this feature. The acknowledgments could be sent only when the node had successfully processed the data. The scalability and fault tolerance can be realized by clustering the same type of nodes. In this solution the consumers and producers are not known to each other and there can be more than one consumer belonging to a single queue. The power of AMQP comes from the ability to create queues, to route messages to queues and even to create routing rules dynamically at runtime based on the actual environmental conditions. This feature would enable to realize an SLA based, fault tolerant execution for workflows.

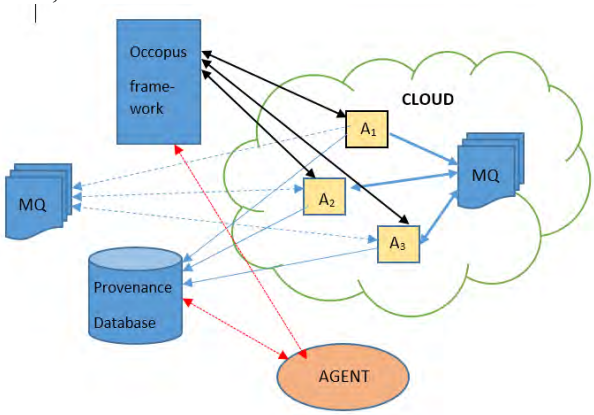


Figure 2. Possible architecture with Message Queueing

In Fig. 2 a possible architecture for executing scientific workflows with Occopus can be seen. The abstract model of the scientific workflow consists of 3 tasks (A_1, A_2 and A_3) for each an individual VM is started in the cloud. These are provided by the Occopus framework. Also Occopus does the monitoring of the resources, as well as the infrastructure upgrading. All of the tasks are communicating with the MQ (message queue), so they are not aware of each other. The MQ can be positioned also in the cloud on a VM or on external storage. It depends on the amount of data that must be shared between the tasks and the geographic location of the VMs. The Agent is only responsible for monitoring the workflow execution according to the predefined constraints (the input, output format of the data, time constraints, etc.) and according to the SLA to request a virtual infrastructure change from the Occopus framework (for example to start more or less replicas of the tasks).

VI. FAULT TOLERANCE METHOD BASED ON WORKFLOW STRUCTURE

Concerning time critical applications a reliable fault tolerant method should be provided. In this section we lay down the bases of our fault tolerant framework that uses replicas in order to ensure that time critical workflows can be successfully terminated before the soft or hard deadline with a probability of p. In our solution every task in a workflow is assigned certain number of replicas. The number of replicas is determined by the estimated execution time of a task, the structure of the workflow, the failure zone of a task (which is the affected tasks in the case of a failure of a given task), and the estimated failure detection time and resubmission time. Before Occopus starts to build the infrastructure the algorithm is executed to determine the number of replicas of each task and the infrastructure is built. When during execution unexpected situation happens (for example too many failures occurring, or even that there is no error) than the number of replicas can be changed accordingly since Occopus is able to upgrade the infrastructure. Determining the exact details of our algorithm is our future work.

VII. CONLUSION

In this paper we have introduced Occopus, a one click cloud orchestrator framework, which supports distributed applications to be executed in single or multi homed clouds. We have investigated the advantages and problems of having scientific workflows being executed with Occopus and gave a proposal for a communication middleware which would provide a reliable, fault tolerant workflow execution environment. We gave a first insight for a fault tolerant method that can be used with Occopus as well and which detailed work out determines our future research direction.

REFERENCES

[1] AMQP Advanced Message Queuing Protocol, Protocol Specification, Version 0-9-1, 13 November 2008.

[2] E. Curry, Message-Oriented Middleware, in: Q.H. Mahmoud (Ed.), Middleware for Communications, John Wiley and Sons, Chichester, England, 2004, pp. 1–28.

[3] A. Videla, J.W. Williams, RabbitMQ in Action: Distributed Messaging for Everyone, MEAP Edition Manning Early Access Program, 2011.

[4] P.T. Eugster, P. Felber, R. Guerraoui, A.-M. Kermarrec, "The many faces of publish/ subscribe", ACM Comput. Surv. 35 (2) (2003) 114–131.

[5] K. Geihs, Middleware challenges ahead, IEEE Comput. 34 (6) (June 2001) 24–31.

[6] M. Albano et al. Message-oriented middleware for smart grids, Computer Standards & Interfaces 38 (2015) 133–143.

[7] G. Kecskeméti, M. Gergely, A. Visegrádi, Zs. Németh, J. Kovács, P. Kacsuk, One Click Cloud Orchestrator: bringing Complex Applications Effortlessly to the Clouds, WORKS 2014.

[8] A. Bala, I. Chana, Fault Tolerance- Challenges, Techniques and Implementation in Cloud Computing, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012

FPGA Kernels for Classification Rule Induction

P. Škoda* and B. Medved Rogina*
* Ruđer Bošković Institute, Zagreb, Croatia
pskoda@irb.hr

Abstract - Classification is one of the core tasks in machine learning data mining. One of several models of classification are classification rules, which use a set of if-then rules to describe a classification model. In this paper we present a set of FPGA-based compute kernels for accelerating classification rule induction. The kernels can be combined to perform specific procedures in rule induction process, such as evaluating rule coverage, or estimating out-of-bag-error. Since classification problems are getting increasingly larger, there is a need for faster implementations of classification rule induction. One of the platforms that offer great potential for accelerating data mining tasks is FPGA (field programmable gate array), which provides the means for implementing application specific accelerators.

Key words - FPGA, dataflow, machine learning, classification rules

I. INTRODUCTION

Classification is one of the fundamental tasks in machine learning, and is used in a wide variety of application. One of the fundamental models of classification is classification rules [1]. Classification rules express the classification model as a set of IF-THEN rules, and enable implementation of fast classifiers with high throughput. Classification rule induction is usually performed in one of two ways: (a) by constructing a decision tree and then extracting rules form it; or (b) by using a covering algorithm. Applications of classification, as well as the field of data mining in general, are faced with continuing increase in dataset sizes. This drives efforts to develop new faster and more efficient algorithms, as well as developing new hardware platforms that provide high computational power [2].

One hardware platforms that is gaining more ground in computing is the field programmable gate array (FPGA). FPGA. FPGAs are digital integrated circuits designed to be user-configured after manufacturing [3], [4]. They enable fast development and deployment of custom digital hardware, and thus allow implementation of custom computational architectures that can be reconfigured on demand.

The most suitable computational mode for FPGA is the dataflow model. Most computer systems implement a control-flow model, in which the computation is described by a sequence of operations that are executed in specified order. In dataflow model, computation is described by chain of transformations applied to a stream of data [5]. The dataflow is described in form of a graph. Nodes of the graph represent operations, and edges represent data links between the nodes. The dataflow graph maps naturally to the hardware implementation. Nodes/operations are translated to functional hardware blocks, while edges are translated to data buses and signals. By dividing the graph into stages separated by registers, it's transformed into a pipelined structure suitable for implementation on FPGA.

In this paper we present a novel set of kernels for accelerating classification rule induction that uses a variant of covering algorithm. The kernels are targeted for dataflow architecture realized on FPGA.

II. RELATED WORK

While there has been a lot of activity in implementing classification rules on FPGAs, mostly for network filtering applications, there is no published work known to authors related to rule learning. There is, however, some work in the closely related problem of decision tree induction [6]–[8].

Narayanan et al. [7] have used FPGA to implement decision tree induction system for binary classification. They implemented Gini impurity computation, as the most intensive part of the process. The rest of the algorithm is executed on a PowerPC CPU embedded in the FPGA device.

Chrysos et al. [8] implement frequency table computation on FPGA as part of HC-CART system for learning decision trees. Frequency table computation is implemented in Frequency Counting module, which receives attribute-class label pairs. The pairs are used to generate addresses for frequency table locations which contents are then incremented by one. In their implementation, all attribute-class label pairs received by the kernel have to be prepared by program running on CPU, and then preloaded to memory attached to FPGA. Input data is transferred from CPU to FPGA memory many times in the course of program execution.

In our previous work [9], we implemented a 2D frequency table computation using FPGAs. The kernel receives two input streams – one for address and one for class labels – and counts occurrences of pairs in FPGA's internal memory. The dataset is held in FPGA attached SDRAM. We built upon this work by expanding it to multiple attribute streams [10], which allow computation of several tables simultaneously, and by implementing efficient data transfer for the computed tables.

In this paper we present out work

This work was supported in part by Maxeler University Programme, and by Croatian Science Foundation under the project number I-1701-2014.

TABLE I. BASIC INFORMATION ON MAXELER VECTIS-LITE FPGA BOARD

FPGA	Xilinx Virtex-6 XCVSX475T
Off-chip RAM (LMem)	6 GiB (6×1 GiB) DDR3-800 SDRAM
On-chip RAM (FMem)	~ 4 MiB Block RAM
CPU ↔ FPGA bandwidth	2 GB/s
LMem ↔ FPGA bandwidth	38.4 GB/s max.

III. DATAFLOW ENGINE ARCHITECTURE

The accelerator is implemented in form of a dataflow engine (DFE) which consists of at least one kernel and a manager. The computation is implemented by kernels, while the role of manager is to organize data movement within the DFE. Kernels are instantiated within the manager, and data links between the kernels, external RAM, and CPU are defined.

The frequency table computation was implemented using the Maxeler platform [11]. The platform includes an FPGA board, drivers, and API. The DFE is coded in Java, using the API which provides objects and methods that are translated to hardware units by the compiler. The compiler translates the Java code into VHDL, and generates software interface for CPU code [12]. The DFE was implemented on Maxeler Vectis-Lite board, which contains a single FPGA, and on-board SDRAM. The board is connected to the host PC workstation via PCIe bus. Basic information on the board is shown in Table I.

A. Count conditional kernel

Architecture of the *ComputeFreq-MS* kernel is shown in Fig. 1. The kernel has two scalar inputs for parameters: one for number of items to process (*items*), and one for stream length in items (*strmLen*). Stream length sets the number of items to read from the on-board SDRAM, which requires all accesses to be made in 96 byte blocks, i.e. 24 item blocks (4 bytes per item). The kernel processes the first *items* elements from the stream, while the rest are read from memory and ignored.

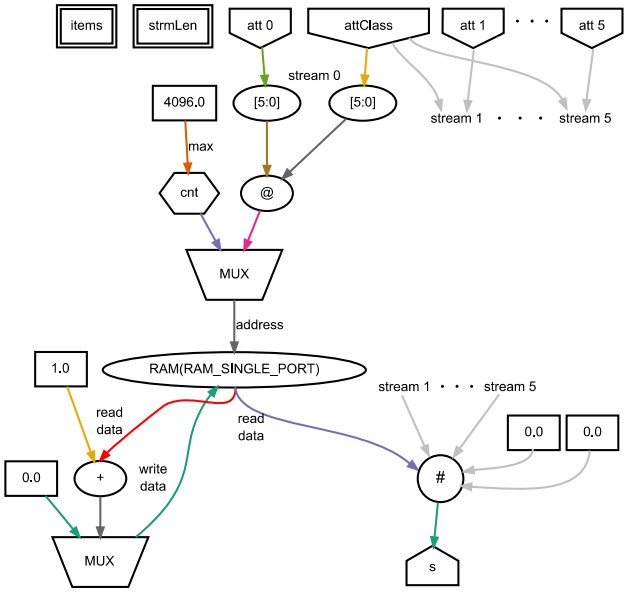


Figure 1. ComputeFreq-MS kernel architecture for up to 64 unique attribute and class values ($N_A = 6$, $N_C = 6$).

The kernel consists of six identical frequency counter structures. Each frequency counter receives two input streams – one for attribute and one for class values. Both streams carry 32 bit unsigned integers. The input streams are sliced so the low N_A bits retained from the attribute, and N_C bits are retained from the class values. The retained bits are concatenated to form the frequency table address. In the table address, the attribute value is the high, and the class value is the low part of the address word.

The kernel has a total of seven stream inputs, six for attribute (att_0 – att_5), and one for class (*class*) value streams. Each attribute value stream is connected to a single frequency counter. The class value stream fans out to all frequency counters, so that they all receive identical class value streams.

B. Increment conditional kernel

The frequency tables are stored in on-chip memory – block RAM (BRAM). Each frequency counter has one BRAM which is addressed by the address formed from the low bits of attribute and class streams. The addressed location is read, its value incremented by one, and written back to the same location. The BRAM is configured as single port RAM with read-first synchronization. Due to access latencies of the BRAM itself, and the registers inserted by the compiler, there are latencies inside the loop. To ensure correct counting, the kernel's input streams are throttled to allow input of one element each m clock cycles, where m is the internal latency of the loop.

After all elements of interest are streamed in, the contents of the BRAMs are streamed out to the host main memory through stream output s. The output stream s is made by joining output streams from all frequency counters (s_0 – s_5) into a one vector stream. The output vector is padded to 8 elements, i.e. two additional dummy elements are added to it. The padding enables efficient conversion of the output stream to word width of 128 bits which is used in communication over PCIe bus.

As BRAMs are read, they are at the same time reset to zero. During the read phase, the BRAMs are addressed by a counter, and their write ports are switched from increment-by-one values to constant zero. In this case there are no loop dependencies, and the output stream is not throttled – it outputs one element every clock cycle. At the same time, any remaining elements in the input stream are read un-throttled (one per clock) and ignored.

C. Logic simple kernel

A single *ComputeFreq-MS* kernel is instantiated in manager. Input streams *attClass* and att_0 – att_5, are linked

TABLE II. FPGA RESOURCE USAGE BY THE DFE

Resource	Used	Total available	Utilization
LUTs	34,724	297,600	11.67 %
Flip-flops	50,064	297,600	16.82 %
DSP blocks	0	2,016	0.00 %
Block RAM	226	2,128	10.62 %

to on-board SDRAM. All input streams use linear memory access pattern. Output stream *s* is linked to computer main memory. Source addresses for input streams $att_0 - att_5$, *attClass*, values for scalar inputs *items* and *strmLen*, and destination address for output stream *s* are defined through the generated DFE interface.

The kernel parameters were set for up to 64 unique attribute and class values (N_A = 64, N_C = 64). Each frequency table holds 4096 32-bit words. Due to the frequency counter's loop latency, it can process one item (attribute - class pair) every five clock cycles. Kernel clock frequency is set to 300 MHz. With six attribute streams, the DFE can process up to 360×10^6 elements per second. FPGA resource usage for the DFE is given in Table II.

IV. Experimental Results

A. Test environment

The kernel was benchmarked using code form C4.5 Release 8 decision tree learning program [13]. Parts needed to load the datasets, and computing the frequency tables were extracted from the program and used for benchmarking. The frequency table computation was parallelized using OpenMP [14] by distributing attributes of the dataset over the threads. This included replicating the frequency table data structure to accommodate multi-threaded execution. The original *ComputeFrequencies* function was modified to use the replicated data structure. Since the attributes were distributed over threads, multiple invocations of *ComputeFrequencies* function were necessary if number of attributes exceeded the number of threads.

Functions for transforming the dataset to the appropriate format, and loading it to the DFE were added to the benchmark program. For execution on DFE, a new function was created by modifying the original *ComputeFrequencies* function. Modifications involve removing the computation loop and replacing it with function calls to the DFE. Received results are transformed into the frequency table data structure used by the parallelized software implementation. Since the DFE processes six attributes in parallel, the *ComputeFrequencies* function is invoked once for every group of six attributes.

Time measurement was added to the *ComputeFrequencies* function. Entire run time of the function is measured. The function's throughput is calculated from the measured times using the formula:

$$T_{a,i} = \frac{a\,i}{t_{a,i}} \tag{1}$$

where a is number of attributes, i is number of items, $t_{a,i}$ is function execution time, and $T_{a,i}$ is the calculated function throughput for the dataset with a attributes, and i items.

A set of 102 synthetic datasets were used in the benchmark. Number of attributes ranged from 6 to 1,536, in a geometric sequence 6×2^i, with i = 0, 2... 7. Number of items ranged from 2048 to 4×2^{20}, in a geometric sequence 2^i, with i = 11, 12... 22. The dataset size was limited to 3 GiB, i.e. to 768×2^{20} elements. Datasets were generated randomly, with uniformly distributed values in range of 1-63 for attribute values, and 0-64 for class values.

Benchmark program was compiled using gcc compiler, version 4.4.7. Benchmarks were run on Intel Xeon E5-1600 workstation, with 16 GiB DDR3-1600 RAM, under Centos 6.5 Linux OS.

B. CPU benchmark results

For the baseline CPU performance, the measurements were conducted on a six-threaded parallelized implementation of *ComputeFrequencies* function. Measured execution times are given in Table III, and are shown in Fig. 2. For datasets with over 128×2^{10} items, execution time scales approximately linearly with number of items. There is a larger increase in execution in area of 16×2^{10} – 32×2^{10} items, for datasets with 96 or more attributes, and 32×2^{10} – 256×2^{10} for datasets with less than 96 attributes.

This increase in execution is more obviously visible in function throughput, shown in Fig. 3, as a sharp drop of the throughput. From the figure it's visible that for datasets with fewer items, the throughput increases with increasing number of items. Once the dataset exceeds a certain number of items (dependent on number of

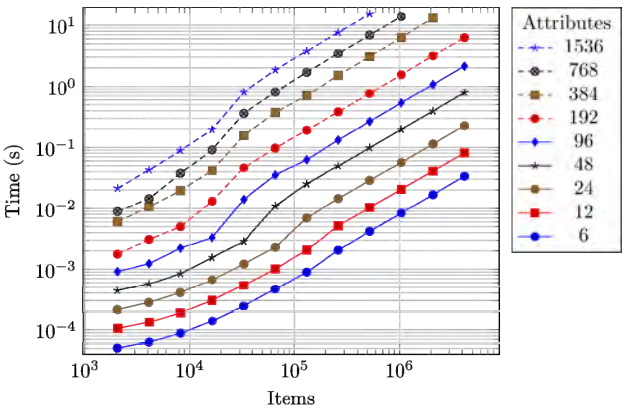


Figure 2. Execution time on CPU as function of number of items, measured for datasets with 6 to 1,536 attributes

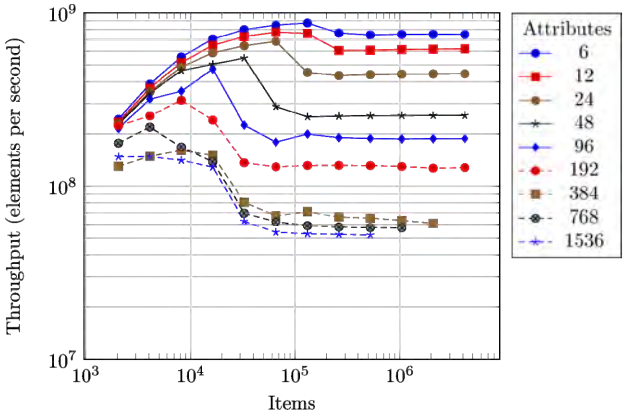


Figure 3. Throughput on CPU as function of number of items, measured for datasets with 6 to 1,536 attributes

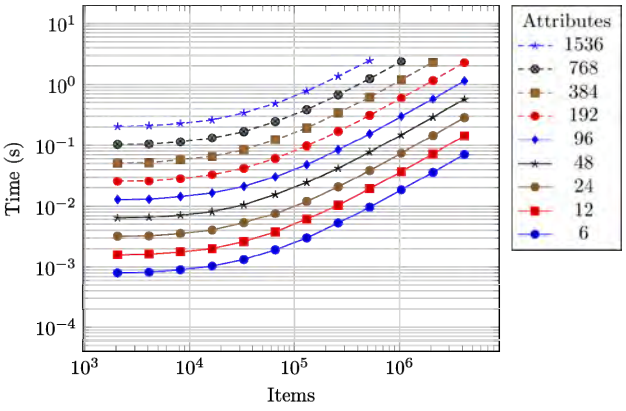


Figure 5. Execution time on DFE as function of number of items, measured for datasets with 6 to 1,536 attributes

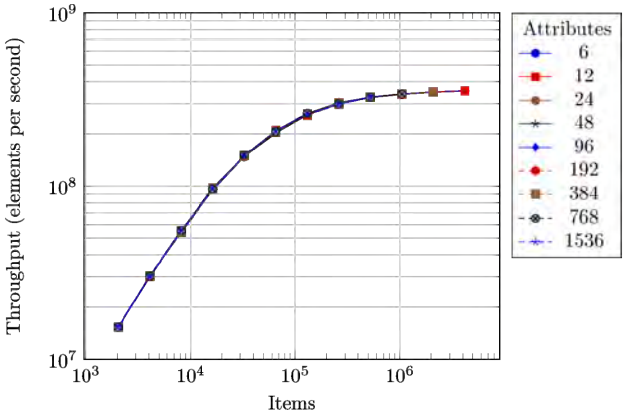


Figure 4. Throughput on DFE as function of number of items, measured for datasets with 6 to 1,536 attributes

attributes), throughput drops, and quickly stabilizes to an approximately constant value. As number of attributes increases, function throughput decreases, assuming equal number of items. The number of items at which peak throughput is recorded, also decreases with increasing number of attributes, as well as the peak throughput value itself. For datasets with 384 or more attributes, there is no increase of throughput. Throughput just drops from one approximately constant value to another.

Maximum measured constant (invariant to number of items) throughput is 617×10^6 elements/s, achieved on dataset with 12 attributes. Minimal measured constant throughput is 52.2×10^6 elements/s, measured on dataset with 1,536 attributes.

The drop in function throughput is most likely a consequence of interaction of CPUs cache system and main SDRAM. Smaller datasets have a smaller likelihood of cache miss, which translates into higher throughput for smaller datasets. Another important factor is the data structure for storing ht dataset. The dataset is stored in attribute major order, which leads to memory accesses with stride. With more attributes, the stride is larger, and consequentially, so is the likelihood of cache miss.

C. *DFE benchmark results*

DFE was benchmarked under the same condition as the six-threaded software implementation. Execution times measured on DFE are given in Table IV, and shown in Fig 4. Execution time on DFE scales approximately linearly with number of items in the dataset, on datasets with 512×2^{10} or more items. On datasets with fewer items, the execution times asymptotically approach a certain minimal value, dependent on number of attributes. For a fixed number of items, the execution times scale linearly with number of attributes.

The kernel throughput graph, shown in Fig 5, shows that curves for all number of attributes form a tight group. The curves for different number of attributed cannot be distinguished one from another. On datasets with 512×2^{10} items, the throughput is approximately constant at 340×10^6 elements/s. The measured throughput is close to the theoretical maximum of 360×10^6 elements/s.

The DFE performs better with datasets with larger number of items, close to theoretical maximum when number of items is 512×2^{10} or more. This is a consequence of communication and control overheads between the DFE and the host computer. The overhead is independent from number of items, and its influence diminishes as the number of items increases. As can be seen from Table IV, execution times are approximately constant for datasets with 8,192 or less items. The minimum DFE execution time was calculated from these measurements, amounting to 801.1 µs.

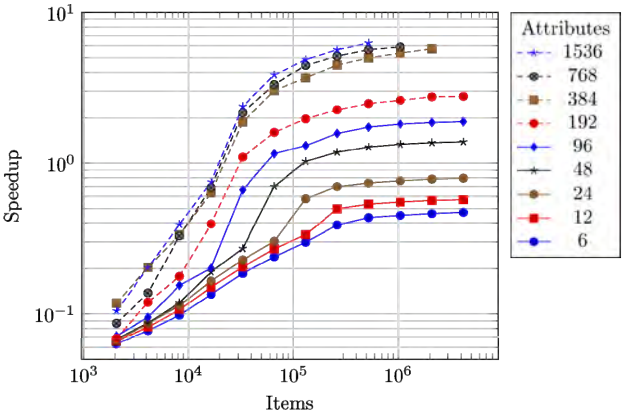


Figure 6. Speedup, measured for datasets with 6 to 1,536 attributes

D. *Comparison of the results*

DFE and CPU results were compared by computing speedup, i.e. ratio of execution times on CPU and on DFE. The speedup is shown in Fig 6. For most of the test dataset sizes, execution on DFE is slower than on CPU. For datasets with fewer than 48 attributes, the DFE is slower no matter what the dataset size is. For 48 attributes the speedup exceeds unity value for numbers of items 128×2^{10} or more. The maximum speedup is achieved on datasets with 384 and more attributes. Speedups on these datasets are close in value, suggesting that by further increasing number of attributes will result in negligible increase in speedup. These values can then be used as approximation of the upper bound on speedup. The maximum speedup measured is 6.26×, achieved on dataset with 1,536 attributes and 512×2^{10} items.

TABLE III. MEASURED EXECUTION TIMES ON CPU

Number of items	Number of attributes / Execution time								
	6	12	24	48	96	192	384	768	1,536
2,048	50.59 µs	105.3 µs	215.0 µs	434.9 µs	916.9 µs	1.769 ms	6.049 ms	8.887 ms	21.32 ms
4,096	63.31 µs	133.5 µs	280.4 µs	571.0 µs	1.237 ms	3.096 ms	10.59 ms	14.37 ms	42.51 ms
8,192	88.21 µs	190.1 µs	403.6 µs	847.8 µs	2.228 ms	5.027 ms	19.56 ms	37.46 ms	89.34 ms
16,384	139.2 µs	302.8 µs	668.8 µs	1.557 ms	3.321 ms	13.08 ms	41.85 ms	91.15 ms	195.6 ms
32,768	245.6 µs	538.0 µs	1.219 ms	2.867 ms	13.99 ms	46.18 ms	156.3 ms	360.6 ms	803.1 ms
65,536	462.1 µs	1.014 ms	2.296 ms	10.92 ms	35.06 ms	97.45 ms	373.2 ms	808.1 ms	1.860 s
131,072	899.0 µs	2.066 ms	6.949 ms	25.08 ms	62.94 ms	191.5 ms	706.2 ms	1.702 s	3.795 s
262,144	2.056 ms	5.184 ms	14.48 ms	49.73 ms	132.1 ms	382.5 ms	1.518 s	3.467 s	7.658 s
524,288	4.215 ms	10.35 ms	28.61 ms	98.80 ms	267.1 ms	766.7 ms	3.079 s	6.987 s	15.44 s
1,048,576	8.365 ms	20.48 ms	56.80 ms	197.3 ms	537.8 ms	1.553 s	6.359 s	14.01 s	–
2,097,152	16.70 ms	40.80 ms	113.3 ms	394.1 ms	1.071 s	3.176 s	13.21 s	–	–
4,194,304	33.49 ms	81.27 ms	226.3 ms	787.0 ms	2.139 s	6.303 s	–	–	–

The total performance of the DFE was estimated by calculating the weighted average of speedups, weighted by number of elements in the dataset:

$$S = \frac{\sum_a \sum_n a\, n\, s_{a,n}}{\sum_a \sum_n a\, n} \quad (1)$$

where S is the average speedup, and $s_{a,n}$ is the speedup on dataset with a attributes and n items. The average speedup is 4.1.

For further comparison, execution efficiency was computed for CPU and DFE. The efficiency is defined as:

$$E_{a,n} = \frac{m\, f}{F_{a,n}} \quad (1)$$

where $E_{a,n}$ is efficiency and $F_{a,n}$ is throughput for dataset with a attributes and n items, f is clock frequency, and m is number of threads or number of streams on CPU and DFE respectively. Peak throughput values were used to calculate efficiency.

The software implementation (CPU baseline) was executed on six threads, on CPU clocked at 3.2 GHz. Peak throughput was 875×10^6 elements/s. The DFE uses six streams, and was clocked at 300 MHz. Peak throughput on DFE was 355×10^6 elements/s. Execution efficiency on CPU is 21.9 clock cycles per element, and on DFE it's 5.08 clock cycles per element. Of the 5.08 clocks, 4 are consequence of the frequency counter's internal loop latency.

V. CONLUSION

In this paper we presented a multi-streamed compute architecture for 2D frequency matrix computation, implemented on FPGA platform. This multi-streamed architecture is an advancement on our previous work [9].

Benchmark results reveal that CPU outperforms DFE for smaller datasets. This is a consequence of control and communication latencies between the FPGA board and the host computer. Minimal time required to execute any action on DFE is approximately 800 µs, regardless of dataset size. Small datasets are easily processed by the CPU in less time. The DFE outperforms CPU for larger datasets that have at least 48 attributes, 32×2^{10} items, and 6×2^{20} elements. Best speedup achieved by DFE is 6.26×. The DFE is more efficient in processing data, requiring only 5.08 clock cycles per dataset element, while CPU requires 21.9, when performing at peak efficiency. Of the DFE's 5.08 clock cycles, 4 cycles are consequence of frequency counter's internal loop latencies.

TABLE IV. MEASURED EXECUTION TIMES ON DFE

Number of items	Number of attributes / Execution time								
	6	12	24	48	96	192	384	768	1,536
2,048	800.6 µs	1.597 ms	3.220 ms	6.385 ms	12.86 ms	25.69 ms	51.29 ms	102.5 ms	203.4 ms
4,096	817.1 µs	1.634 ms	3.248 ms	6.559 ms	13.00 ms	25.90 ms	51.89 ms	104.2 ms	209.2 ms
8,192	898.9 µs	1.780 ms	3.544 ms	7.133 ms	14.39 ms	28.33 ms	58.38 ms	113.6 ms	227.3 ms
16,384	1.031 ms	2.014 ms	4.043 ms	8.187 ms	16.47 ms	33.04 ms	65.85 ms	131.9 ms	260.4 ms
32,768	1.322 ms	2.626 ms	5.338 ms	10.55 ms	21.03 ms	42.13 ms	83.68 ms	167.0 ms	337.5 ms
65,536	1.929 ms	3.758 ms	7.576 ms	15.54 ms	30.35 ms	60.91 ms	123.2 ms	243.9 ms	483.4 ms
131,072	3.009 ms	6.153 ms	11.98 ms	24.43 ms	48.17 ms	97.42 ms	192.3 ms	381.3 ms	779.6 ms
262,144	5.277 ms	10.41 ms	20.70 ms	42.10 ms	83.98 ms	169.3 ms	338.4 ms	675.5 ms	1.356 s
524,288	9.664 ms	19.35 ms	38.62 ms	77.18 ms	154.0 ms	308.5 ms	616.5 ms	1.236 s	2.467 s
1,048,576	18.55 ms	37.08 ms	74.19 ms	147.9 ms	296.4 ms	594.5 ms	1.184 s	2.374 s	–
2,097,152	36.03 ms	72.09 ms	144.2 ms	288.3 ms	576.1 ms	1.154 s	2.307 s	–	–
4,194,304	70.96 ms	142.0 ms	284.1 ms	567.9 ms	1.136 s	2.272 s	–	–	–

The kernel can be further improved by vectorizing inputs stream in the same manner its output was vectorized. This will allow efficient processing of more elements in parallel, and better utilization of available memory bandwidth. Another improvement would be reducing or eliminating the internal loop latencies. This can potentially quadruple its performance assuming that operating frequency of the kernel does not drop due to higher design complexity.

In future work, in addition to the stated improvements, this kernel will be integrated in the C4.5 program as accelerator unit. However, to achieve significant performance gains the communication latency between host computer and FPGA will have to be compensated for. This will most likely involve adding support for batch processing of a series of small dataset, and require some modifications of the original algorithm. Overall, with additional improvements the kernel is expected to outperform the CPU for a wider range of dataset sizes.

REFERENCES

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed. Morgan Kaufmann, 2006.

[2] A. N. Choudhary, D. Honbo, P. Kumar, B. Ozisikyilmaz, S. Misra, and G. Memik, "Accelerating data mining workloads: current approaches and future challenges in system architecture design," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 1, pp. 41–54, Jan. 2011.

[3] I. Kuon, R. Tessier, and J. Rose, "FPGA Architecture," *Found. Trends Electron. Des. Autom.*, vol. 2, no. 2, pp. 153–253, 2008.

[4] Mark L. Chang, "Device Architecture," in *Reconfigurable Computing: The Theory and Practice of FPGA-based Computation*, S. Hauck and A. DeHon, Eds. Morgan Kaufmann, 2008, pp. 3–27.

[5] J. B. Dennis, "Data Flow Supercomputers," *Computer (Long. Beach. Calif).*, vol. 13, no. 11, pp. 48–56, Nov. 1980.

[6] P. Škoda, B. Medved Rogina, and V. Sruk, "FPGA implementations of data mining algorithms," in *MIPRO, 2012 Proceedings of the 35th International Convention*, 2012, pp. 362–367.

[7] R. Narayanan, D. Honbo, G. Memik, A. Choudhary, and J. Zambreno, "An FPGA Implementation of Decision Tree Classification," in *2007 Design, Automation & Test in Europe Conference & Exhibition*, 2007, pp. 1–6.

[8] G. Chrysos, P. Dagritzikos, I. Papaefstathiou, and A. Dollas, "HC-CART: A parallel system implementation of data mining classification and regression tree (CART) algorithm on a multi-FPGA system," *ACM Trans. Archit. Code Optim.*, vol. 9, no. 4, pp. 1–25, Jan. 2013.

[9] P. Škoda, V. Sruk, and B. Medved Rogina, "Frequency Table Computation on Dataflow Architecture," in *MIPRO, 2014 Proceedings of the 37th International Convention*, 2014, pp. 357–361.

[10] P. Škoda, V. Sruk, and B. Medved Rogina, "Multi-stream 2D frequency table computation on dataflow architecture," in *MIPRO 2014, Proceedings of the 38h International Convention*, 2015, pp. 288–293.

[11] "Maxeler Technologies," 2015. [Online]. Available: http://www.maxeler.com/. [Accessed: 23-Dec-2015].

[12] O. Pell and V. Averbukh, "Maximum Performance Computing with Dataflow Engines," *Comput. Sci. Eng.*, vol. 14, no. 4, pp. 98–103, Jul. 2012.

[13] J. R. Quinlan, "C4.5 Release 8," 1993. [Online]. Available: http://rulequest.com/Personal/c4.5r8.tar.gz. [Accessed: 23-Jun-2013].

[14] OpenMP Architecture Review Board, "OpenMP Application Program Interface," 2011. [Online]. Available: http://www.openmp.org/mp-documents/OpenMP3.1.pdf. [Accessed: 05-Feb-2014].

VISUALIZATION SYSTEMS

Prototyping of visualization designs of 3D vector fields using POVRay rendering engine

J. Opiła*
*AGH University of Science and Technology
Department of Applied Computer Science, Faculty of Management, Cracow, Poland
jmo@agh.edu.pl

Abstract – There is a persistent quest for novel methods of visualization in order to get insight into complex phenomena in variety of scientific domains. Researchers, ex. VTK team, achieved excellent results; however, some problems connected with implementation of new techniques and quality of the final images still persist.

Results of inspection of number of visualization styles of 3D vector field employing POVRay ray-tracing engine are discussed, i.e. hedgehogs, oriented glyphs, streamlines, isosurface component approach and texturing design. All styles presented have been tested using water molecule model and compared concerning computing time, informativeness and general appearance. It is shown in the work that Scene Description Language (SDL), domain specific language implemented in POVRay is flexible enough to use it as a tool for fast prototyping of novel and exploratory visualization techniques. Visualizations discussed in the paper were computed using selected components of API of ScPovPlot3D, i.e. templates written in the SDL language. Results are compared to designs already implemented in VTK.

Keywords - POVRay, vector field visualization, ScPovPlot3D, visual data analysis, VTK.

I. INTRODUCTION

In the recent years both, computational and experimental methods deliver a lot of data. Most productive sciences includes astronomical sky surveys, engineering, econometrics, medical sciences including medical imaging and this condensed list is hardly complete. Often collected or computed data are in the form of 3D vector field, static or dynamic, on all spatial scales, for example ocean currents, wind speed or electrostatic molecular level fields. Vector fields can be described by differential equations which sometimes can be reduced to a simple formula e.g. for gravitational, or electrostatic field. However in numerous cases they can be obtained by measurements only for example winds distribution [1], [2]. Reliable analysis of such a data requires intensive usage of computer data processing and subsequent visualization step.

After many years of development ([3], [4]) visualization of 3D vector fields still may be improved according to characteristic of a specific case, e.g. by implementation of novel hybrid designs. Thus visualization library should allow not only for minor adjustments like color or proportions of data glyphs but for experimenting with brand new styles as well. This is why visualization software has to be flexible as much as possible. Unfortunately, ready-made software, always offers only closed list of variants of charts and graph layouts (e.g. [13], [14], [15], [17]), as all of them have to be explicitly compiled into the machine code of the final application or library. As a result, efficient and fast prototyping of novel designs is usually difficult, even if using low level visual toolkits, and sometimes impossible without writing many lines of a new code. Another less obvious limitation results from obligation to be member of the development team. All these weaknesses can be overcome by choice of efficient renderer (in order to obtain high quality images) equipped with scripting high level language (to avoid compilation overhead) and published as FLOSS software (to avoid license restrictions). Utilization of the POVRay [16] raytracing engine for this goal was discussed in paper [21] where results of inspection of possible extensions to known potential field visualization procedures using ScPovPlot3D templates library [22] are presented[1)].

II. VIRTUAL LABORATORY

In order to inspect problems described above, one needs to conduct some experimental research. Because all "experiments" to be done require computer only, name "numerical experiments" fits, as well as descriptive term "virtual laboratory".

A. *The choice of rendering engine.*

As in the previous work [21], the POVRay [16] has been chosen as a rendering engine, because all reasons mentioned there remained valid, i.e. only POVRay met all requirements. Amongst examined: software platforms were: Blender [5], BioBlender [6], Diderot [7], gnuplot [13], ParaView [15], RasMol [9], Shadie [10], [11], [12], VTK [15]. Some of mentioned above packages just cannot produce visualization of 3D vector field but molecules only [6], [9] or have disappointing output ([13], [15]), the other requires tedious low level coding in Python or C++ [5], [17]. For resume of this analysis refer to [21].

The work was supported by AGH University of Science and Technology.

[1] Gallery of experimental visualizations in full color as well as source files is freely available for download from [23].

B. VTK versus POVRay

While in this paper ScPovPlo3D library is often compared to VTK [17], it is mainly for reference not for contradiction. Flexibility of POVRay and written in SDL[2] ScPovPlot3D library is used for fast prototyping and testing, but of course VTK should be used to build efficient end user tool. This sequence allows avoiding learning of complicated library in C++ and laborious coding phase. In fact, adding novel designs may be possible by extending base classes of VTK but sometimes it may affect whole class hierarchy. In this case it may require not only excellent programming skills but also membership in the development team. On the other hand, testing of many variants of the visualization is much easier in dedicated domain language and often requires adjustment of single line of code only. I hope, that proposed here novel designs may someday be introduced into VTK.

C. POVRay and SDL

POVRay [16] is the raytracing program known for producing stunning photorealistic images. The output is controlled by text files containing statements written in specific to POVRay domain specific language dubbed by developers "Scene Description Language" – SDL. SDL as domain specific language is built into POVRay, and its full documentation is available on POVRay's home page and dedicated User Forum. SDL implements a couple of predefined graphics primitives (i.e. readymade "shapes"), simple [3] and complex [4] ones, all of them defined as mathematical entities. Objects inserted into scene description file are thus parameterized by a set of attributes and can also be scaled, rotated, translated and skewed. Because POVRay is founded on mathematical modeling it offers to the end user a number of valuable numerical methods like root solvers (including Sturmian solver), spline functions, fractals and vector analysis functions (ex. gradient or divergence finding). While SDL is good enough for programming of even very complex scenes, it lacks some features characteristic for modern languages like constants, structures or classes. Another problem is connected with coordinate system employed by the program, which not only is left-handed but rotated too. Because of these weaknesses only those skilled in programming and 3D aware users are able utilize its full power. Luckily, SDL implements many constructs supporting efficient Application Programming Interface (API), like macros, inline functions, function pointers, conditionals, loops and arrays. Furthermore, POVRay is equipped with comfortable Integrated Development Interface[5] (IDE), featuring amongst other syntax coloring, word completion and simple post-mortem debugger.

D. ScPovPlot3D

Few years ago stable version of libraries for POVRay, named ScPovPlot3D (**S**cientific **POVRay Plot**s in **3D**) had been released [22]. In stable branch it is capable of rendering a wide range of 3D charts [27], [28]. Work on improvements still continues in experimental branch including procedures for a.o. fuzzy numbers [25], hybrid charts [26], vector field and potential visualization. Moreover helper modules delivers easy to use helper items like smart cameras ("Cameras.inc" library), handful of coordinate systems visualizations ("CoordsSys.inc"), legend keys as well as format enriched text ("TextExt.inc").

The toolkit is addressed to a wide audience due to an assumed balance between ease of use, flexibility and quality of produced images. For example, less experienced users can prepare custom charts in minutes, just by modifying included examples, while skilled users may tune code of the scene to obtain desired effect. In the following chapters prototyping of visualization styles of 3D static vector fields using the POVRay, complemented with ScPovPlot3D library is discussed.

E. Choice of exemplary vector field source

In order to make a fair comparison of variety visualization designs the same vector field produced by selected molecular structure should be used for all tested styles. The selected structure should obey at least two, contradictory, constraints. Firstly, it should be complex enough for making field distribution unpredictable based solely on intuition. Secondly, it needs to be as simple as possible to be computable on regular workstation, which shortens testing cycle. Based on these assumptions arbitrary and simplified (i.e. flat) water cluster comprised from 6 water molecules has been chosen, see Fig.1. The structure is referred across the paper as $(H_2O)_6$.

F. Water

Water, or more properly *hydrogen oxide*, appears to be simple chemical compound. However elusive quantum

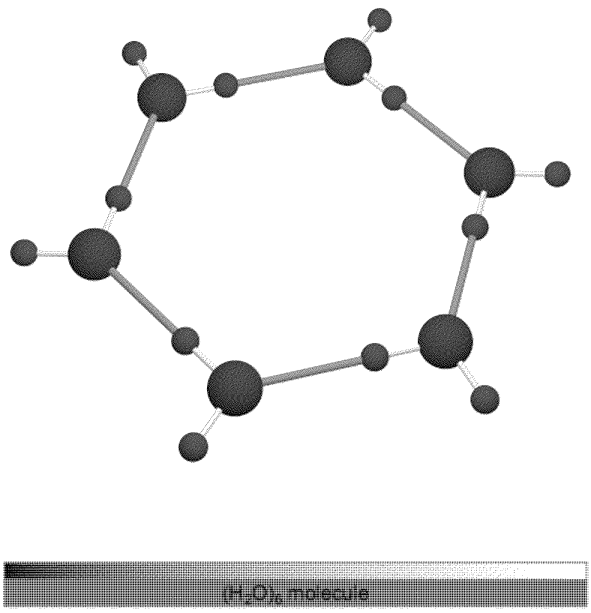


Figure 1. Visualization of flat, six molecule water cluster $(H_2O)_6$. Atoms are marked as spheres scaled according to respective ionic radius: hydrogen as smaller while oxygen larger. Bonds are rendered using thin cylinders. Hydrogen bonding is shown in darker tint.
Source: own using ScPovPlot3D.

[2] SDL - Scene Description Language is Domain Specific Language (DSL) used by ScPovPlot3D Library.
[3] ex.: plane, cylinder, torus.
[4] 3D Bezier patch, isosurface, fractals,
[5] Windows™ version only

interactions turn water into key life ingredient due to number of unique, often anomalous, properties. Thus structure, properties and possible phases of water are under continuous research. Thorough discussion based on almost 2500 references is presented on WWW sub-site [24] of London South Bank University. Numerous other resources are available as well.

Number of crucial for life (more than 70) properties of water is connected with ability to create two hydrogen bondings per molecule and anomalous angle between O-H bonds (which depending on source and water phase varies in range 104.45°-109.5°). Even in liquid phase H_2O molecules may build *N*-molecule clusters, particularly when approaching melting (freezing) point. Electric field in such arrangements becomes complicated and unpredictable based on intuition. Thus water is an interesting object for testing vector field visualization procedures.

It should be stressed, that electrostatic potential and resulting electric field distribution around real water cluster is generated dynamically by electric charge whose spatial distribution is described in terms of continuous charge density function governed by principles of the quantum mechanics. While evaluation of relevant wave functions is deliberately out of the scope of this paper some of presented herein procedures are designed for visualization of externally computed or measured data.

Although in the paper classic electrostatic field is taken into account conclusions remain the same for any vector field, no matter if physical[6] (e.g. gravity or magnetic field), or just representing a property measured or computed at space points, for instance stream of tourists entering highways network during summer holidays[7] or change in incomes per capita over an area of the state.

G. *Electrostatic vector field*

Visualization of molecules is a real challenge. First of all, world in nanoscale is governed by quantum mechanics so phenomena like hybridization of orbitals [13] are often observed. In particular nanoscale objects, due to the continuity of electric charge distribution, does not sport a well defined (i.e. rigid) surface like a billiard ball so a special convention should be adopted, e.g. specific value of potential or ionic radius resulting from measurements to enable drawing "surface" of the molecule. It should be noted, that ionic radius depends on the electronic structure of an element in a given molecule. Another problem is connected with numerical calculations, namely, in equations very small values (ε_0, q, r) are placed one in denominator and the other in nominator which is a serious source of round-off errors. This requires special attention at designing source code.

Accepting assumption discussed above, calculation of static Coulomb field produced by a given molecule requires summation over all ***M*** atoms comprising it (1). For the case presented in this paper ***M***=18:

$$\vec{E}(\vec{r}) = \sum_{i=1}^{M} \frac{n_i}{4\pi\varepsilon_0} \frac{e}{\left|\vec{r} - \vec{r}_{0,i}\right|^2} \frac{\vec{r} - \vec{r}_{0,i}}{\left|\vec{r} - \vec{r}_{0,i}\right|} \quad (1)$$

where $\vec{r}_{0,i}$ is position of an '*i*-th' atom, $\vec{r}$ points at observation point, $\boldsymbol{\varepsilon_0}$ is vacuum permittivity, '***e***' denotes elementary charge and $\boldsymbol{n_i}$ is "multiplicity of elementary charge" per atom '***i***' including sign. Exemplary for oxygen O^{2-}, n_O=-2, but for H^+, n_H=+1.

Charts presented in the next part of the paper require passing scalar value (e.g. field strength) in the form of POVRay "function". However, functions in POVRay have many limitations and resembles lambda expressions known from Python rather than fully featured subroutines. Although formula (1) seems to be simple, coding it in the form of POVRay "function" is not. Because loops and temporary variables are not allowed inside POVRay functions whole definition of the function, including header and statements, has to be explicitly injected into the code of the scene at runtime. The solution is to design dedicated function factory using #macro{} syntax completing this task. The code of function factory for Coulomb potential is presented in the Fig. (2). Although code of the function factory is short, after expansion may inflate to hundreds of lines for larger structures. Computation of electric field vector $\vec{E}(\vec{r})$ may be done directly using (1) however field strength may be computed two way using first (1) and then computing resulting vector length |E| or indirectly employing built into POVRay function factory *fn_Gradient*(). For details refer POVRay documentation available from within IDE.

```
#macro CreateVFC(_PTs)  // creates Coulomb potential function from network of neutral molecules given by structure table _PTs[][]:
// The result is given in [V]
    function(x,y,z){ // use: #declare VVV = CreateVFC(PT  table) in scene code
        qeff*(_PTs[1][0]/sqrt(pow(x-_PTs[1][1],2) + pow(y-_PTs[1][2],2) + pow(z-_PTs[1][3],2))
            #for(i, 2, _PTs[0][0])
                + _PTs[i][0]/sqrt(pow(x-_PTs[i][1],2) + pow(y-_PTs[i][2],2) + pow(z-_PTs[i][3],2))
            #end
        )
    }
#end // CreateVF()
```

Figure 2. Function factory for Coulomb potential defined as macro injecting raw code into scene file at runtime. Relevant data are defined in the global table _PTs[][]. First index, i=1…N (first row for i=0 has special meaning), enumerates all molecules in the structure, and the second has following interpretation (***i***-index of the molecule): [i][0] charge, [i][1,2,3] x, y and z coordinates of observation point, [i] [4] L-J ionic radius R, [i][4], L-J radius of minimum of potential well [i][5], potential well ε [i][6]. In case Coulomb field is computed, L-Jpotential relevant data remains not used.

[6] i.e. governed by some physical principle or equation.
[7] such an analysis may help to redesign network of highways in Croatia.

Code snippets discussed in following sections are reduced to the most important lines. Full source code, required to re-compute included graphics as well as full color image gallery is available on-line [23]. Most calculations were done on HP ProBook 430 laptop.

III. VECTOR FIELD VISUALIZATION

After many years of development vector field visualization still is a demanding task [3], [4], [18], [20], [32], [33]. During that time researchers implemented many styles of different complexity and field of application. In particular one may enlist: glyphs of different shape oriented or not, hedgehogs, streamlines, streaklines, streamtubes, blob tracing, stream ribbons, isosurfaces of magnitude of the field, warping and color mappings. Majority of them, however, utilize pure geometry and plain color sometimes enhanced with simplistic transparency mainly due to computational efficiency [19]. Most often visualization toolkits allow for so called hybrid designs, i.e. designs describing two or more different properties in single image, e.g. using hedgehog style magnitude of the field may be proportional to the length of the spine while texture connected with value of potential at given space point. P(x,y,z) [18]. Most[8)] of available in VTK vector field visualization styles are incorporated into ScPovPlot3D library. Due to specific capabilities of POVRay, at least four styles may be proposed for implementation: symmetric glyphs tinted according to vector length and direction, so called "emitting media" design, variants of texturization and superhybrid design utilizing translucency. Presented raster images were computed using 1200 by 900 pixel resolution and adaptive antialiasing with supersampling[9)].

A. Glyphs (re)invented

Spatial distribution of the vector field **E**, given by a function **E**=E(r) (not necessarily continuous) is being spontaneously visualized using glyphs. "Glyphs" are simple geometric objects spherically symmetric (unoriented, e.g. spheres) or having axial symmetry (oriented - e.g. cones, cylinders, arrows) [4], [18], [20], [32]. Glyphs are implemented in variety of graphics packages, e.g. gnuplot [13] or VTK [17].

Unfortunately this design has a few shortcomings. Firstly it easily gets cluttered with increasing number of glyphs, secondly, it shows field in selected points only, thirdly, usually it is very hard to decipher true location of glyphs in 3D space from rendered 2D image as well direction and magnitude of the field, and finally designation of "hot spot" of the glyph is not so obvious.

Possible solution is based on high end 3D renderer. Thus field direction and magnitude may be encoded as components of RGB color, using primary and complementary color concept. Glyphs should be spread on the regular grid over a planar rectangle, what allows for easy location of glyphs versus visualized molecule. This solution requires excellent colorful output. Field data may be decoded from computed glyphs using graphics viewer revealing RGB components of pixel under the cursor, e.g. GIMP and then referring to the "decoding table". Stack of planes may be required to map whole 3D vicinity of the molecule. Optionally color distribution may help spot regularities of the field distribution. Exemplary visualization in grayscale is shown in the Fig. 3.

B. Media design

Another possible visualization style is based on pseudoparticle idea. "Particles" in 3D computer graphics, implemented as "media" in POVRay, simulate presence of 0-sized objects scattering, absorbing or emitting light. These objects may get specific color due to programmer's demand. Similar paradigm has been implemented in very old program HPOV [14], designed for visualization of hybridized electron orbitals using spherical harmonics.

While media calculations, especially scattering, in POVRay are generally rather slow, computation of tiny slice of 3D space using emitting media only is relatively fast. Hence visualization is reduced to internal point calculations now. Important drawback of this design is restrictions put on media by Scene Description Language syntax. Generally, whole process is programmed using emitting media filling a probe object, e.g. flattened box. Density of the media must be passed to the media description as POVRay function what sometimes may be difficult to accomplish. Specially crafted color scales allows for fast and precise visualization of 3D vector field magnitude distribution. Exemplary image is shown in the Fig.4. In this case very flat box was used as probe object. However by use of larger containers and compact color scales the shape of constant field magnitude (or scalar vector field derivative, e.g. divergence), may be shown.

C. Isogradient textured by potential

Usually texture mapping is applied to simple probe objects like planes or boxes. This section is devoted to texture mapping of "isosurface" or "equipotential surface" if the potential field (e.g. electrostatic) is under consideration. POVRay offers object dedicated for this purpose called *isosurface*{} [16], whose definition is

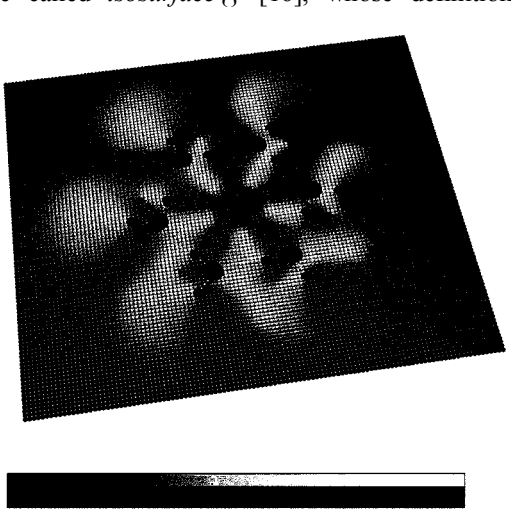


Figure 3. Revealing distribution of vector field using unordered glyphs. Direction and magnitude of the field is encoded in RGB components of glyph's color. Symmetry of the field is clearly visible. Computation time: 29s. Source: own.

[8] excluding warping and solid ribbons.

[9] POVRay command line was: +UA +W1200 +H900 +FNG +A +AM2 +R5 +TH. For details on switches refer to documentation.

described in [21] as well as details of usage.

Isosurface visualization style may be simple or hybrid. In the first case only shape of equipotential surface is computed and texture may be used to make surfaces with different threshold values distinct. Thus a few surfaces may be shown in a single figure.

In the second case texture may be calculated dynamically using, for example, value of the gradient of the field over the surface. That way two different functions may be visualized at the same time as discussed in [21]. In this paper another attempt is made, namely the try to texture isosurface of field magnitude by value of potential. The result is shown in Fig.5. Computation time was almost 3 hours. This problem may be overcome using cellular trilinear interpolation method [30], [31] and [21].

Computation of isosurface requires repetitive calculations of field magnitude for every pixel of output image even for empty space what is very inefficient. This may be solved by splitting whole search box into tiny cells, then rejecting from analysis those where no piece of isosurface can be found.

Trilinear interpolation function is implemented now in ScPovPlot3D using refactored formulation[10] (2) [30], [31], relevant geometry and symbols are displayed in the Fig.6. Because (2) uses only simple arithmetic, i.e. additions and multiplications, it is computed relatively fast. However, computation of values for all nodes of the computational grid inside POVRay takes a long time due to limitations of its scripting language. Such a grid should be then computed using fast workstation or, perhaps, supercomputer running optimized command line driven parallelized program, and then imported to POVRay.

Test image (isosurface of gradient tinted by potential value) presented in the Fig. 6, has been rendered using 50×50×16 cell grid, hence potential value in 40000 nodes was calculated. Because algorithm is approximated, some shape irregularities may be spotted on the calculated surface what will be addressed in the future research.

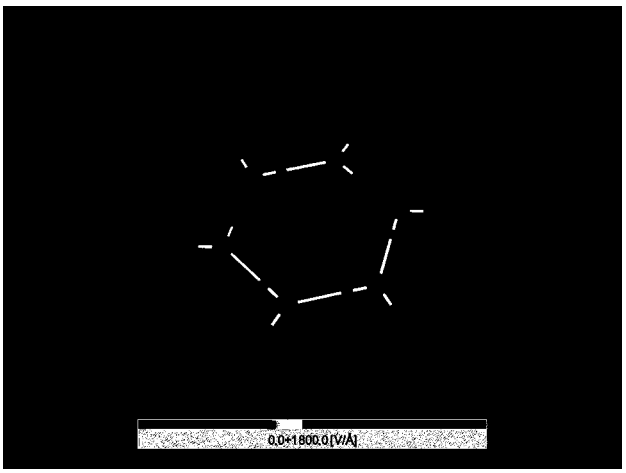


Figure 4. Media representation of field magnitude distribiution. Thanks to implemetation of translucency, the water cluster, ie. source of the field is not fully obscured. Computation time was 35s. Source: own.

[10] Presently defining formula is explicitly symmetrical versus values in corners of an elementary box.



Figure 5. Hybrid representation of constant field magnitude surfaces textured after potential value. three surfaces were computed, for 1, 10 and 20 [V/Å]. Computation time: 2h 52m 36s. Source: own.

$$w(i,\lambda_x) = \begin{cases} 1-\lambda_x & ,i=0 \\ \lambda_x & ,i=1 \end{cases}$$

$$Q(\lambda_x,\lambda_y,\lambda_z) = \tag{2}$$

$$= \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} w(i,\lambda_x)w(j,\lambda_y)w(k,\lambda_z)\cdot q_{ijk}$$

D. Superhybrid design, fluxlines

The term superhybrid design means stacking at least three information layers in the single image. In the image shown in Fig.8. the field is described using tubelines, textured according to decimal logarithm of absolute value of potential; diameter reflects constant value of electric field flux, $\Phi_E = \vec{E}\cdot S\cdot\vec{n}$, however clipped at lower and upper limits to keep the picture clear. This genre of tubelines deserves thus name **fluxline**. Finally, three mutually perpendicular planes, parallel to the Cartesian coordinate system planes but shifted, are textured by

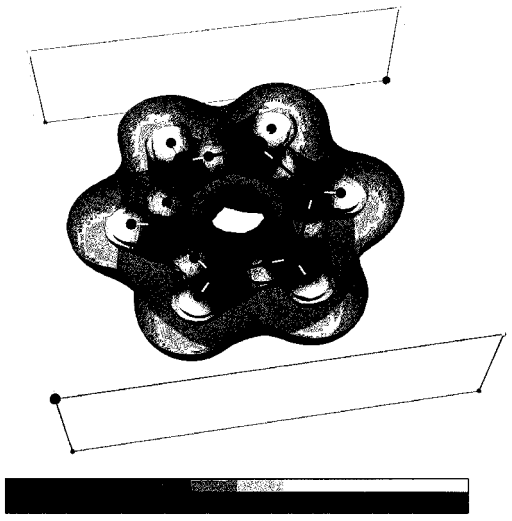


Figure 6. Hybrid representation of constant field magnitude surfaces textured after potential value. Three surfaces were computed, for 20, 10 and 1 [V/Å] (inside-out). Shown is bounding box as well. Computation time: 0h 1m08s. Source: own.

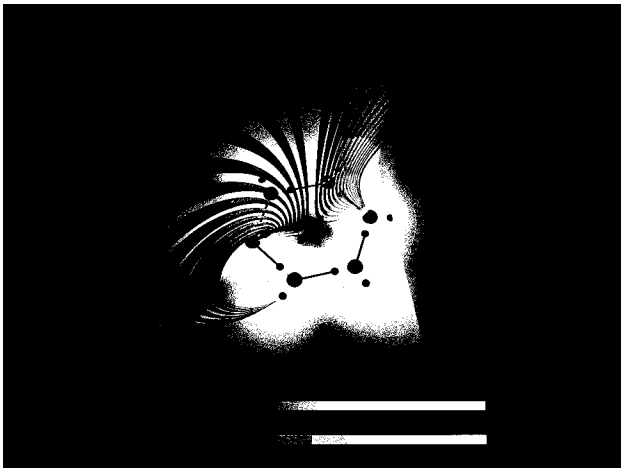


Figure 8. Six fold superhybrid representation of electrostatic field. Fluxlines follows vector field and are textured by decimal logarithm of absolute value of potential (lower scale), diameter is inversely proportional (and clipped) to the square root of strength of the field (upper scale), texture on three perpendicular planes follows decimal logarithm of field magnitude (upper scale again). Flux lines were computed for Φ=0.01 [V*Å]. Computation time: 14min39s. Source: own.

decimal logarithm of field magnitude again, however this time texture map introduces general insight into field symmetry. Texture used on planes is partially translucent so fluxlines remains partly visible behind them, which may be used as 7th information layer.

IV. Conclusion

High quality 3D render program equipped with efficient scripting language and supplementary API may be used for fast prototyping of complex visualization styles. Here elaborated designs should be incorporated into production toolkits like VTK.

In order to improve numerical efficiency external program should be used in selected cases. API of ScPovPlot3D package is ready for visualization of data computed externally.

Due to relatively short computation time texture map style, fluxlines with derivatives and media design are preferred for fast introductory analysis of new structures.

References

[1] Y-M. Wang, N.R. Jr Sheeley, A.G. Nash, *Latitudinal distribution of solar-wind speed from magnetic observations of the Sun,* Nature 347, 439 - 444 (04 October 1990); doi:10.1038/347439a0,

[2] A. Lampert, B. Bernalte Jimenez, G. Gross, D. Wulff, T. Kenull (2015), *One-year observations of the wind distribution and low-level jet occurrence at Braunschweig, North German Plain*. Wind Energ., doi: 10.1002/we.1951.

[3] R. Crawfis, N. Max, B. Becker, B. Cabral, "*Volume rendering of 3D scalar and vector fields at LLNL,*" in Supercomputing '93. Proceedings , vol., no., pp.570-576, 15-19 Nov. 1993, doi: 10.1109/SUPERC.1993.1263506,

[4] C.M. Wittenbrink, A.T. Pang, S.K. Lodha, *Glyphs for visualizing uncertainty in vector fields*, in Visualization and Computer Graphics, IEEE Transactions on , vol.2, no.3, pp.266-279, Sep 1996, doi: 10.1109/2945.537309,

[5] Blender, v.2.76b, www.blender.org, accessed 2016-02-15,

[6] BioBlender, http://www.bioblender.eu/, accessed 2016-02-15,

[7] *Diderot* http://diderot-language.cs.uchicago.edu, acc. 2015-01-27

[8] G. Kindlmann, C. Chiw, N. Seltzer, L. Samuels, J. Reppy, *Diderot: a domain-specific language for portable parallel scientific visualization and image analysis*, URI: http://people.cs.uchicago.edu/~glk/pubs/pdf/Kindlmann-Diderot-VIS-2015.pdf, accessed 2016-02-15,

[9] RasMol, http://www.openrasmol.org/, accessed 2016-02-15,

[10] Shadie project, https://code.google.com/p/shadie/ (followed from Diderot home page) accessed 2015-02-04,.

[11] M. Hašan, *Short but comprehensive introduction to Shadie*, http://miloshasan.net/Shadie/shadie.pdf, accessed 2015-02-04,

[12] M. Hašan, *Volume rendering of dosimetric distribution and biological response from 3D/4D treatment and delivery*: http://miloshasan.net/Shadie/aapm-talk.zip, accessed 2015-02-04

[13] gnuplot, http://www.gnuplot.info, accessed 2016-02-08,

[14] A. Paltanavi-Źius, *Hybridized orbital preview*, HOPV.exe, 2006, http://www.adomas.org/hopv/, accessed 2015-02-09,

[15] ParaView, http://www.paraview.org/, accessed 2016-02-08

[16] POVRay, http://povray.org, accessed 2016-02-08,

[17] VTK, *Visualization Toolkit*, http://www.vtk.org/, accessed 2016-02-02,

[18] B. Geveci, *An introduction to visualization using VTK. Vector visualization*, Vis 2004, URI: http://goo.gl/nc2YGw, accessed 2016-02-16,

[19] *How to display transparent objects (in VTK)*? URL: http://goo.gl/O0zkHq, accessed 2016-02-07,

[20] CMISS: *Visualizing fields at points using glyphs*, URL: http://goo.gl/dUPp7L

[21] J. Opila, *Prototyping of visualization styles of 3D scalar fields using POV-Ray rendering engine*, in: 38th Int. Conv. on Inform. and Comm. Technology, Electronics and Microelectronics (MIPRO), 2015, doi: 10.1109/MIPRO.2015.7160285

[22] J. Opiła, *ScPovPlot3D POVRay templates,* http://scpovplot3d.sf.net, accesed 2016-02-08,

[23] http://sourceforge.net/projects/scpovplot3d/files/Dev/4MiPro39.zip/download, accessed 2016-02-08,

[24] M. Chaplin, *Water structure and science*, http://goo.gl/7B1WXJ, accessed 2016-02-18,

[25] I. Skalna, B. Rębiasz, B. Gawel,, B. Basiura, J. Duda, J. Opiła, T. Pełech-Pilichowski, *Advances in fuzzy decision making*, Springer, ISBN: 1434-9922, DOI: 10.1007/978-3-319-26494-3,

[26] J. Opiła, "*Business data visualization using ScPovPlot3D, POVRay based toolkit*", Wybrane zastosowania technologii informacyjnych wspomagających zarządzanie w organizacjach,, 2015, Wyd. Politechniki Częstochowskiej, Edt. L. Kiełtyka, R. Niedbał, ISBN: 978-83-7193-626-5,

[27] J. Opiła, *Some remarks on visualisation of irregular or dispersed data* (pol. *Kilka uwag o wizualizacji danych nieregularnych lub rozproszonych*) in: Diagnozowanie stanu środowiska, red. J. K. Garbacz. Bydgoszcz: BTN, 2012, ISBN: 978-83-60775-34-9,

[28] J. Opiła, I. Skalna, T. Pełech-Pilichowski, *Interval arithmetic for irregularly distributed data visualisation by kriging* in: Enterprise Management-Theory and Practice, XIV International Science. Conference.: 22–23 Nov. 2012, Kraków,

[29] J.E. Lennard-Jones,, *On the determination of molecular fields*, Proc. R. Soc. Lond 1924, A 106 (738): 463–477,

[30] P. Bourke, http://paulbourke.net/miscellaneous/interpolation, accessed 2015-03-23,

[31] R. Wagner, *Multi-Linear interpolation*, http:\\bmia.bmt.tue.nl\people\BRomeny\Courses\8C080\Interpolation.pdf, accessed 2015-03-23

[32] E. McQuinn, A. Chourasiaba, J.P. Schulzeda, J.B. Minsterca, *GlyphSea:visualizing vector fields*, (2013), URI: http://goo.gl/E71dlq, accessed 2016-02-08,

[33] M. Bailey, *Vector visualization*, Oregon State University, March 2015, URI: http://goo.gl/DlQbNk, acc. 2016-02-08,

New Cybercrime Taxonomy of Visualization of Data Mining Process

M. Babič*, B. Jerman-Blažič*
* Laboratory for Open Systems and Networks, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
babicster@gmail.com

Abstract - Data Mining is the process of identifying new patterns, insights in data and knowledge discovery, and is at the intersection of multiple research areas, including Machine Learning, Statistics, Pattern Recognition, Databases, and Visualization. With the maturity of databases and constant improvements in computational speed, data mining algorithms that were too expensive to execute are now within reach. Data visualization is a general term that describes any effort to help people understand the significance of data by placing it in a visual context. Patterns that might go undetected in text-based data can be exposed and recognized easier with data visualization software. Exploring and analyzing the vast volumes of data becomes increasingly difficult. Information visualization and visual data mining can help to deal with the flood of information. There is a large number of information visualization techniques which have been developed over the last decade to support the exploration of large data sets. In this paper, we propose a classification of information visualization and visual data mining techniques. Fractals and graph theory are very popular in different areas. We present a new method for estimating fractal dimension for network and new Taxonomy of visualization of data mining process application in Cybercrime activity.

Keywords: Visual Data mining, Network, Modeling, Fractal geometry, taxonomy, Cybercrime,

1. Introduction

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all areas of human endeavor, from the mundane to the more exotic. Little wonder, then, that interest has grown in the possibility of tapping these data, of extracting from them information that might be of value to the owner of the database. The discipline concerned with this task has become known as data mining. The fundamental algorithms in data mining and analysis form the basis for the emerging field of data science, which includes automated methods to analyze patterns and models for all kinds of data, with applications ranging from scientific discovery to business intelligence and analytics. Data mining is the process of discovering insightful, interesting, and novel patterns, as well as descriptive, understandable, and predictive model s from large-scale data. Data mining is an interdisciplinary exercise. Statistics, database technology, machine learning, pattern recognition, artificial intelligence, and visualization, all play a role. And just as it is difficult to define sharp boundaries between these disciplines, so it is difficult to define sharp boundaries between each of them and data mining. At the boundaries, one person's data mining is another's statistics, database, or machine learning problem.

Statistical techniques alone may not be sufficient to address some of the more challenging issues in data mining, especially those arising from massive data sets. Nonetheless, statistics plays a very important role in data mining: it is a necessary component in any data mining enterprise. In this section we discuss some of the interplay between traditional statistics and data mining. With large data sets (and particularly with very large data sets) we may simply not know even straightforward facts about the data. In summary, while data mining does overlap considerably with the standard exploratory data analysis techniques of statistics, it also runs into new problems, many of which are consequences of size and the non traditional nature of the data sets involved.

Pattern recognition aims to make the process of learning and detection of patterns explicit, such that it can partially or entirely be implemented on computers. Automatic (machine) recognition, description, classification (grouping of patterns into pattern classes) have become important problems in a variety of engineering and scientific disciplines such as biology, psychology, medicine, marketing, computer vision, artificial intelligence, and remote sensing. In almost any area of science in which observations are studied but the underlying mathematical or statistical models are not available, pattern recognition can be used to support human concept acquisition or decision making. Given a group of objects, there are two ways to build a classification or recognition system (Watanabe 1985), supervised, i.e., with a teacher, or unsupervised, without the help of a teacher. Interest in pattern recognition has been renewed recently due to emerging applications which are not only challenging but also computationally more demanding, such as data mining, document classification, organization and retrieval of multimedia databases, and application for prevent of cybercrime as biometric authentication (i.e., face recognition and fingerprint matching).

Data visualization is a quite new and promising field in computer science. It uses computer graphic effects to reveal the patterns, trends, relationships out of datasets. Human has a long history with basic data visualization, and data visualization is still a hot topic today. The history of visualization was shaped to some extent by available technology and by the pressing needs of the time, they include: primitive paintings on clays, maps on walls, photographs, table of numbers (with rows and columns concepts), these are all some kind of data visualization – although we may not call them under this name at that time. Visualization is the graphical presentation of information, with the goal of providing the viewer with a qualitative understanding of the information contents. It is also the process of transforming objects, concepts, and numbers into a form that is visible to the human eyes. When we say "information", we may refer to data, processes, relations, or concepts.

In computer and network science, network theory [5] is the study of graphs as a representation of either symmetric relations or, more generally, of asymmetric relations between discrete objects. Network theory is a part of graph theory. It has applications in many disciplines including statistical physics, particle physics, computer science, electrical engineering, biology, economics, operations research, climatology and sociology. Applications of network theory include logistical networks, the World Wide Web, Internet, gene regulatory networks, metabolic networks, social networks and epistemological networks.

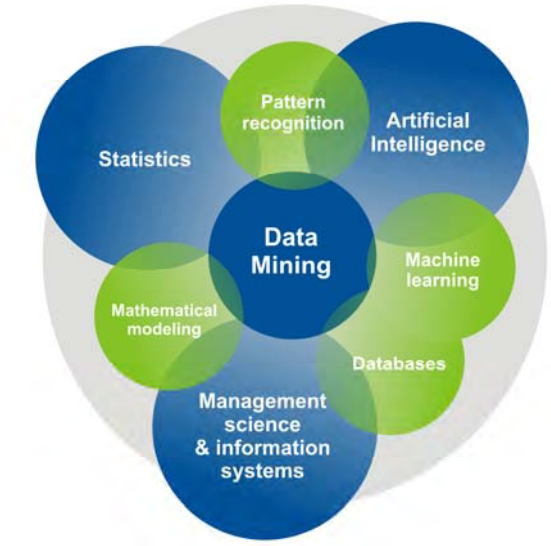


Fig. 1: Data Mining present as an interdisciplinary discipline encompassing a blend of statistical, artificial intelligence, and management science & information systems disciplines for pattern recognition, mathematical modeling, and databases activities

In this article we present new cybercrime incident taxonomy of visualization of data mining process.

2. Experimental method

We use data mining as an interdisciplinary discipline encompassing a blend a blend of statistical, artificial intelligence, and management science & information systems disciplines for pattern recognition, mathematical modeling, and databases activities. After this, we use network theory to present taxonomy of cyber incident. On Fig. 2 is presented incident of cyber crime.

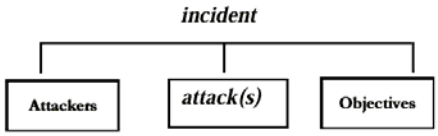


Fig. 2. Incident of cyber crime include attackers, attack(s) and objective

A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Dendrograms are often used in computational biology to illustrate the clustering of genes or samples. The agglomerative hierarchical clustering algorithms available in this program module build a cluster hierarchy that is commonly displayed as a tree diagram called a dendrogram. They begin with each object in a separate cluster. At each step, the two clusters that are most similar are joined into a single new cluster. Once fused, objects are never separated. The horizontal axis of the dendrogram represents the distance or dissimilarity between clusters. The vertical axis represents the objects and clusters. The dendrogram is fairly simple to interpret. Remember that our main interest is in similarity and clustering. Each joining (fusion) of two clusters is represented on the graph by the splitting of a horizontal line into two horizontal lines. The horizontal position of the split, shown by the short vertical bar, gives the distance (dissimilarity) between the two clusters. Looking at this dendrogram, you can see the three clusters as three branches that occur at about the same horizontal distance. Fig. 3 is can we present as dendrogram and dendrogram presented as network.

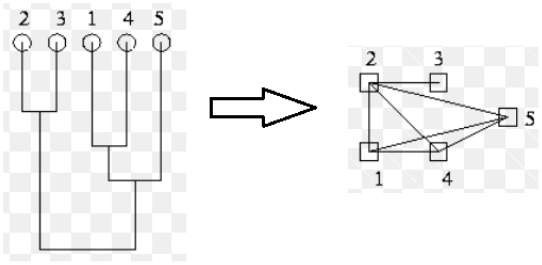


Fig. 3: Dendrogram presented as network

In network we calculated topological property, density. Network density" describes the portion of the potential connections in a network that are actual connections. A

"potential connection" is a connection that could potentially exist between two "nodes" regardless of whether or not it actually does. The density *D* of a network is defined as a ratio of the number *N* of edges *E* to the number of possible edges, giving

$$D=2E/N\times(N-1).$$

3. Results and discussion

Hierarchical clustering is one method for finding community structures in a network. Fig. 4. Present Computer and network cyber incident taxonomy. Incident include attackers, tools, vulnerability, action, target, unauthorized result and objective.

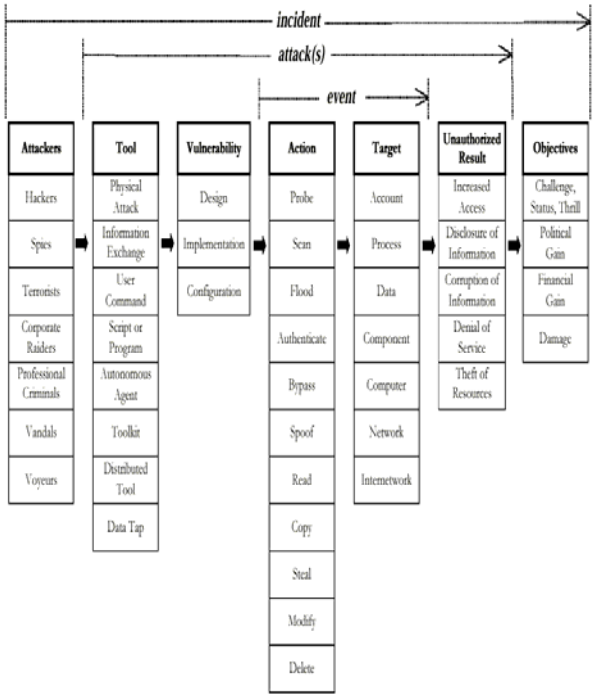


Fig. 4. Computer and network cyber incident taxonomy

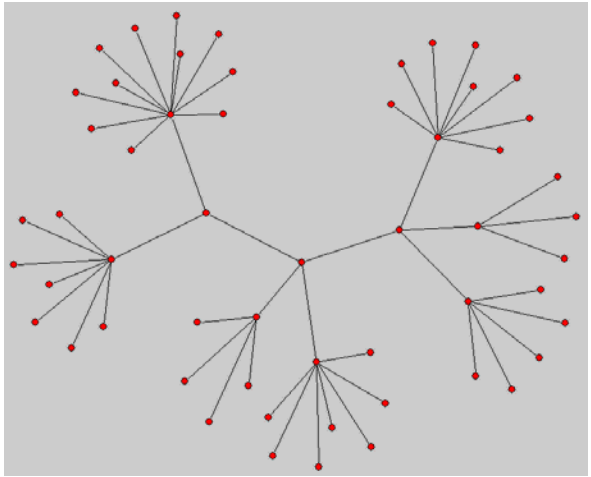


Fig. 5: Computer and network cyber incident taxonomy presented as network

On Fig. 5 is presented Computer and network cyber incident taxonomy presented as network. On this network we calculated topological properties density. Density of network is 0.03570248. In this paper, we presented a Systemic Taxonomy of Cyber incident and its network research. Our taxonomy of Cyber incident was an extended taxonomy of existing taxonomies of Cybercrime. A dendrogram of Cyber incident was presented with method graph theory. We described Cyber incident with topological properties of network density. Density present complexity of network.

4. Conclusion

This paper has contributed to the conceptual and empirical understanding of global cyber wars and crimes. Different motivations of hackers, source characteristics and target country characteristics lead to different likelihoods of attacks on different organizations. Timely reporting of cyber attacks to authorities is thus likely to strengthen the rules of law and help combat cyber threats in the long run. Cooperation and collaboration among national governments, computer crime authorities and businesses are critical to combat cyber attacks. If national governments work with one another as well as with business communities to modify institutions by defining appropriate policies for the security of the digital world, it will result in lower transaction costs. With topological property density we describe Computer and network cyber incident taxonomy.

5. Literature

[1] Shukla, D., Verma, Kapil, Dubey, Jayant and Gangele, Sharad, (2012 c): Cyber crime Based Curve Fitting Analysis in Internet Traffic Sharing in Computer Network, International Journal of Computer Application (IJCA), Vol.46(22), pp. 41-51.

[2] Gangele, Sharad, Verma, Kapil and Shukla, D.,(2014): Bounded Area Estimation of Internet Traffic Share Curve, International Journal of Computer Science and Business Informatics (IJCSBI), Vol. 10, No. 1, pp. 54-67.

[3] Shukla, D., Verma, Kapil and Gangele, Sharad, (2012 a): Iso-Failure in Web Browsing using Markov Chain Model and Curve Fitting Analysis, International Journal of Modern Engineering Research(IJMER) , Vol. 02, Issue 02, pp. 512-517.

[4] Shukla, D., Verma, Kapil and Gangele, Sharad, (2012 b): Least Square Curve Fitting Applications under Rest State Environment in Internet Traffic Sharing in Computer Network, International Journal of Computer Science and Telecommunications (IJCST), Vol. 03, Issue 05, pp. 43-51.

[5] SWAMY, M.N.S. & THULASIRAMAN, K.:Graphs, Networks, and Algorithms.Wiley (1981).

Visual representation of predictions in software development based on software metrics history data

Blažo Popović*, Adis Balota * and Dženan Strujić **
* Faculty for Information Technology, University Mediterranean, Podgorica, Montenegro
** Nilex Development, Nilex AB I Helsingborg, Helsingborg, Sweden
popovic.blazo@gmail.com, adis.balota@unimediteran.net, dzenan.strujic@nilex.se

Abstract - Software who is being developed for more than a year, is categorized in the group of large projects, and is entering the area of critical and risky in terms of their successful completion. Because of those reasons, they require constant monitoring by the software manager. In order to perform adequate monitoring of developed solution and its projects, there are a number of software metrics that provide these information in numerical form, but very few of them is displayed to software manager in visual form. Predictions for future solution development, and its visual representation, based on historical data gathered from the same system, are usually not included in such systems. These kind of systems must have the ability to predict state of the developed system in the future, and adapt its output upon having analyzed his software metrics history data. In this paper, we propose a intelligent system that will analyze software metrics history data of the solution, and make predictions and visual outputs to support software manager decisions. The results show that using this decision support tool, as a form of intelligent system, helps software managers in their decision making during project management, and in reducing overall project risk.

I. INTRODUCTION

One of the most important tasks of the software manager in project management is risk management throughout all stages of its development [1, 2]. In order to successfully monitor the risk in all individual phases of the software product development it is necessary to have appropriate measures, in respect of which it is possible to quantify the size of the risk [1, 2]. These measures are in software engineering reflected in the form of software metrics.

In today's software engineering, there are a number of development tools and environments that are used for software design and software metrics presentation [5, 17-20]. Those tools helps the software manager significantly in project management. Graphic representation and visualization of software metrics parameters offers the possibility that large amount of numbers that is not in systematic form, become shaped, categorized and graphically or visually expressed.

With development of tools for graphical representation of software metrics parameters, there is a possibility for more efficient risk management and control of software projects [2]. Previous approaches to solving the aforementioned problems give limited results [8-16], and that fact requires that software managers use one or more tools for software metrics representation and analysis [2], in order to see the state of the developed software code. In addition, these tools and models do not offer fully graphical interpretation of its numbers in a way that meets the demand of software managers in real-time.

Recognizing the need of software manager in real-time for risk management in project they lead, there is an open possibility to create decision support system in form of an prototype, intelligent rule-based expert system. This intelligent system will have the ability to analyze software metric history data, represent them in the visual form, and make predictions for future software development based on this data. With this system that analyses history software metric data gathered from the system that is being developed, software managers will have opportunity to anticipate and cope with the risk they face, and more easily predict future software development on the project based on it.

Paper has the following structure. In section II, we will give an overview of software metrics in enterprise systems, and introduce the metrics that will be monitored, collected and used in proposed rule-based expert system. Section III will make an overview of visualization techniques used for graphical representation of data, and give the details of the visualization tool that we will use in synergy with our rule-based expert system. In section IV we will introduce prototype of proposed rule-based expert system, its architecture, knowledge base rules, facts, inference engine, explanation facility, with its prediction and visualization capabilities. Section V will give the discussion of the proposed solution, results, and limitation of the same. Paper ends with conclusion in section VI.

II. SOFTWARE METRICS IN ENTERPRISE SYSTEMS

Software metrics is of great importance in the life of development of a software product. Using software metrics in various stages of development of software product can provide a clear picture to the software manager about state and condition of project that he leads. Additionally, software metrics provides information about software product, process, what kind of effort is invested, what effort should be made to complete the launched project or product, what is the quality of code that

developers are producing, and finally, he gets the information needed to guide the process of software testing.

Software metrics must be used in project management so that it can be properly performed, and because of risk that he wears with itself. Appropriate software metrics will provide inputs to software managers to make decisions about how to further advance the development of software products. There are a large number of categories in which one can classify software metrics. One of these categorizations is done in [1, 2].

Two basic software metrics are SLOC and function points. The former, in its purist form, counts the number of the typed code lines, and later represents functional user requirements in numerical form as function points [3]. In enterprise projects, these metrics is not suitable because software is written in object oriented manner. Because of that fact, appropriate object oriented software metrics has to be employed. The most widely used set of object oriented measures is proposed by Chidamber and Kemerer [2, 4], and those measures include: WMC (weighted methods per class), DIT (depth of inheritance tree), NOC (number of children), CBO (coupling between object classes) and RFC (response for a class).

Since focus of our paper are enterprise projects developed in Microsoft Visual Studio 2013 development environment, we will observe enterprise metrics that is produced by its plug in Code Metrics Viewer [5]. Code Metrics Viewer offers following metrics to developers and software managers: maintainability index (MI), cyclomatic complexity (CC), depth of inheritance (DIT), class coupling (CBO), and lines of code (LOC). Their values and interpretation can be found in [6].

Following the good software manager practice, every software managers want their developers to produce code following their established coding standard, that will guarantee high-quality software components, from methods to custom types. In order to establish such routine and discipline, Code Metrics Viewer gives to developers the possibility to measure their code quality, so that software manager is satisfied with code quality and coding standard. This is also true for software manager, who can view, collect and analyze software metrics data from projects and solution in numerical form. In large enterprise projects, number of produced code lines grows constantly. Number of methods per project and per solution, including number of developed types by project and per solution grows day by day. In our previous work [7], we have demonstrated that growth of the methods and types per project and per solution in enterprise development does not increases its complexity, does not make it more difficult for maintenance, testing, or refactoring. This was all true if software manager is aware of the need for employment of adequate software metric and need for established coding standard for developers. Rules in coding standard employed MI, CC and LOC values of software metric that could not be broken by developers during software development.

However, today software metric tools do not provide means to try to predict the development process based on collected software metrics history data. This is the first important fact in which lies the idea, and basis for building for our prototype intelligent rule-based expert system that we will propose in section IV.

III. SOFTWARE VISUALIZATION TECHNIQUES

The field of software visualization is very broad. Visualization of software has found its application in everyday software engineering, various algorithm creation and representation, presentation of object-oriented code, software diagnostic, just to name a few.

Software visualization has advanced from two-dimensional graphics, over three-dimensional representation [8], up to virtual environments [9]. The biggest challenge in the field of software visualization is how to represent object oriented way of thinking, that is, how to represent classes, objects, interfaces, inheritance, polymorphism and other object oriented pillars. For this particular matter, there has been a significant amount of research and propositions in the science filed, and some of the ideas can be observed in [8, 10-16]. The bottom line of all this work is to find a way to represent software in visual way, so that software managers can visually analyze the state of the developed software product, and the quality of the produced software code.

Following the science field in software visualization, the software industry and practice developed tools for representing software visualization. In the domain of the C# programing language and its natural development environment Microsoft Visual Studio 2013, these tools include: Atomique [17], Nitriq [18], CodeIt.Right [19], NDepend [20]. All these tools have their own way of analyzing software code based on software metrics, and represent it in a visual or graphical form. However, it has been noticed that none of them includes any kind of prediction module for future software development in the developed solution.

Further enhancement and help in graphical representation and visualization on web sites and web pages has been brought by several graphical libraries. These libraries include: D3.js [21], FusionCharts.js [22], jqPlot [23], Google Charts [24], and Highcharts [25], just to name a few. In our proposed solution, we will use Highcharts library for drawing graphics. All mentioned libraries have the ability to draw different kinds of charts to web pages, tablets or mobile phones. These charts include: line charts, bar charts, area charts, column charts, pie charts, scatter charts, bubble charts, synchronized charts, dynamic charts, 3D charts, polar charts, spider web charts, wind rose charts, box plot charts, waterfall charts, gauges, and in particular libraries, heat and tree maps. Above listing of options in libraries makes them the perfect ground for use in visual representation of software metrics history data.

It can be brought to notice that all of these software tools and products do not offer the possibility to try to visualize the predictions of future software development based on software metrics history data. This is the second important fact for building our prototype intelligent rule-based system.

IV. PROPOSED SOLUTION

In our previous work [7], we demonstrated that usage of software metrics in enterprise development is of crucial importance in success of leading, monitoring and evaluating quality in software development, which in turn gives the software manager the precise control of the developed project and solution in whole. Further more, since software metrics gives information what is going on with the developed system, it of great importance for software manager to collect it, analyze it, and make business decisions after that. In the same work, we proposed software metrics that had to be used in the developed systems, which represented the development standards that had to be satisfied by all developers. After collecting software metrics history data from the system, we made the analysis and proved that development standards were respected by the developers, and there was no increase of complexity of the system aside from fact that solution was growing day by day in size of number of new methods and types per project, and per whole solution. The analysis in our work [7] did not include any kind of future software development predictions or recommendations.

The solution in question was developed by 50 developers that were spread all over the world. The leading software manager was not physically present all the time in person for questions and answers about future development decisions and predictions for all other development team leaders, who were leaning on his decisions between development iterations. Since the solution has been developed for more than a one year, it has grown so large in size that controlling and monitoring of the solution had to be delegated to other team leaders, aside from leading software manager. In order to do that, the expertise and experience of the leading software manager had to be caught in some form so that it could provide answers and guidelines to developers team leaders regarding future software development. The solution was found in our proposition to develop of prototype rule-based expert system, who had to catch and represent the logic of human expert, and help in decision making for other team leaders. Since the developers were geographically dispersed, we proposed that the system had to be a web based solution, so that every team leader could easily connect online on system and find the answers that he needed.

After narrowing the problem domain, the requirements for the expert system were elicitated. The proposed prototype of our intelligent system had to have ability to give state of the developed projects and overall solution based on collected software metrics history data. After making the analysis of software metrics history data, they had to be visualized on the web page with some form of visualization technique, where we proposed Highcharts library. The proposed system have to make predictions for next day based on performed analysis on software metrics history data, and number of days before current iteration ends. Prediction of the software development for the next day had to be represented in visual form so that team leader can visually conclude what will probably happened in tomorrow software development, without the need to check the numbers by hand. Proposed intelligent system must have the ability to explain how it has reached its decision on demand by team leader. Ultimately, the proposed expert system must give the visual output to team leader about current level of the risk in developed project in question, after making the predictions for software development.

After analyzing all the requirements for the intelligent system, and knowing the fact that most of the developers in company are using programing language C# to develop the solution, we proposed a solution that is based on Microsoft technologies. Development environment is Microsoft Visual Studio 2013. Since developed solution had to be a web application, it is based on ASP.NET MVC 4.5 and ASP.NET WEB API 2 frameworks. Database system for the intelligent system will be Microsoft SQL Server 2012 Express Edition. Our choice to go for ASP.NET WEB API 2 framework is made from gathered requirements for intelligent system, and its capabilities for creating RESTfull web services in easy manner, using ordinary http verbs: GET, POST, PUT and DELETE.

Rule-based systems and web-based systems are important part in theory of intelligent systems and artificial intelligence [26, 27], and have found their usage in lots of important works [28-34]. In our solution, we had to make modification and use combination of both in order to create solution that will fit the needs of this real life situation.

The proposed architecture of prototype intelligent rule-based expert system is shown in Figure 1.

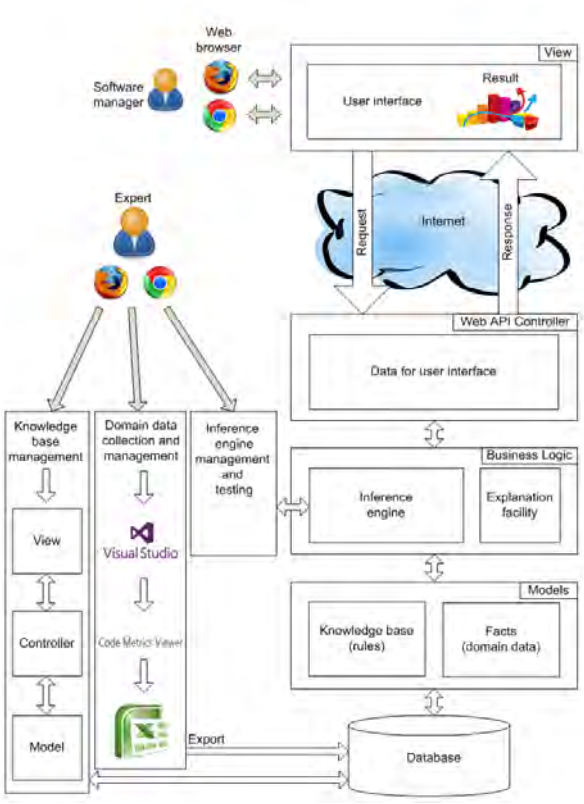


Figure 1. Architecture of prototype (web) rule-base expert system

A. Knowledge engineering

In order to make our prototype of intelligent system to work, knowledge for it had to be acquired from human expert, which in our case was the leading software manager. Capturing the experience, knowledge and reasoning of software manager was done using observations during his work and asking questions how did he reach the decision, interviewing him in situations that required his call, and discussing the various case studies that were appearing during software development process. Knowledge was then represented as a set of typical if-then rules, that could be used later in inference process. Human expert has the ability to insert, edit and test the rules, so that intelligent system can represent his logic as closely as possible. For this purpose, human expert uses web browser as his interface when working on knowledge engineering.

B. Collecting of Software Metrics History Data

Second important part of our intelligent system is ability to gather software metrics history data from the developed software solution, so that it can be used in the inference process. The production solution is being developed in Microsoft Visual Studio 2013 environment using C# programming language. Leading software manager is using Code Metrics Viewer plug-in for this environment, in order to track the quality of developed software code, represented in numerical form. The process of collecting software metrics history data begins with capturing software metrics data in numerical form represented in Code Metrics Viewer, upon which they are exported to the Microsoft Excel file. Each working day must be represented as one file, which contains the measured software metrics data from the production solution. After the files have been created, the files are being parsed and data is exported to SQL Server database, consolidated and aggregated. In this way, the database has the necessary data, in expert system terms – the facts, that are going to be used in inference process.

C. Inferencing in our prototype intelligent system

Coding standard is the very important in development of software systems. As it has been shown in our previous work [7], proposed parameters for software metrics give the expected results and move the development of the solution in desired solution, without the rise of the overall solution risk growth. The metrics in question is maintenance index (MI), cyclomatic complexity (CC) and lines of code (LOC). Proposed parameters of this metrics are: MI lower limit 50, CC maximum 10, and LOC upper limit 20 and critical limit more than 25. Only external indicator is end of iteration in days, which will help in predictions for software development.

Inferencing mechanism in our system is forward chaining technique, which means that it is using knowledge based rules and data to reach the decision, in our case, the prediction of future software development.

One example of inferencing for a particular project in production solution, based on its software metrics history data is represented in following manner:

IF Quality_MI = High
AND Quality_CC = Low
AND Quality_LOC = Low
THEN CodeQuality = Low

IF IterationEndsInDays > 0
AND IterationEndsInDays < 8
THEN IterationInState = Ending

IF CodeQuality = Low
AND IterationInState = Ending
AND CodeQualityIndicator = “Required”
THEN Prediction is “Refactor”
AND Effort is “Increased”
AND Productivity is “Increased”
AND PotentialForIterationBreach = “High”
AND Risk = “High”

Human expert has the ability to train the inference engine, in order to help the intelligent system to reach his decision making behavior.

D. Explanation facility

Inference engine is hidden from end user of the intelligent system. The human expert is the only one who has the ability to view and modify the rules in knowledge base and to test the inference engine how it reaches its decision. For end user of the system, it is often needed to explain how the system reached its solution that he predicted. In our proposed system beside the inference engine is standing the explanation facility, which will follow the rules that were fired by the inference engine, and based on them will generate most possible human language representation of the reached solution.

Returning to the example in previous section, the explanation facility would give back to end user following text:

“*Known facts for last 7 days: Average code standard for maintenance index is High. Average code standard for cyclomatic complexity is Low. Average code standard for lines of code is Low. Requirements for this iteration: Coding standard. Conclusion: Coding standard violated. Prediction: Action is refactor solution. Effort will be increased. Productivity will be increased. Iteration deadline breach potential is High. Risk is high.*”

Language variables like “refactor”, “increased”, “lower”, “high” will have their graphical/visual representation in the web page in the form of arrows, so that hey can implicitly induce the team leader to make adequate decision.

In this way, decision making of software managers and his fellow team leaders that are managing other developers will be significantly shortened and improved.

E. User Interface

Proposed intelligent system has the user interface in form of the web page so that team leaders can interact with it. User interface of the prototype intelligent system is shown n Figure 2.

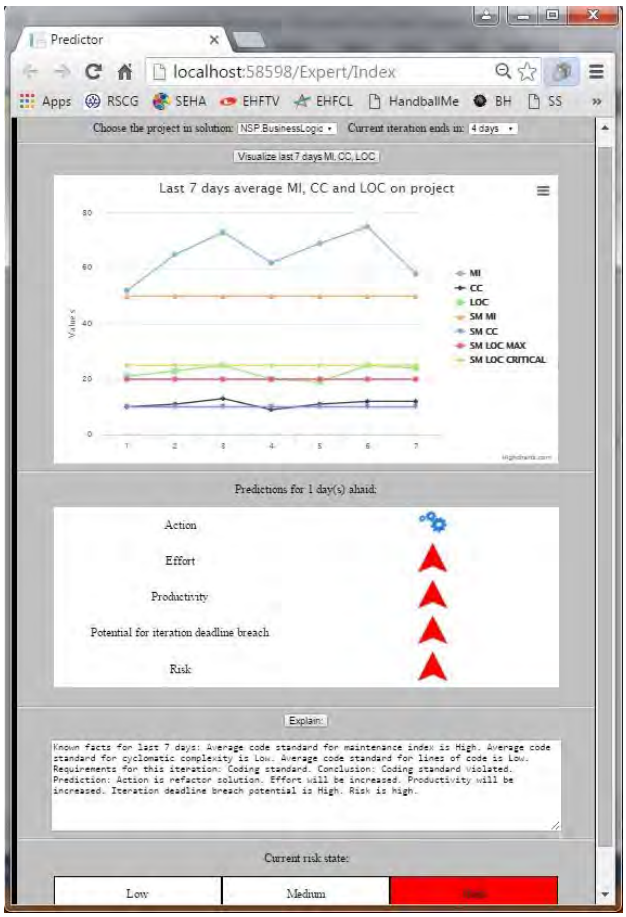


Figure 2. Interface for prototype rule-based expert system

User interface is fairly simple. It is a web page that consists of the following elements: dropdown box for choosing the project in developed solution and dropdown box for iteration end; button bellow it to trigger the visual representation of the current state of the project in terms of average MI, CC and LOC using Highcharts library; bellow the this graph there is visual representation of predictions made by inference engine, and a button for explaining how was the decision reached; lowest element of the page has the visual representation of the current risk on the examined project. All elements were written in HTML code.

V. Discussion Of The Solution And Results

The proposed intelligent system is in prototype phase when speaking in terms of expert system building methodology. Therefore, it is to young to give the precise picture of its ability to replace in full the human expert, our leading software manager. The thorough and exhaustive testing and validation of the expert system is yet to come in future months. Current testing of our prototype by human expert also shows that existing rules will have to be modified using fuzzy logic in the future, so that it can overcome their current strict boundaries. The limitation factor is also time and availability of the human expert and the developers team leaders who have lot of work to cover in their everyday work. In order to move on the development and testing of the prototype, it was given to 7 team leaders to try to work with it in their limited free time on local machines. After spending some time with the prototype, team leaders were given the questionnaire to answer questions regrading usefulness of the proposed rule-based expert system. The results are given in Figure 3.

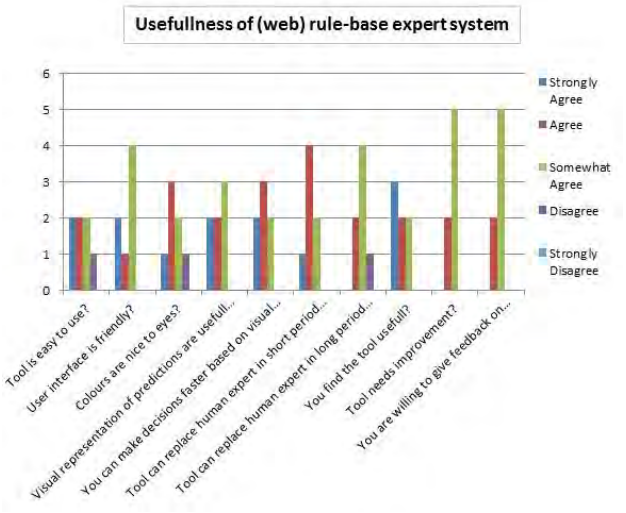


Figure 3. Questionnaire results from team leaders

The questionnaire included 10 questions with five level scale: strongly agree, agree, somewhat agree, disagree and strongly disagree. Questions that were given to team leaders are following: 1) Tool is easy to use?, 2) User interface is friendly?, 3) Colors are nice to eyes?, 4) Visual representation of predictions are useful to you?, 5) You can make decisions faster based on visual output?, 6) Tool can replace human expert in short period of time?, 7) Tool can replace human expert in long period of time?, 8)You find the tool useful?, 9) Tool needs improvement?, 10) You are willing to give feedback on improvement?

Results from analysis of questionnaire show following facts. On the first question, 57.1% of the team leaders would find the tool easy to use. Same percent of people find the colors used in the web interface appealing. User interface is friendly to 42.9% of examinees. Visual representation of predictions are useful to 57.1% percent of examinees. Five team leaders or 71.4% find that hey can make their decisions faster and that the tool could replace the human expert in short period of time. In this moment, tool is still not ready for long term replacement of the human expert, which can be seen from the fact that only 2 team leaders share the positive opinion about that question. Five team leaders or 71.4% of the total number of the examinees, would find the tool useful in their everyday work. The opinion that tool needs improvement was found with 2 team leaders or 28.6%, and 5 team leaders or 71.4% would somewhat agree with that fact. On the other hand, the 14.3% of the examinees did not find the tool easy to use, in the same percent did not like the colors used in web interface, and the same percent of the team leaders does not think that a tool could make replacement of the human expert in the long period of time. Lastly, all examinees showed similar interest and readiness to give feedback for expert system improvement, that was asked in last question of the questionnaire. Limitation of our proposed prototype of rule-based expert system is the fact that it is produced only for small number of people. In other words, in this

moment, only a closed group of users can use it and test its usefulness for everyday work and operations.

VI. CONLUSION

Software metrics has an important role in development of large software solutions. Its visualization and visualization of predictions of software development can play vital part of todays software manager tasks. In this paper we proposed the prototype of an rule-based expert system with capabilities of visual representation of predictions in future software development, based on software metrics history data. The results showed that prototype of the rule-based system would bring benefits for team leaders working on the software project. Benefits include visual representation of state of the developed project, visual representation of predictions for future software development, faster decision making of software managers, reduction of the risk in overall project based on visual representation of software metrics history data, which makes decision making more reliable, then analyzing numerical representation of the same. Our future work will be based on continuation toward ending the proposed expert system, and its finalization in terms of production implementation.

REFERENCES

[1] D.Galin, Software Quality Assurance: From theory to implementation, Pearson Adisson Wesley, ISBN: 0-201-70945-7, 2004.

[2] I. Sommerville, Software Engineering, 9th edition, Addison-Wesley, ISBN-13 978-0-13-703515-1, 2011.

[3] D. Longstreet, Function Point Analysis Training Course, [online], available at: http://www.softwaremetrics.com/Function%20Point%20Training%20Booklet%20New.pdf, [accessed: 21.10.2015. at 23:55].

[4] N. E. Fenton, S. L. Pfleeger, Software Metrics: A Rigorous & Practical Approach, second edition, PWS Publishing Company, ISBN: 0-534-95425-1.

[5] Microsoft MSDN – Code Metrics Values, [online], available at: https://msdn.microsoft.com/en-us/library/bb385914.aspx, [accessed: 25.12.2015. at 13:35h].

[6] Code Metrics Viewer Values and Meaning, [online], available at: https://codemetricsviewer.wordpress.com/2011/06/26/how-to-interpret-received-results/, [accessed: 23.11.2015. at 15:28].

[7] Dz. Strujic, A. Balota, B. Popovic, Software metrics as the backbone of reducing risk in development of Enterprise software solutions, XXI science and business conference YU INFO, Kopaonik, Serbia, Proceedings, pp.151-156, ISBN: 978-86-85525-15-5, 2015.

[8] A. Marcus, L. Feng, J. I. Maletic, 3D Representations for Software Visualization, In: Proceedings of the 2003 ACM symposium on software visualization (SoftVis'03). ACM Press, New York, pp 27-36.

[9] C. Knight, M. Munro, Comprehension with[in] Virtual Environment Visualizations, In: Proceedings of the Seventh IEEE International Workshop on Program Comprehension, pp 4-11, 1999.

[10] J. Maletic, A. Marcus, M. L. Collard, A Task Oriented View of Software Visualization, Proceedings of the First International Workshop on Visualizing Software for Understanding and Analysis (VISSOFT '02), 0-7695-1662-9/02, IEEE, 2002.

[11] T. A. Ball, S. G. Eick, Software Visualization in the Large, IEEE Computer, Vol. 29, No. 4, April 1996, pp. 33-43, 1996.

[12] M. Lanza, S. Ducasse, Understanding software evolution using a combination of a software visualization and software metrics, L'objet-8/2002, LMO'2002, pp. 135-149, 2002.

[13] D. Holten, R. Vliegen, J. J. van Vijk, Visualization of Software Metrics using Computer Graphic Techniques, s.l., s.n.

[14] M. Compton, Visualization of Software Metrics, [online], available at: http://marlenacompton.com/wp-content/uploads/2009/05/visualization_of_software_metrics.pdf, [accessed: 04.10.2015. at 11:25h], 2009.

[15] G. Langelier, H. Sahraoui, P. Poulin, Visualization-based Analysis of Quality for Large-scale Software Systems, ASE'05, November 7-11, 2005, Long Beach, California, USA, pp. 214-223, ACM 1-58113-858-X/04/06, 2005.

[16] M. Balzer, O. Deussen, C. Lewerentz, Voronoi Treemaps for the Visualization of Software Metrics, Association for Computing Machinery, pp. 165-172, ACM 1-59593-073-6/05/0005, 2005.

[17] Atomiq, [online], available at: http://www.getatomiq.com/features, [accessed: 03.09.2015. at 14:20h].

[18] Nitriq Code Analysis, [online], available at: http://www.nitriq.com/features, [accessed: 03.09.2015. at 15:30h].

[19] Submain CodeIt.Right, [online], available at: http://submain.com/codeit.right/features.aspx#1, [accessed: 03.09.2015. at 16:15h].

[20] NDepend, [online], available at: http://www.ndepend.com/, [accessed: 03.09.2015. at 17:20h].

[21] D3 Data Driven Documents, [online], available at: http://d3js.org/, [accessed: 07.10.2015 at 01:23h].

[22] Javascript charts for web and mobile, [online], available at: http://www.fusioncharts.com/, [accessed: 08.10.2015. at 00:34h]

[23] jQuery plotting library, [online], available at: http://www.jqplot.com/, [accessed: 09.10.2015. at 23:45h].

[24] Google Charts library, [online], available at: https://developers.google.com/chart/interactive/docs/, [accessed: 12.11.2015. at 23:42h].

[25] Highcharts library, [online], available at: http://www.highcharts.com/, [accessed: 11.11.2015. at 20:15h].

[26] M. Negnevitsky, Artificial Inteligence: A Guide to Inteligent Systems, second edition, Addison-Wesley, ISBN: 0-321-20466-2, 2005.

[27] S. Russel, P. Norvig, Artificial Intelligence: A Modern Aproach, third edition, Prentice Hall, ISBN-13: 978-0-13-604259-4, 2010.

[28] P. Sosnin, Question-Answer Programming and Modeling in Expert Systems, Proceedings of the IASTED International Conference Artificial Inteligence and Applications (AIA 2009), pp. 208-123, Insbruck, Austria, 2009.

[29] M. S. P. Babu, N. V. R. Murty, S. V. N. L. Narayana, A Web Based Tomato Crop System Based on Artificial Intelligence and Machine Learning Algorithms, International Journal of Computer Science and Information Technologies, Vol. 1 (1), 2010, pp. 6-15, ISSN: 0975-9646.

[30] I. M. Dokas, Developing Web Sites For Web Based Expert Systems: A Web Engineering Approach, In: Proceedings of the Information Technologies in Environmental Engineering (ITEE'2005), September 25-27 2005, Otto-von-Guericke-Universitat Magdeburg, pp. 202-217.

[31] N. Askarov, K. Kozhakhmet, L. Atymtayeva, Analysis Of Web-Based Applications For Expert Systems, Computer Modeling and New Technologies, 2011, Vol. 15, No. 4, pp. 41-45.

[32] A. Abraham, Rule Based Expert Systems, Handbook of Measuring System Design, edited by Peter H. Sydenham and Richard Thorn, John Wiley & Sons, Ltd. ISBN: 0-470-02143-8, 2005.

[33] V. Rafe, M. H. Goodarzi, A Novel Web-based Human Advisor Fuzzy Expert System, Journal of Applied Research and Technology, Vol. 11, No. 1, pp. 161-168, 2013.

[34] Y. Duan, J. S. Edwards, M. X. Xu, Web-based Expert Systems: Benefits and Challenges, Information and management, Vol. 42, No. 6, pp. 799-811, 2005.

Interaction with virtual objects in a natural way

I. Prazina, K. Balic, K. Prses, S. Rizvic and V. Okanovic*
*Faculty of Electrical Engineering Sarajevo, Sarajevo, Bosnia and Herzegovina
iprazina1@etf.unsa.ba, kbalic1@etf.unsa.ba, kprses1@etf.unsa.ba, srizvic@etf.unsa.ba, vokanovic@etf.unsa.ba

Abstract - Digital technologies are an efficient tool for visualization and presentation of the virtual objects. Interaction with the virtual objects for their presentation can be very important, especially in a domain of the presentation of cultural heritage. Implementation of natural interactive online 3D visualization is difficult process. This paper describes the interaction techniques for manipulating 3D objects. The paper gives one solution for natural interaction with virtual objects using Leap Motion as a sensor and WebGL for presenting virtual hand and 3D objects.

I. Introduction

Implementing 3D visualization is difficult process. One of the main problems is making users engaged in a 3D scene. When we have geometric data of 3D objects and environments we must implement means of interaction which will be helpful for a user and give him/her best experience.

There are different ways of interaction that use new technologies but only those which are easy and fun to use, and those which users accept will remain.

In this work two ways of interaction, with a 3D object, using Leap Motion will be presented. One with visible and one with hidden virtual hand. Here will be presented the way of interaction with virtual scenes where a user can chose in a HTML page which scene he wants to see.

These interactions are implemented using Three.js for online 3D presentation and Leap Motion as a sensor.

II. Related work

In [1] we described one interaction using Leap Motion where the virtual hand has dynamics of the real hand. This type of interaction had problems where users attention was mostly on the virtual hand not on the 3D object.

Use of Leap Motion in interaction with 3D objects is researched in following papers.

Gesture recognition with Leap Motion is described in papers [2], [3] and [4].

In [2] Kinect's and Leap Motion's pros and cons are given. The authors state that Leap Motion gives data as a set of relevant hand points and pose features that is easier to use, while Kinect data is more general and less accurate.

In [3] use of Leap Motion in sign language gestures recognition is presented. Authors find that Leap Motion can be used in sign language recognition with some problems when some of the fingers is not visible to the sensor.

In [4] the authors use the Leap Motion API to detect key press and to make keyboard like musical instrument.

The before mentioned works find Leap Motion easier to use in comparison with similar devices because of its well documented API and high level data. One of the problems Leap Motion has is small interaction area and problems with hidden fingers' position estimation.

In [5] authors give detailed analysis of the Leap Motion precision and reliability. One of the parameters that affects Leap Motion's performance is hand-sensor distance. Authors find that Leap Motion is less accurate if the hand is more than 250mm above the controller.

In [6] authors propose use of Kinect in human gestures recognition in interaction with 3D virtual environments.

III. Two ways of interaction

The interaction described in [1] distracted users from a 3D object. Users where more focused on the virtual hand and its dynamic. The virtual hand was simulated using JavaScript physics engine. Collision properties where added to the virtual hand and to the 3D object. The virtual hand's position and rotation is calculated using physics engine and Leap Motion's sensor (Figure 1).

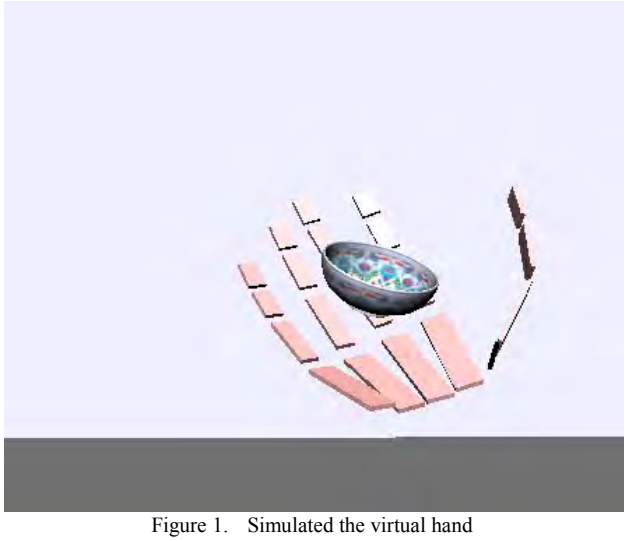


Figure 1. Simulated the virtual hand

To improve user's experience but keep fun of use in the new interaction hand dynamics is not simulated and

the hand is not displayed in the 3D object's scene. The virtual hand is now displayed in the main scene.

Main scene contains two sub-scenes in which a 3D object is displayed. In first interaction, in the main scene a user can see his/her hand moving on top of the HTML page. The virtual hand in this interaction is used like mouse cursor. User can now chose a scene by tapping on it. When tapping gesture is detected, and a sub-scene is picked the focus of interaction is moved to the sub-scene.

In the sub-scene a 3D object is displayed. The object can be rotated using Leap Motion's data.

Special logic for cases when hand is lost accidentally, and when hand intentionally leaves sensor space is implemented. In the implementation focus of interaction stays at active scene if hand is lost accidentally, and in case when hand is intentionally moved from sensor area main scene is selected as active.

IV. CASE STUDY

A. *The White Bastion project*

The fortification known as ”White Bastion” is one of the most impressive and important historical sites in Sarajevo. In this site a significant number of artifacts was found digitized. 4D Virtual presentation of White Bastion aims to present the historical development of this cultural heritage object through digital stories combined with interactive 3D models of the Bastion in various time periods.

Digitized findings and their reconstruction are stored as two 3D objects. On the HTML page both scenes for digitized finding object and reconstruction object are displayed. Structure of the project is displayed in Figure 2.



Figure 2. Project structure

B. *Natural interaction implementation*

Implementation of the idea is made using two libraries. Three.js is used as a library for 3D scenes and objects presentation in web browser. Leap Motion's API is used for data acquisition.

The initial interaction is when a user is picking which scene he/she wants to see. In the interaction hand is displayed as a set of cube meshes which represent finger's bones. Similar idea of virtual hand presentation is used as in [1]. Each bone in hand is mapped as one mesh, and for each mesh position and rotation is acquired by the sensor.

Position of a finger's meshes are updated in Leap Motion's loop. In the first loop iteration initial mesh position and rotation is set using data from Leap Motion. In every next iteration a mesh position is set using a data from new frame. In each iteration Leap Motion gives new frame with current positions and rotations. Position and rotation update is done for each bone's mesh in each finger. In this way virtual hand’s movement is solved.

There are two events that are critical for the interaction: when hand is placed above the sensor and when the hand leaves sensor’s space. On first event meshes are created/recreated. On second event all hand’s meshes are destroyed.

When device has registered the hand in the interaction box, the corresponding function is called. This function creates meshes of the hand in the main scene that represents whole screen. Web page gives options to interact with one of two objects:

(1) Reconstructed 3D model,

(2) 3D model of the excavated object.

By moving our hand and placing it above the object with whom we want to interact on double tap gesture (as we tapping the imaginary screen) with hand, we recognize gesture and detect which scene is chosen using hand position. After that, we remove a virtual hand from the main scene and place it in a chosen scene so we can interact with the desired object (Figure 3).

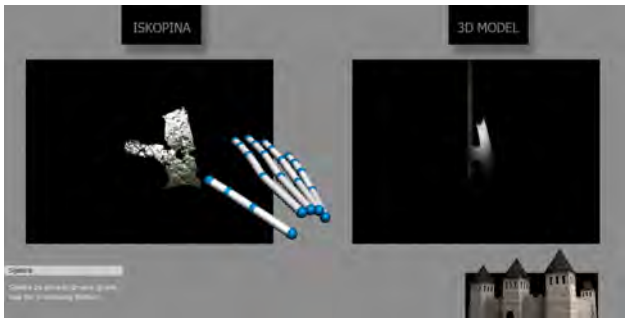


Figure 3. Main scene interaction

After scene is picked the new type of interaction applies. Virtual hand is not displayed in the scene because it would obstruct the view of the 3D object. Rotation of the hand is used to rotate 3D object. This way the interaction is easier because we don’t need to “grab” the objects. Only constraint is that our hand has to be within interaction box (area) of sensor. This gives us flexibility to see every side and every edge of the object.

Logic in cases of hand lost is implemented in following way; if sensor accidentally loses our hand, nothing happens until it is recognized that user wants to leave the scene. We predict if user moved hand out of interaction area accidentally, he/she will return his/her hand in 1.5 seconds. In that case, we do nothing. The 3D object's scene is still active. If wait-time is longer then 1.5

seconds, we conclude that the user's intention is to leave the scene. Then the virtual hand is removed from the object's scene. When a hand is placed on top of the sensor again, virtual hand will be placed in the main scene and the active scene is main scene. Described steps are presented in diagram (Figure 4).

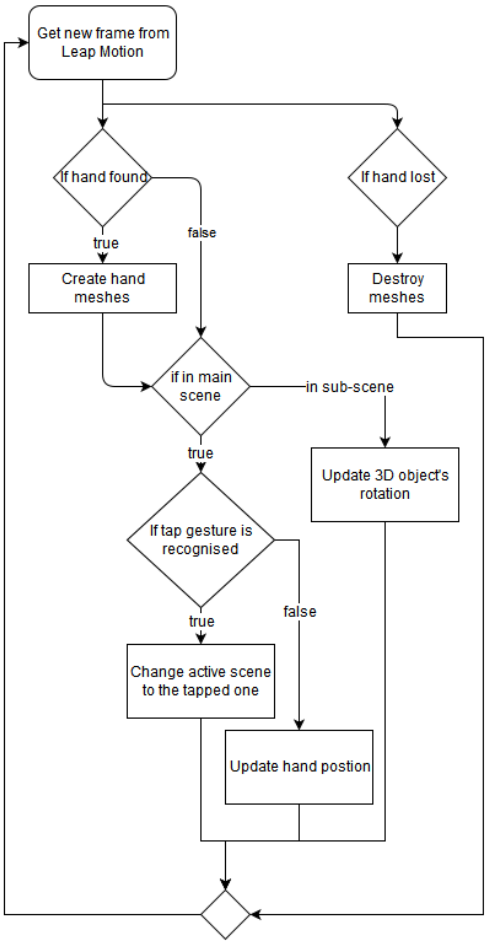


Figure 4. A 3D object web page from White Bastion project

The described interactions are added to the White Bastion project. In the White Bastion project there is a web page for each of the digitized archaeological findings. The web page's layout is made of two three.js scenes; in the first scene 3D model of the excavated object is presented and in the second scene the reconstruction of the object is displayed. In the initial version of the project objects can be viewed from all sides using mouse for movement and rotation. In the new version the Leap Motion interaction is added. An example web page is presented in Figure 5.

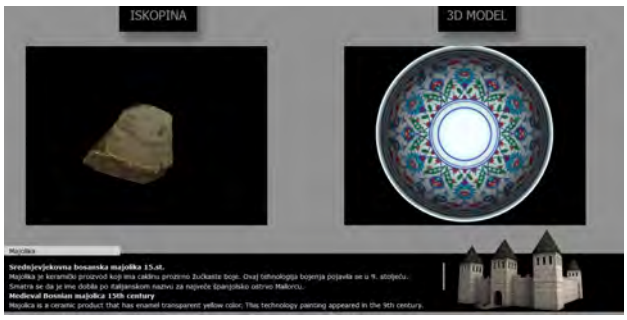


Figure 5. A 3D object web page from White Bastion project

V. EVALUATION

Using qualitative user evaluation, interactions are evaluated. In the evaluation the interaction is compared with initial mouse interaction and the simulated hand interaction. Participants in the analysis are selected so they have different educational backgrounds, age and gender.

Twelve participants have been interviewed. Interview was prepared as set of open questions. Participants had freedom to express their impressions and questions are used as a guidelines. Hypotheses that are being checked in the interview are presented in Table 1.

TABLE I. LIST OF HYPOTHESES

Mark	Hypothesis
H1	Leap Motion interaction is easier to use than mouse
H2	The virtual hand draws attention from 3D object
H3	Leap Motion interaction is more engaging

List of asked questions was enough for getting participants' opinions, impressions and to test our hypotheses. List of questions is given in Tables 2.

TABLE II. QUESTIONS FOR HYPOTHESES H1, H2 AND H3

Mark	Question
Q1	What is your opinion on interaction with a virtual object using mouse? (What are advantages and disadvantages?)
Q2	What is your opinion on interaction with a virtual object using Leap Motion?(What are advantages and disadvantages?)
Q3	What is easier to use and why?
Q4	How displayed virtual hand affects your experience?

In the interview most of the time participants spent in interactions with Leap Motion. Participants were unfamiliar with Leap Motion and at first they found it hard to use. Once they get familiar with it, they used it with ease. In the case of the simulated virtual hand participants had fun using it but also they had problem to see every side of the object, because virtual hand was obstructing the view.

Participants found mouse interaction familiar, easy to use but less exciting. This disproved our hypothesis H1. Even though participants found mouse easier to use at a same time they found the Leap Motion interaction more interesting and immersive, which proves our H3. Some of them made statements like: “Leap Motion is not precise as mouse but it is interesting, fun and has potential”, “When using the Leap Motion with hand displayed I focus only on the hand”, “Leap Motion can make people interested in viewing all these objects”, “In the case when the hand is not visible I can focus on 3D object even more than when I use mouse” etc. (H2 and H3 proven).

All participants claimed that mouse interaction is more precise and reliable but dull and less immersive. They had feeling of immersion when using Leap Motion, and that experience made them interested in the real object and historical site where they have been found.

VI. CONCLUSION

In this paper improved interaction with small 3D object scenes was presented. User evaluation's participants have accepted and used the interaction with ease. As in the previous works users had some problems with the sensor's accuracy and performance. Even with that problem in mind participants were more engaged and motivated to know more about 3D objects. Participants stated that they had fun using the application and that it made them interested in the history. This interaction shown potential as it was accepted by people with different age and education.

REFERENCES

[1] I. Prazina, S. Rizvić, Natural interaction with small 3D objects in virtual environments, Submitted to 20th Central European Seminar on Computer Graphics 2016.

[2] P. Zanuttigh, G. Marin, F. Dominio, Hand gesture recognition with leap motion and kinect devices, ICIP, Paris, France, 2014.

[3] L. E. Potter, J. Araullo, L. Carter, The leap motion controller: A view on sign language, 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration, Adelaide, Australia, 2013.

[4] J. Han, N. Gold, Lessons learned in exploring the leap motion sensor for gesture-based instrument design. Internation Conference on New Interfaces for Musical Expression, UK, 2014.

[5] J. Guna, G. Jakus, M. Pogacnik, S. Tomazic, J. Sodnik, An analysis of the precision and realiability of the leap motion sensor and its suitability for static and dynamic tracking.,Sensors, 2014.

[6] J.O. Kim, M. Kim, K. H. Yoo, Real-time hand gesture-based interaction with objects in 3d virtual environments. International Journal of Multimedia and Ubiquitous Engineering, 2013.

[7] A. Ranfelt, Build a simple 2d physics engine for javascript games, IBM developerWorks, 2012.

[8] Three.js. http://threejs.org. Accessed: 2016-02-14.

[9] Cannon.js. http://www.cannonjs.org. Accessed: 2016-02-14.

[10] Leap motion. http://developer.leapmotion.org. Accessed: 2016-02-14.

Bone Shape Characterization Using the Fourier Transform and Edge Detection in Digital X-Ray Images

Diego Sušanj[‡], Gordan Gulan[#], Ivica Kožar[†] and Željko Jeričević[‡]
[‡]Department of Computer Engineering/Engineering Faculty, Rijeka, Croatia
[#]Medical School Rijeka, Croatia
[†] Department of Computer Modeling/Civil Engineering Faculty, Rijeka, Croatia
zeljko.jericevic@riteh.hr

Abstract - From the series of digital X-ray images we extracted bone edges using information entropy based edge detection algorithms. The extracted edges are series of two dimensional coordinates from image space. The space series were tested for shapes of medical interest inside the knee joint. In particular, the existence of straight edges inside the knee joint was tested for by using digital filtering and analytic computation of first and second derivative in the Fourier domain. This kind of analysis was used because it provides rotational invariance for the sought bone shape, and allows the statistical comparison of shapes from different images. The real life examples are taken from medical practice using X-Ray imaging of series of knee joints in order to illustrate the analysis procedures and their medical relevance for real data.

I. INTRODUCTION

In our previous paper [1] we described use of Shannon's information theory [2] for the detection and analysis of edges in digital images. Our numerical experiments confirmed that the use of information entropy measures for edge detection yields a better continity of the edges then some classical operators like Sobel.

The main goal in our work presented here was to achieve continuity of the edge in low contrast and noisy X-ray images, which is a first necessary step for shape analysis of femur (thigh bone) in a knee joint. There are some valid medical reasons to expect that the portion of the bone in knee joint is straight to accommodate sliding motion during the joint operation.

II. THEORY

The X-ray images are coming in all sorts of different orientation and sizes and analysis which aims to become massive and automated must resolve those problems. Problem of different sizes can to a large extent be reduced by organizing data into age, gender and sex based groups. The problem of different orientation, on the other hand, has to be solved computationally. It is not reasonable to expect that every X-Ray image coming from medical practice will be oriented in the same way in two dimensional image space. If we want to meaningfully compare the images, it is necessary to develop analysis strategy which will be shift and rotationally invariant. Our approach is based on computing the second derivative of the bone edge. For a straight edge section, the first derivative will be constant, while second derivative will be zero regardless of orientation of the edge in image. The problem of increased high frequency context in the derivatives was solved by doing most of our numerical processing in the Fourier domain, using filtering and taking the derivative in the Fourier space [3].

III. METHOD

Taking derivative of experimental functions in the Fourier domain is the equivalent of taking the analytical derivative of a theoretical function. When the derivative is computed in the Fourier domain, all the frequencies contained in the sequence are used in computation [3]. From the Fourier transform $F(\omega)$ of function $f(t)$, the Fourier transform of derivative $H(\omega)$ is calculated by multiplications of frequencies with factor $-i\omega$ (1)

$$F(\omega) \underset{\div(-i\omega)}{\overset{\cdot(-i\omega)}{\rightleftarrows}} H(\omega) \qquad (1)$$

In contrast, the numerical derivative in space domain uses only neighboring points and is very sensitive to the noise in individual points [4]. The following series of digital images illustrate the steps in our processing sequence. The images are collected from medical practice in Orthopedic Clinic of Lovran, Croatia.

The Fig. 1 shows X-ray image with superimposed edge detected using information entropy [1]. Edge is continuous, but it is not smooth, and an attempt to compute the derivative numerically will result in noise enhancement and highly oscillating curve. In order to avoid this type of problem, our first step was to filter out higher frequencies from the edge data. The Fig. 2 shows the same image and same edge data after filtering in the Fourier domain.

The edge data are series of coordinates in two dimensional image space and are treated as an experimental function in parametric form with x and y functions defined using parameter s. The parameter s is a space index for sequence of points in the edge.

$$\frac{dy}{dx} = \frac{dy}{ds} / \frac{dx}{ds} \qquad (2)$$

According to the chain rule, the derivative of function in parametric form is calculated using the (2).

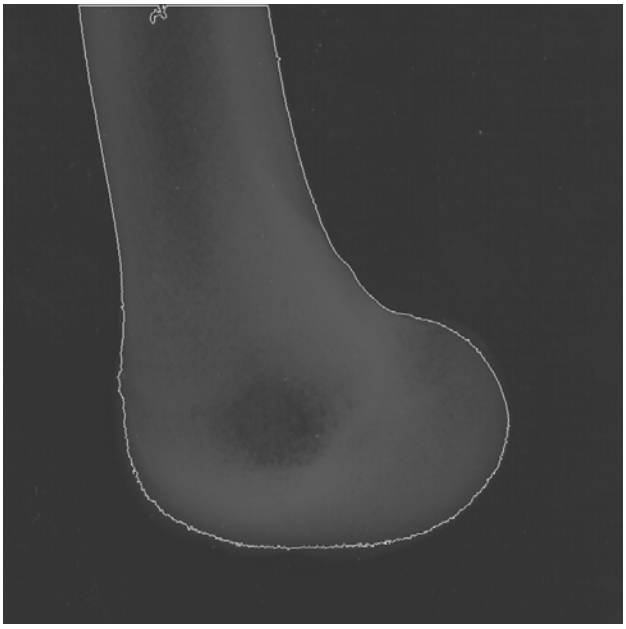


Figure 1. Image with superimposed edge detected by information entropy



Figure 2. Image with superimposed filtered edge detected by information entropy

Values for first derivative of function x and y are shown in Fig 3.and the second derivative for function x and y in Fig. 4. Both graphs are for the filtered edge shown in Fig. 2.

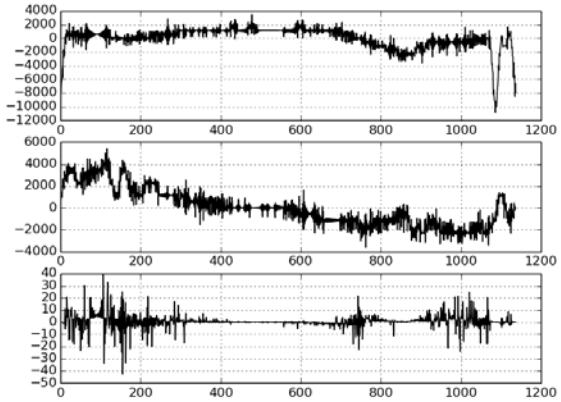


Figure 3. Top: First derivative of x. Middle: First derivative of y. Bottom: First derivative computed as in (2)

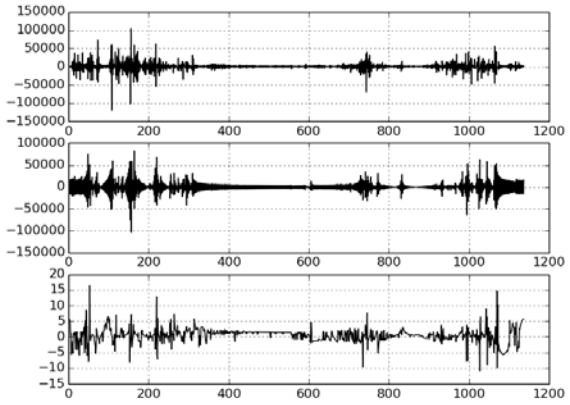


Figure 4. Top: Second derivative of x. Middle: Second derivative of y. Bottom: Second derivative computed as in (2)

Threshold of derivative (only values close to zero are shown) are superimposed on bone image, showing flat regions of the bone edge in the Fig. 5.

The numerical experiments were performed on a set of about 20 different images in order to asses the validity of our approach. With first and second derivatives already computed, we also tested for the curvature (κ) of edge function which is defined in (3)

$$\kappa = \frac{\dot{x}\ddot{y} - \dot{y}\ddot{x}}{\left(\dot{x}^2 + \dot{y}^2\right)^{3/2}} \qquad (3)$$

.

The thresholds of curvature (only values close to zero are shown) are superimposed on bone image, showing flat regions of the bone edge in the Fig. 6.

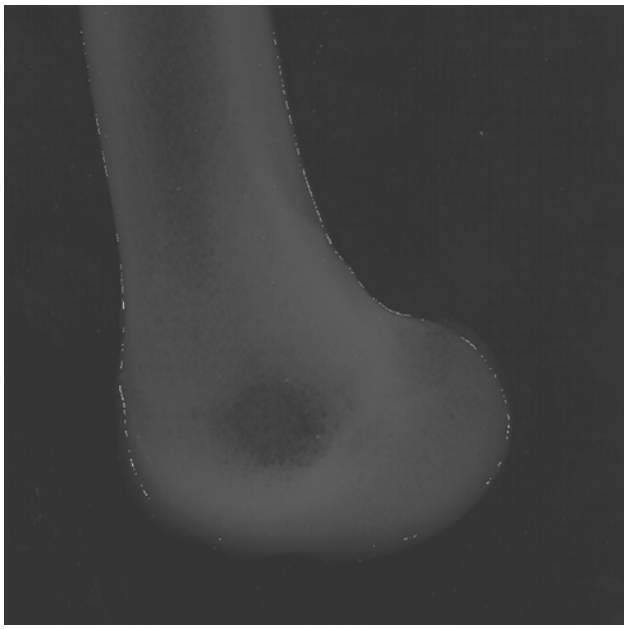


Figure 5. White values indicate second derivative close to zero.

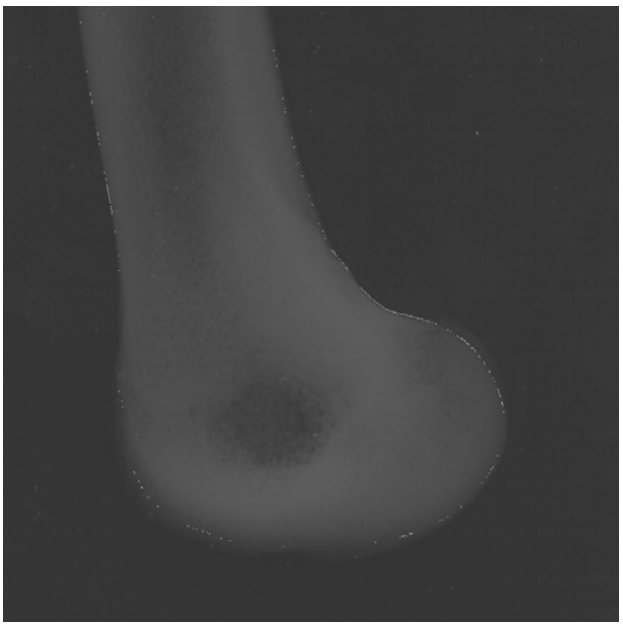


Figure 6. White values indicate curvaturee close to zero.

IV. CONLUSION

The information entropy based edge detection algorithm [1] has been combined with digital filtering, analytical computation of first and second derivative and curvature to find straight line sections of a knee joint. The combination of techniques used allows construction of shift and rotational invariant processing scheme and comparison of different data sets. In this way we are building the elements for an automated system for X-ray digital image analysis in order to conduct clinical studies with large number of patients. First testing was done on a set of 20 X-ray images of femur bones. Our results strongly indicate that on the basis of better edge continuity for information entropy operator it is possible to construct shift and rotational invariants and search for the straight sections of femur bone in knee joint. From the medical standpoint, the shape analysis of multiple knee joints will provide orthopedic surgeon with detailed statistics about variability in the shape of a knee joints and presents a valuable tool for clinical studies.

ACKNOWLEDGMENT

This research was supported by Croatian Science Foundation (HRZZ) Grant No. 3432

REFERENCES

[1] Diego Susanj, Vjeran Tuhtan, Luka Lenac, Gordan Gulan, Ivica Kozar, Zeljko Jericevic "Using Entropy Information Measures for Edge Detection in Digital Images", DC VIS MIPRO (2015) 368-371 paper #3016

[2] Shannon, C.E. "A Mathematical Theory of Communication". *Bell System Technical Journal*, **27** (1948): 379–423.

[3] Weaver, H.J. “Applications of Discrete and Continuous Fourier Analysis”, 1992, Krieger Pub Co, Malabar, FL, pp 375.

[4] Gonzalez, R.C, Woods, R.E., “Digital Image Processing”, 3rd Ed. 2008, Pearson Prentice Hall, Upper Saddle River, pp 497.

GIS in the e-government platform to enable state financial subsidies data transparency

Mirjana Kranjac*, Uroš Sikimić**, Ivana Simić***, Milan Paroški****, Srđan Tomić*****
* University of Novi Sad, Faculty of technical sciences/Department for transport, Novi Sad, Serbia
** Politechnico di Milan/Department for engineering, Milan, Italy
*** Provincial secretary for economy, employment and gender equality, Novi Sad, Serbia
**** University of Novi Sad, Faculty of technical sciences/Department for electronic, Novi Sad, Serbia
*****University Nikola Tesla, Faculty for engineering management, Belgrade, Serbia
mirjana.kranjac@ftn.uns.ac.rs, uros_sikimic@hotmail.com, ivana.simi@vojvodina.gov.rs, milan.paroski@ftn.uns.ac.rs, srdjan.tomic@fim.rs

Abstract - The authors of the paper present an expanded model of the e-government. The model includes a new function within e-government platform and this is visibility of the state aid activities. Visualization of financial subsidies given by the different state administration levels (national, provincial and local) is presented with the Geographic information systems – GIS tools. Such approach makes the activities of the public administration transparent to all deljenje stakeholders and contributes the e-democracy evolution. In the paper is presented pilot attempt of the Provincial government of Vojvodina to implement such an innovative tool into the communication with citizens. The scientific contribution of this paper is to present GIS tool which includes new spatial visibility into the transparency of state activities.

I. INTRODUCTION

Communication between citizens and government is a crucial element of the democracy in a state. Citizens require permanent exchange of information with its government and transparency of all significant data. They are extremely interested into different types of state aids. There is a big competition to get state subsidies from all stakeholders and they are strictly observing the whole competing process and they analyze in details the list of entities selected to be subsidized. For the government is very good to present the data about deliver of state aid in such a way which will make user satisfied with the impression that the financial support is delivered rightly and that the process and results are transparent to everybody. All activities which belong into presented process are important part of the e democracy. E democracy could be explained as democracy enabled by implementation of different ICT tools which makes it fast responsive, user friendly and available to all citizens. Every possibility to have interactive approach to these results is appreciated a lot from the users. E government is a new approach which is entering new possibilities into permanent dialog between state administration and its human surroundings enabled by wide range of ICT tools. It is enabled by the intensive and fast development of the ICT sector. One of its functions is assistance to the state aid delivered. The state is spending a big part of its budget collected from the taxes to the various state aids. Engagement of the state aid is a part of strategic development vision of states and this makes its delivering process as much important. It directly activates many social groups like: public and private companies, entrepreneurs, different associations, NGOs etc. For some of them state aid is the only possible resource of financing and it enables their activities and living of their members at one side and political influence on other. Some, like old handicrafts, cannot survive without financial assistance. They are still keeping old professions due to state subsidies. It promotes accountability as an essential factor within society and offers better policy decisions.

Many governments improved transparency of their work at local, regional and national levels. Last few years is this process visible in many countries. Democratic society supported by strong civil sector is doing monitoring of public resources spending and is a strict censor of all irregularities. Many social groups are interested in allocation of state resources. Here is a big responsibility of the governments to satisfy citizens who want information about their money spending [1]. A primary objective which will make possible the modernization of state aid is making granting given by state transparent. Data available to everyone make citizens feel that they are part of democratization process, that they influence state decisions and control state activities, specially those connected with financial and other supports [2]. Implementation of emerging technologies, focus is on ICT, can visible foster the transparency of state domain.

This paper presents new aspect of ICT tools utilization, not yet used in this way, to spatially present state data, what makes them user friendly to citizens.

The research object is e government in the domain of state subsidies.

The authors set a hypothesis that the spatial state aid data presentation enabled by GIS improves state aid transparency.

The task which should be researched is to involve new type of spatial presentation into e government platform which will foster visibility and transparency of the governmental activities.

Methodology used by authors to reach outcomes is: collecting data, connecting data with geo coordinates, analyzis of connecting data, geo referenced visualization of data.

Authors main goal in the paper is to present how is transparency of public data about the state aid possible and useable for the citizens. This will expand e government functionalities from the present functions which could be described as "the utilization of information technology (IT), information and communication technologies (ICT), and other web-based telecommunication technologies to improve and/or enhance on the efficiency and effectiveness of service delivery in the public sector" [3] to the new level which will enhance visibility of the governments financial assistance. !!!

The specific goal is to present a new model of using geographic information systems - GIS within one domain of the e government platform which will enable visible, spatial presentation of the state aid to the all stakeholders.

This paper describes new model which is involved in the official web site of Vojvodina region to present visualized data about financial support for employment of unemployed in the region of Vojvodina for the year 2014. This is a good example which proves expanding possibilities of e-government by using GIS applications.

II. LITERATURE OVERVIEW

There are many ways to present different data. Mostly used are excel and word tables and geographical presentations. In many cases data are linked to the geographic issues and their presentations without link to the specific georeferenced points, lines or polygons are not understandable to human beings. Man like to have clearly visualized data which make him better understand a problem, analyze them more successful, present solutions and follow the changes. Changes could appear as changes through time (visits of tourist through years) and changes through space (incomes of people in different areas). The geographic information systems are created to make spatial data more understandable by there presentation in geographic maps. They are software for visualization of spatial data [4]. Geographic data are special in term that a location is their significant attitude and this distinguishes them from other sorts of data. Because of that GIS brings new benefits which enables many additional analysis and conclusions, usually unavailable in other softwares which do not have geographic components and cant present data in geographic maps [5, 6].

E-government as an important ICT branch gives enormous platform to introduce all benefits of using GIS. To understand its role, e-government service should be explained. E-government is a new concept with many explanations and discussions in papers, books. Some authors characterize Some definitions characterizes e-government as a group of new processes based on the use of information and communications technology, which should bring some advantages to the existing government actions. Its role is to enable more responsive and efficient government with transparent results of its activities

E-government concept covers few functions. One of them is to reform functioning of governmental institutions by enabling better common work among them and more efficient business between these institutions and the focus group, citizens. Its service to civil society should be better through ICT introduction into the government work.

Some of services which is new and very popular, well accepted by citizens is called: one-stop shop. It is available as part of the World Wide Web service. One-stop shop is service always requested from customers, namely it offers to them to finish the whole complicated administrative demands, by visiting only one place, one-stop shop on WWW. It is like visiting all offices only at one site, without running confused around and looking for the sites where different services could be finished. Users are doing all very fast and transparently [7].

Growing of e-governance is being fosterted by emerging implementation of IC technologies. ICT brought Internet as a new approach to many services. Internet platform added new services which enreaced . e-government processes. Involvement of GIS added a new dimension of visible reproduction of spatial information and easier interactions with the society. This made user friendly use of e-government. World Wide Web sites witch have incorporated GIS services have some new applications like:

- Government-to-business: It presents economic and land development, licensing.
- Government-to-citizen: It enables online service information.
- Government-to-government: It improves communication inside and among the government departments.

They enable citizens and governments to exploit resources wisely. They involve them into the planning process. They make them responsive effectively to the emerging situations and to prevent them etc. [8, 9].

Typical GIS architecture is presented by Kumar [10]. It is showing GIS architecture useful for the state implementation (Figure 1). Such an architecture presents how government, private sector and citizens could be linked to a GIS platform designed as support desk for the e government functioning. He stresses that automatization of the government service is of big significance and that core created with GIS framework makes it possible.

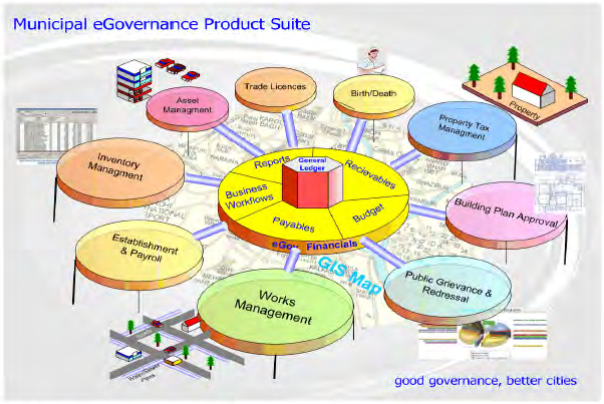


Figure 1. Municipal e-governance product suite. Source: Nadhamuni (2009).

Figure 1. GIS architecture for the E-government [11]

Local communities have essential role to generate numerous data inputs. They need a holistic output, full of information which is achievable by GIS. GIS expands e-government services with possibilities of making decisions, analyzing geographic data and clear presentation of it with many reports. GIS is solving problems which were obstacles to the work of administration like: problems of locations, urban and other planning, technical design etc. What is necessary is to have good knowledge experts and GIS software experts competent to work together and design algorithms which will reach the solutions of problems like: designing development plans, monitoring and reacting to changes and doing corrections [12].

Using of GIS within e-government system is explainted by some authors on the following manner:

- The ineraction between government and citizen is offen geographically based
- E-geography is within e-government concept [13].

Policy design and e-democracy are new sectors enabled by GIS. Some authors present how using GIS fosters citizens to be included into the urban planning and redevelopment of plans. Some case studies which they investigated proved that culture and knowledge potential are influencing of e-government platform. The gap between citizens and government could be reduced or eliminated when using GIS [14, 15].

Some authors [16] present framework of unified e-government management platform based on the Web-GIS It is easily implementable and a good technology for e-government information management. Their vision is economic development with Web based GIS. It should be site for competition of local communities. Private investments will be dependent on the displayed, open source spatial data presented in a e–government platform. Researching, analayzing, monitoring and evaluation will be done without being on the real place [17].

III MODELS OF GIS IMPLEMENTATION

Since year 2009 Secretariat for Economy, Employment and Gender Equality, on the basis of the decision of the Provincial Government of Vojvodina financially supports self-employment, employment, professional practice, and public works for the unemployed in Vojvodina region. These aides are defined in the Program for the active politic for the employment and have task to arise awareness about entrepreneurship, to incrise number of entrepreneurs and jobs within existing companies. Professional practices give opportunities to young unemployed professionals to prove them within companies and be employed after the practice is finished. Non profit organizations organize public works to employ unemployed to solve some social problems: gerontohousewifes, for different cleanings.. The target groups are two groups of unemployed people:

1. Potential entrepreneurs
2. Unemployed who have found companies ready to employ them.

This paper analysis susidies for employment in already successful companies which need more employees.

In this paper authors present how the ICT technics could be useful tool for the spatial data presentation, in this case presentation of data about geograhic distribution of the subsidies for employment. GIS software is included into the official site of the Provincial secretary for economy, employment and gender equality to make data spatial visible and to enable interactive use of them by all stakeholders, special those who participated in the competition to reach jobs.

Such an approach is:

1. Good way for the Secretary itself to follow timeline of results, to analyze effects, make follow up conclusions and correct its future actions.

2.It is one platform which researchers could use for different kind of investigations. Such activities will enable them to be included into state administration activities and suggest novelities, innovations and improvements. Such spatial mapping of the dispersion of new employed makes possible comparison through more years. It will give exact picture of possible social, economic, political influences when cross cutting different layers GIS makes very simply to do.

Such general, holistic insight into data is a base for different improvements.

The following part of the text presents different maps which are available to users due to their interactive requests. Maps contain spatial data and represent a small group of possibilities which are reachable by GIS use.

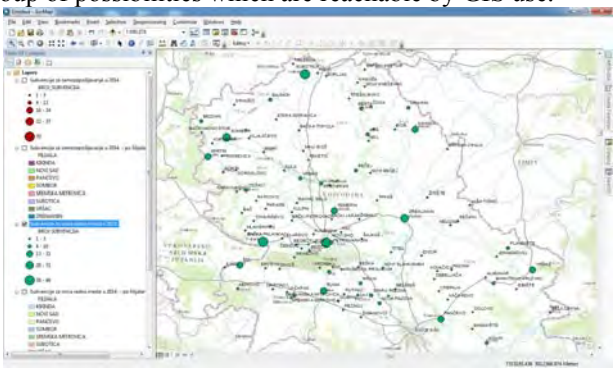


Figure 2. Subsidies for employment in 2014 (source: authors)

In Figure 2 is shown dispersion of financial subsidies for employment in 2014. At the left side of the Figure is table of contents where a layer which is marked is highlighted in blue color. The graduated presentation is chosen that circles are symbolizing number of giving due to the location of local community. All circles are green. The color green is selected for the data for new employed within existing companies. The circle`s size reflects the number of subsidies. The green color is used for the number of subsidies for new employment.

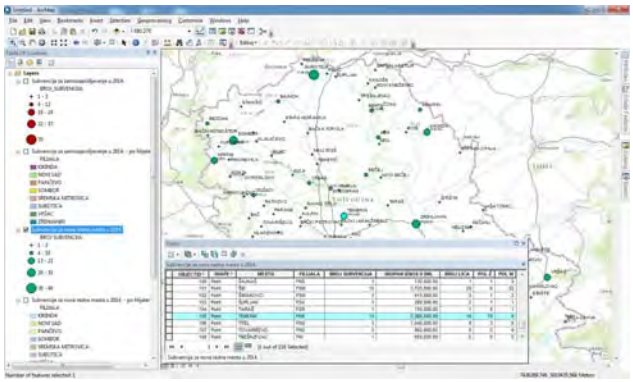


Figure 3. Subsidies for employment in 2014 with attribute table (source: authors)

Figure 3 has the attribute table which contains details connected to a selected location, here the location is: local community. Fifure 3 is showing the case when one local community is selected, in Figure 3, this is Temerin. Its circle changes color and becomes light green. This causes that GIS is opening the table with data. Data which are visible are:

- The name of the local office
- Number of subsidies
- Sum of each subsidy
- Total sum for this local office
- Number of persons who got it
- Number of females of the total giving
- Number of man of the total giving.

Object selected in the table is automatically visible on the map.

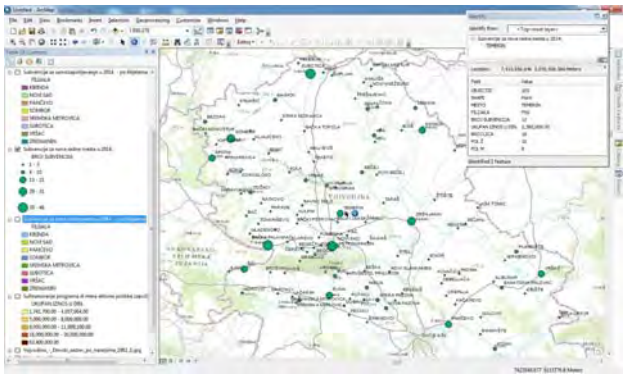


Figure 4. Subsidies for employment in 2014 with identification of one town (source: authors)

Figure 4 is new possibility of GIS to do identification. It is selected as Identify tool in GIS. When the certain local community is selected, user is receiving all information from the previous table, but only for the selected local community. He does not need to open whole attribute table to get them.

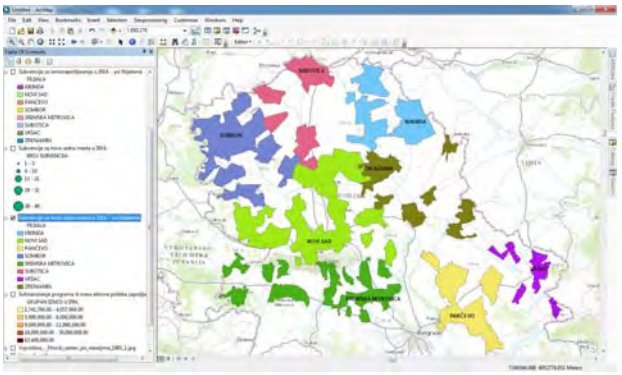


Figure 5. Subsidies for employment in 2014 within branches of National Employment Agency (source: authors)

In Figure 5 are shown data about subsidies for employment by branches of the National Employment Agency (NEA) in Vojvodina region. On the left side of the Figure appears the table of content with marked appropriate layer. Eight branches of the National Employment Agency (NEA) are colored with different colors to be recognizable due to colors.

All possibilities with presentation of data from attribute tables and data for each selected branch of the NEA can be used for the NEA branches, too.



Figure 6. Subsidies for employment in 2014 with overlapping of NEA branches with locations of villages (source: authors)

Overlapping of layers is a powerful tool of GIS. Figure 6 presents two layers simultaneously, containing different data of the same public call - competition. Overlapping of two layers enables superposition of data. In this case are shown local communities within each NEA which have companies which received state add for the new jobs. The size of the circle, this time colored into green color, corresponds to the total value of the grant delivered to this local community.

In the Figure 7 is displayed the zoomed NEA branch of the local community Pančevo. Settlements which belong to the territory of Pančevo (presented with green curcles) could be seen. The tool which is used to present them is called: Identify. The same procedure can be used for each local community: town or settlement. The table in the right corner contains all details about subsidies for Kovin. Kovin belongs to the NEA branch of Pančevo. This is possible for each settlement.

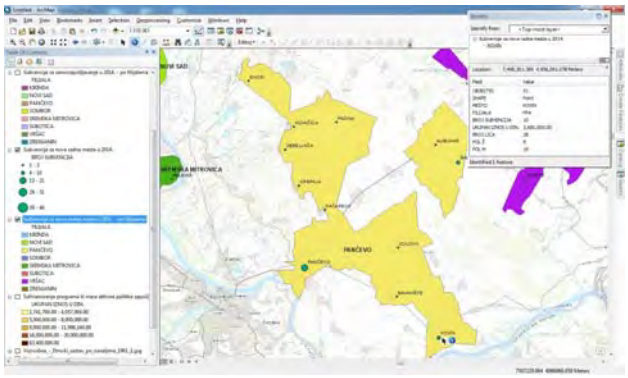


Figure 7. Subsidies for employment in 2014 within zoomed branch of local community Pančevo with settlements (source: authors)

IV. Conclusion

Scenario of the GIS use which is described in the paper is added value to the development of e-democracy. E-government could be significant tool to foster e-democracy. In the era of globalization which rises from the ICT tools contribution, GIS with its capability to visualize data is an additional advantage added to all what e-government offers and makes the set of its services even wider. Such approach which enables citizens' complete insight into the state decisions which must be independent and made according to previously set up criteria should be vision of the e-government development. The visualization in figures 2 to 7 prove the hypothesis that GIS tool implemented within e-government platform contributes to the transparency of the state aid data.The increase of number of visitors of the sight, which was 85% in the first month of its placing within government site, proves the same.

Further research and expansion of the concept to wider interactive, user friendly concept is possible. It could enable fulfilment of special users requests, customized to their needs and cause new actions of the state administration. Some of possibilities how GIS could be used in different sectors like: telecom market [18] research, postal service planning [19] and transport routing [20] and others are described in few papers. Added value of this paper is development of e-democracy with GIS.

References

[1] „State aid transparency for taxpayers", EC Competition policy brief, http://ec.europa.eu/competition/publications/cpb/2014/004_en.pdf

[2] Nicolaides, P, "Transparency is also needed at the European Commission level", *State aid hub EU*, 2014. http://stateaidhub.eu/blogs/stateaiduncovered/post/22

[3] Hai, J, "*Fundamental of development administration*", Selangor: Scholar press. ISBN 978-967-5-04508-0, 2008, pp. 125-130

[4] Goodchild, M, "Twenty years of progress: GIS science in 2010". *Journal of Spatial Information Science*, 2010, pp.12-13.

[5] Heywood, I., Cornelius, S., Carver, S, „*An Introduction to Geographical Information Systems",* (3rd ed.). Essex, England: Prentice Hall, 2006, pp. 230-232.

[6] Perry, M., Hakimpour, F., Sheth, A, „Analyzing theme, space and time: An ontology-based approach", Proceeding: *ACM International symposium on Geographic information systems* (PDF), 2006, pp. 147–154.

[7] Gordon , T, "E-government- introduction", *ERCIM News* No.48, Fraunhofer, FOKUS, 2012, pp. 211-213.

[8] http://www.esri.in/industries/government/egovernance

[9] Nikhil, K, "Citizen centric e-government", *State consultation workshops in partnership with Nasscom,* India, 2012.

[10] Kumar, A, "Participatory GIS for e-governance and local planning: A case study in different environments on its applicability", *Centre for Study of Regional Development*, New Delhi, India, 2013.

[11] Nadhamuni, S. "GIS architecture for the E-government", *Centre for Study of Regional Development*, New Delhi, India, 2009.

[12] Kranjac, M., Sikimić, U., Simić, I., Tomić, S. "*E-government based on GIS platform for the support of state aid transparency*", ICIST 2016, Kopaonik, 2016.

[13] Barr, R, "*E-government and GIS"*, The British Experience

Manchester Regional Research Laboratory, School of Geography, University of Manchester, UK, 2013, pp.15-20.

[14] Moody, R, „Assessing the role of GIS in E-government: A tale of E-participation in two cities", https://www.scss.tcd.ie/disciplines/information_systems/egpa/docs/2007/Moody%2007.pdf , 2005, (accessed 12.10.2015)

[15] Fu, P., Sun, J, "*Web GIS: Principles and applications*", ISBN:158948245X 9781589482456, Esri Press, New York, 2010, pp. 25-28.

[16] Lu, X, "A GIS-Based integrated platform for E-government application, Information Science and Engineering (ICISE), *1st International Conference on* GIS, IEEE, ISBN 978-1-4244-4909-5, 2009, pp. 200-218.

[17] Reddick, Ch, "*Handbook of research on strategies for local e-government adaptation and implementation: Comaparative studies"*, University of Texas, Information science reference, USA, 2009, pp. 632-635.

[18] Kranjac, M., Sikimić, U., Dupljanin, Dj., Tomić, S, "Use of Geographic information systems in analysis of telecommunication market", *Proceeding ICIST 2015*, Kopaonik, 2015, pp. 303-308.

[19] Kranjac, M., Sikimić, Vujaković, M, "Model for calculation of post service efficiency by using GIS", *Proceeding Mipro 2015*, Opatija, ISBN 978-953-233-083-0, pp. 353-359.

[20] Kranjac, M., Sikimić, U., Vučinić, M., Trubint, N, „The use of GIS in the model of implementation of the light train transport in Novi Sad", *Conference Towards a human city,* 2015, Novi Sad, FTN, ISBN 978-86-7892-541-2, 2015, pp. 124-130.

Evaluation of Caching Techniques for Video on Demand in Named Data Networks

Kire Jakimoski, Sime Arsenovski, Lidija Gorachinova, Slavcho Chungurski, Oliver Iliev, Leonid Djinevski and Emilija Kamcheva
FON University/Faculty of Informatics, Skopje, Republic of Macedonia
kire.jakimoski@fon.edu.mk

Abstract - Although the main driver for ICN (Information-Centric Networking) has been the rise in Internet video traffic, interplay between ICN and video is not much researched in the available literature. The type of ICN caching strategies that works well on video workloads as well as how ICN helps in improving the video-centric QoE (Quality-of-Experience) needs more thorough research from the academic community. In this work we research three types of ICN caching strategies or LRU, FIFO and LFU caching strategies and. Detailed evaluation of these three schemes is done in terms of the key performance metrics and important conclusions are obtained. Performance metrics that are researched in the ICN literature include cache hit ratio, throughput and server load. We expansively evaluate in this work a combination of 3 content replacement schemes on 4 different cache sizes of 1GB, 10GB, 100GB, and 1TB. In fact, the focus of our work is to evaluate the above mentioned schemes on video workloads in order to improve the user experience and network performances (lower congestion and lower server load).

I. INTRODUCTION

Information-centric networking concepts are supported in a lot of future internet architectures [1-3]. One of the main reasons for this behavior is that the Internet traffic consumption is mostly content driven. This means that users mainly worried about the type of content they want, and less about the location of the content that they request. As a result of this, several aspects of the networks nowadays could be simplified. One of the aspects is that complex optimizations that are required by CDNs (Content Delivery Networks) could be eliminated. Another aspect is to provide intrinsic trust in content in order to minimize trust in sources. The network could be also enabled to scale to greater volumes. This inspiration for ICN goes together with, and is guided in considerable part, by the remarkable increase in Internet video traffic volumes in the last several years. An example of ICN caching architectures is presented in Fig. 1.

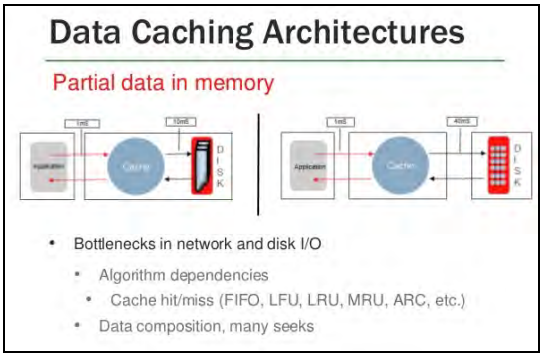


Figure 1. ICN architecture

In-network caching is very important in ICN (Information-Centric Networking) architectures [4] and there are a lot of researches in this subject including two main areas, content placement and content replacement [5, 6]. Strategy of content replacement [7] defines the method of ejection of objects when the cache is full. Authors in [8, 9] are trying to optimize the performance metrics of the content replacement algorithms. In the results obtained for this work three types of strategies are included:

- Least recently used or LRU;
- First-in first-out or FIFO;
- Least frequently used or LFU.

The rest of the paper is organized as follows. Section 2 presents the dataset and topology that was used for getting the simulation results. Section 3 presents the simulation results and Section 4 concludes the paper.

II. DATASET AND TOPOLOGY DESCRIPTION

In this section we explain the dataset and topology that was used for evaluating the replacement algorithms and to analyzing the obtained results. Datasets, as well as router-level network topology and the statistics of video views were gathered from the PPTV platform [10]. Three types of nodes are used in the obtained simulations, servers or content origins, consumers or content requesters and routers.

In the evaluations that are done in this work servers that correspond to the group of the PPTV CDN (Content Delivery Network) servers are used. It must be noted here that P2P mode is not taken into consideration, but only the CDN mode. Actually, 221 servers in total are connected to the network from the same access points as in the real deployment of PPTV. It is assumed that each of these servers is maintaining all of the video contents. Consumers correspond to the PPTV clients that are observed in the dataset.

Each of the clients is connected with the access router that has IP with the longest prefix in common. There are

82,726 included routers and each of them is equipped with a cache of the same size. Routing tables of each router are been populated using the shortest paths as in other research works [11-14]. This means that each of the content requests is routed towards the closest origin server. If the content is found on an on-path router, then the request is "shortcut".

Two main dimensions could be researched in this part, the cache size, and the content replacement strategy. We are taking into account 4 different cache sizes of 1GB, 10GB, 100GB, and 1 TB. So, considering that the total volume of different videos in our dataset is 137 TB, the sizes of caches could store approximately 0.0007%, 0.007%, 0.07%, and 0.7% of contents, correspondingly. Zero-cache scenario is used as a reference point and the following replacement algorithms are implemented: FIFO, LRU, LFU."

III. Simulation Results

In this work Video on Demand through PPTV is used for evaluating three algorithms, FIFO, LRU and LFU in order to understand the congestion in each of them.

According the presented congestion results in Fig. 2 we can see that FIFO algorithm has a value of 0.868421, LRU algorithm has a value of 0.651316, and LFU algorithm has a value of 0.578947. Hence, from the obtained results in Fig. 2, we can conclude that the LFU algorithm gives the best results for congestion from the presented algorithms. It has the lowest value of all three presented cache replacement algorithms.

A. Description of the Algorithm

In the presented model in this work where the cache ejection policy is based on the information for frequency, the node looks at its incoming request vector and compares the request frequency for the items held in the local cache with the requests for other items. The best option for the node regarding to its request vector r is to cache the items with the maximum average request rate. If $F_C(r)$ presents the function that returns the C largest values from r, then the cache probability vector for node k, c_k will be:

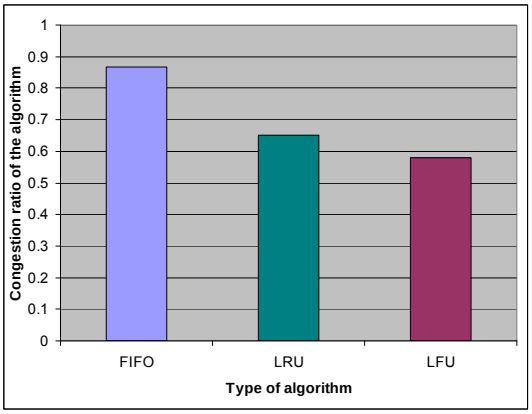


Figure 2. Congestion Ratio of the PPTV

$$c_k^{LFU} = 1_{i \in F_C(r_k)}. \quad (1)$$

where 1 presents the indicator function.

If the average request rates of this LFU model are static and well known, the best policy in steady state is to cache items with high average request rates. But, this is not an optimal strategy because the temporal variations in requests could initiate ejection of the items that are popular in the long run. Hence, this LFU model can provide us with an estimation of an average case not including the upper or lower performance bounds.

B. Cache Hit Ratio of the Algorithms

Average cache hit ratio across all routers could be expressed by the following equation:

$$HitRate = \frac{\sum_{r=1}^{R} \frac{Hit_r}{Hit_r + Misses_r}}{R}. \quad (2)$$

In the above equation Hit_r presents the number of requests that router r deals with from its cache, and $Misses_r$ presents the number of requests that are forwarded upstream because of the cache miss.

Fig. 3 presents obtained results for the cache hit ratio of the LFU algorithm for different cache sizes starting from 1 GB to 1 TB. We can conclude that when the cache size is increased from 1 GB to 1 TB, the cache hit ratio also increases. If the cache size is 1 GB, the cache hit ratio is 0.007324. When the cache size is 10 GB, the cache hit ratio increases to 0.026257. For 100 GB of cache size, the cache hit ratio is 0.059332, and for 1 TB cache size, the cache hit ratio has the highest value of 0.108125.

In the next diagrams we will present the cache hit ratio of all three replacement algorithms for cache sizes of 1G, 10G, 100G, and 1T.

Fig. 4 presents the results of the cache hit ratio of all algorithms when the cache size is 1G. Because the cache size has low value, we cannot clearly see the performances of these algorithms expressed in cache hit ratio. The values of cache hit ratio here are nearly the same. But, in Fig. 5, when the cache size is increased to 10G, we can already notice that the cache hit ratio of the LFU algorithm is drastically increased comparing to other two algorithms. Reason for this behavior is that when the cache size is 1GB, the cache hit ratio and the average reduction of video traffic is lower than 0.009 across all three caching strategies. But, when the cache size is increased to 10GB, the cache hit ratio is significantly improved. LFU is drastically increased comparing to two algorithms, because when a put call is made for a new element, the element with least number of hits or the Least Frequently Used element is evicted. So, LFU is an algorithm that takes a random sample of the elements and evicts the smallest. That is why the cache hit ratio is better in this case.

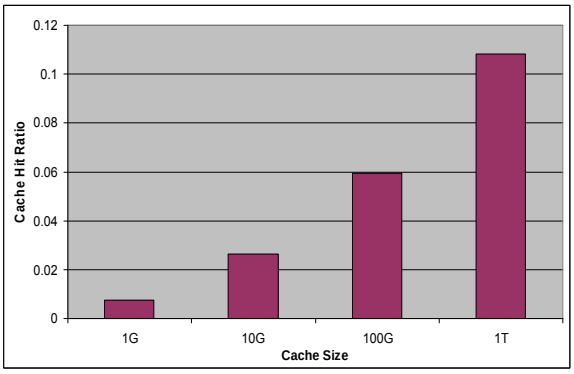


Figure 3. Cache Hit Ratio of the LFU Cache Replacement Algorithm

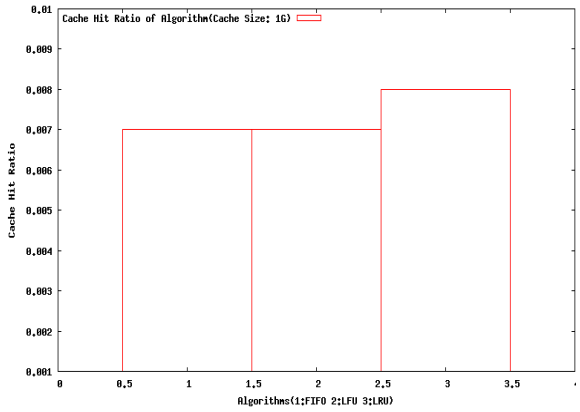


Figure 4. Cache Hit Ratio of the algorithms for 1 GB of cache size

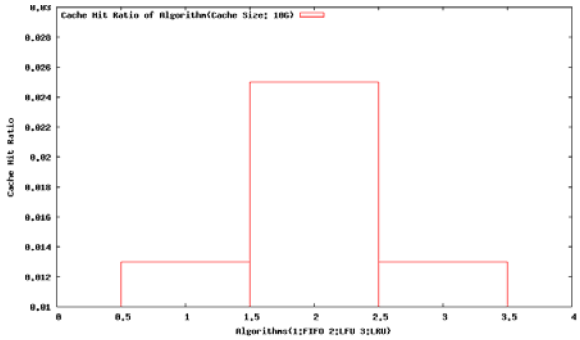


Figure 5. Cache Hit Ratio of the algorithms for 10 GB of cache size

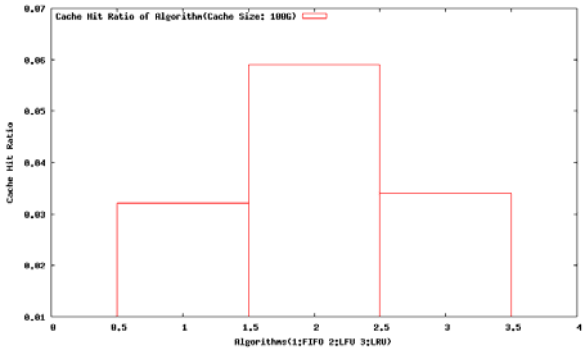


Figure 6. Cache Hit Ratio of the algorithms for 100 GB of cache size

Fig. 6 presents the results of the cache hit ratio for all three algorithms when the cache size is 100 GB. We can conclude here that LFU algorithm has the largest value of cache hit ratio similarly to Fig. 5. But, here the difference to other two algorithms is even greater comparing to Fig. 5.

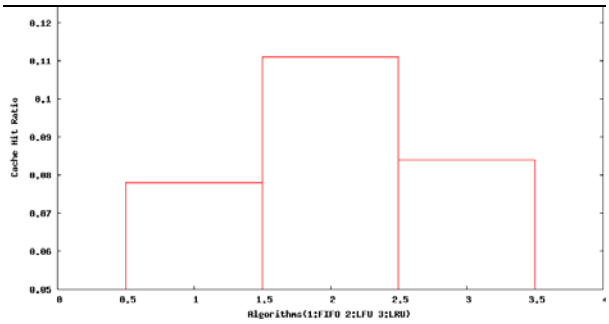


Figure 7. Cache Hit Ratio of the algorithms for 1 TB of cache size

Fig. 7 presents the cache hit ratio of all three algorithms when the cache size is 1 TB. We can conclude here that again LFU algorithm gives the largest value for the cache hit ratio.

Generally in all diagrams from Fig. 4 to Fig. 7, the cache hit ratio increases in all cases for all three algorithms as the cache size increases. If we compare the values of the cache hit ratio of the LFU algorithm to the other two algorithms, they are largest when the cache size is 10G and above.

We can explain the significant benefits of the cache hit ratio at 100GB and 1TB cache sizes in Fig. 6 and Fig. 7 as a result of the large traffic volume of video. So, if the cache hit ratio for 1GB of cache size on each router is less than 0.009 for all three algorithms, for 100GB cache hit ratio is between 0.03 and 0.06 for all three algorithms, and for 1TB is between 0.075 and 0.15.

C. *Analysis of the Performance Metric Throughput*

The throughput as a performance metric presents the number of bits that are sent in one second. It can be calculated with the following equation:

$$MT = \frac{Bytes * 8[bits]}{Time[s]}. \quad (3)$$

In the above equation *Bytes* refer to the number of bytes that are sent or received. This value is multiplied by 8 to give the total number of bits, because one byte has 8 bits. The value of *Time* is expressed in seconds and it refers to the time of the sent or received bytes or bits. So, throughput *MT* is expressed with the number of transmitted bits per second.

Fig. 8 depicts the results of the throughput for the LFU cache replacement algorithm expressed in megabytes or MB. The same results with exact numbers are presented in Table 1. From the obtained results, we can see that the results in throughput are given for each second starting from 1 to 25 seconds. It is clear from the Fig. 8 that the throughput results are variable in the observed 25 seconds. The lowest value for throughput is 109.342 MB on the 2nd second and the highest value is 145.309 MB on the 13th second. So, throughput results

are fluctuating between these two extremes in the presented 25 seconds of the experiment.

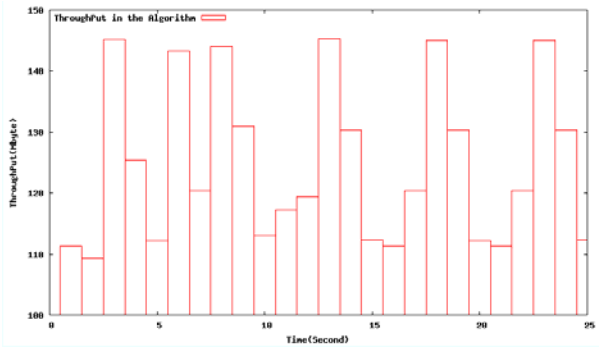


Figure 8. Throughput results of the LFU cache replacement algorithm expressed in megabytes

TABLE I. THROUGHPUT RESULTS OF THE LFU ALGORITHM RELATED TO FIG. 8

Time [s]	*Throughput [MB]*
1	111.345
2	109.342
3	145.091
4	125.345
5	112.235
6	143.345
7	120.342
8	144.092
9	130.945
10	113.023
11	117.345
12	119.342
13	145.309
14	130.345
15	112.253
16	111.345
17	120.342
18	145.049
19	130.345
20	112.237
21	111.345
22	120.342
23	145.019
24	130.345
25	112.253

Hence, we can conclude from the Fig. 8 and Table 1 that throughput results of the LFU algorithm are not stable, and they are varying between 109.342 and 145.309 MB. Anyway, we can also conclude that they are maintaining satisfactory level of throughput, because we don't have sudden drops in throughput.

D. Analysis of the Performance Metric Server Load

Server load as one of the main performance metrics could be calculated using the following equation:

$$Sl = \frac{NumCached\ \mathrm{Re}\ q}{v}. \qquad (4)$$

In the above equation *NumCachedReq* presents the number of requests that are served from some in-network cache. On the other hand, the variable *v* presents all video requests.

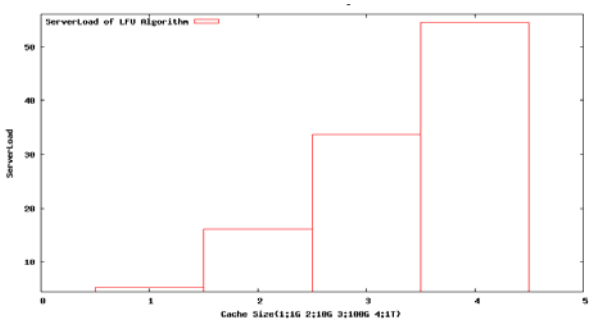


Figure 9. Server load of the cache replacement LFU algorithm expressed in %

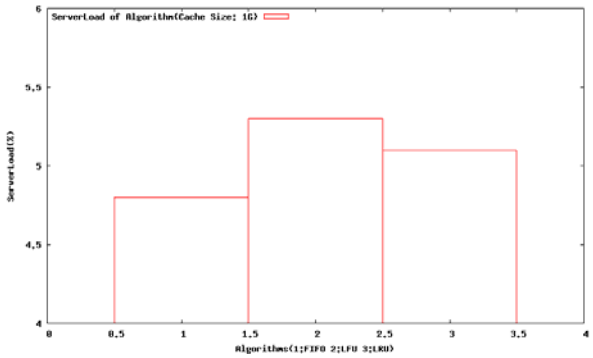


Figure 10. Server load of all three algorithms for cache size of 1G

Fig. 9 presents the server load results for the LFU algorithm expressed in % for different cache sizes from 1G to 1T. It is obvious from the diagram that server load is increased as the cache size increases from 1G to 10G, 100G, 1T. It is clearly shown that the server load for 1G is about 1% and for 1T more than 50%.

In the following diagrams we will present the values of the server load for all three replacement algorithms for different cache sizes.

Fig. 10 presents the results of server load expressed in % for all three algorithms when the cache size is 1G. It is obvious that the server load is largest when LFU algorithm is used, and FIFO gives the lowest server load value. So, in this case FIFO algorithm gives the most favorable results because lower server load gives better overall performances. Anyway, differences between the results of all three algorithms in this case are not so high, because they are moving between the server load of 4.5 and 5.5%.

Fig. 11 depicts the results of server load for all three algorithms when the cache size is 10G. Here we can see that server load of the LFU algorithm is around 16%, but the server load of FIFO and LRU algorithms are almost the same – less than 9%.

Fig. 12 presents the server load results for all three algorithms when the cache size is 100G. From the shown results it is obvious that again LFU algorithm has the greatest server load of around 34%. FIFO algorithm has less than 20%, and LRU algorithm has more than 20% server load.

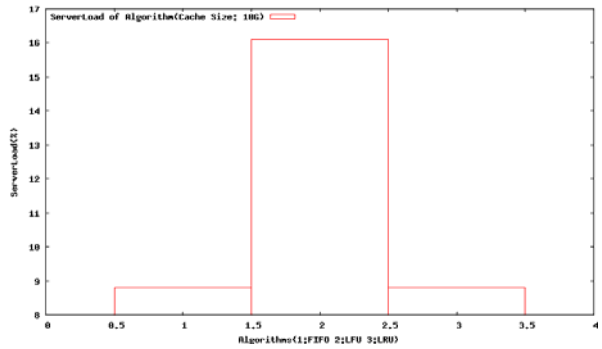


Figure 11. Server load of all three algorithms for cache size of 10G

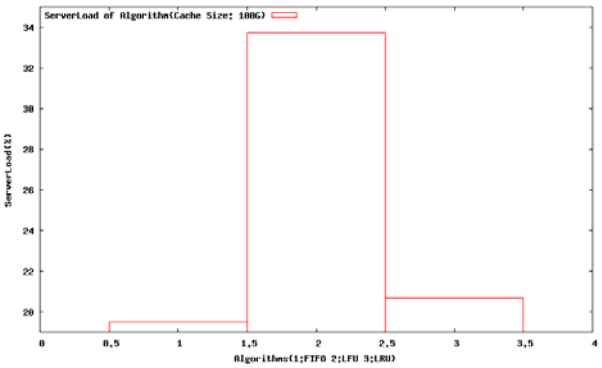


Figure 12. Server load of all three algorithms for cache size of 100G

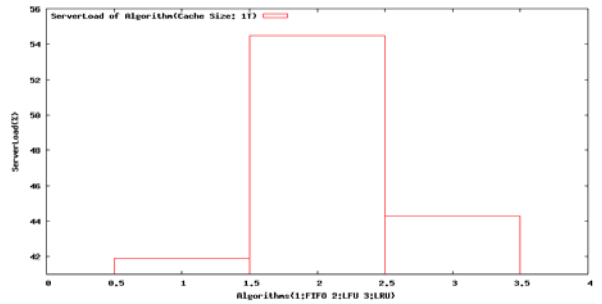


Figure 13. Server load of all three algorithms for cache size of 1T

In Fig. 13 we show the results of server load of FIFO, LFU, and LRU algorithm when the cache size is 1T. Server load for LFU algorithm gives the highest results – more than 54%. FIFO algorithm gives around 42% of server load and LRU around 44%.

From the presented diagrams from Fig. 10 to Fig. 13, generally, we can conclude that the server load is increased when the cache size is increasing for all three algorithms. We can also confirm that the server load of the LFU algorithm is larger than other algorithms for each of the cache sizes. This difference is larger as cache size increases from 1G to 1T. So, we can conclude that LFU algorithm gives the highest results when we are analyzing the performance metric *server load*.

IV. CONLUSION

In this work we have evaluated three cache replacement algorithms, LFU, LRU, and FIFO in order to understand better the congestion for each of them. We have used video on demand through PPTV network to get the necessary results.

From the obtained results we have got the best results of congestion for the LFU algorithm. The congestion was lowest when LFU algorithm was used in the experiments.

After this part of the work we have examined the performance metrics of cache hit ratio, throughput, and server load for the previously selected algorithm with the lowest congestion or LFU algorithm. Additionally, we compared the results of this algorithm to other two algorithms, LRU and FIFO. In this context, we analyzed and gave conclusions particularly for each of the presented diagrams.

Generally, we can conclude that LFU algorithm, which was the algorithm with the lowest congestion, gives satisfactory results in the examined performance metrics. It gives stable results for throughput. Compared to other two algorithms, we can conclude that LFU algorithm gives the best results in the cache hit ratio and the highest results in server load.

REFERENCES

[1] Han, Dongsu, et al. "XIA: Efficient Support for Evolvable Internetworking." *NSDI*. Vol. 12. 2012.

[2] Zhang, Lixia, et al. "Named data networking (ndn) project." *Relatório Técnico NDN-0001, Xerox Palo Alto Research Center-PARC* (2010).

[3] Koponen, Teemu, et al. "A data-oriented (and beyond) network architecture."*ACM SIGCOMM Computer Communication Review*. Vol. 37. No. 4. ACM, 2007.

[4] Psaras, Ioannis, Wei Koong Chai, and George Pavlou. "Probabilistic in-network caching for information-centric networks." *Proceedings of the second edition of the ICN workshop on Information-centric networking*. ACM, 2012.

[5] Zhang, Guoqiang, Yang Li, and Tao Lin. "Caching in information centric networking: a survey." *Computer Networks* 57.16 (2013): 3128-3141.

[6] Kim, Yusung, and Ikjun Yeom. "Performance analysis of in-network caching for content-centric networking." *Computer Networks* 57.13 (2013): 2465-2482.

[7] Podlipnig, Stefan, and Laszlo Böszörmenyi. "A survey of web cache replacement strategies." *ACM Computing Surveys (CSUR)* 35.4 (2003): 374-398.

[8] Abrams, Marc, et al. "Removal policies in network caches for World-Wide Web documents." *ACM SIGCOMM Computer Communication Review*. Vol. 26. No. 4. ACM, 1996.

[9] Ming, Zhongxing, Mingwei Xu, and Dan Wang. "Age-based cooperative caching in information-centric networks." *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on*. IEEE, 2012.

[10] PPTV. http://www.pptv.com/.

[11] Fayazbakhsh, Seyed Kaveh, et al. "Less pain, most of the gain: Incrementally deployable icn." *ACM SIGCOMM Computer Communication Review*. Vol. 43. No. 4. ACM, 2013.

[12] Psaras, Ioannis, et al. "Modelling and evaluation of CCN-caching trees."*NETWORKING 2011*. Springer Berlin Heidelberg, 2011. 78-91.

[13] Rossini, Giuseppe, and Dario Rossi. "A dive into the caching performance of content centric networking." *Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2012 IEEE 17th International Workshop on*. IEEE, 2012.

[14] Tyson, Gareth, et al. "A trace-driven analysis of caching in content-centric networks." *Computer Communications and Networks (ICCCN), 2012 21st International Conference on*. IEEE, 2012.

BIOMEDICAL ENGINEERING

Diagnostic of Asthma Using Fuzzy Rules Implemented in Accordance with International Guidelines and Physicians Experience

Almir Badnjević[1,2,3,4], Lejla Gurbeta[1,2], Mario Cifrek[5], Damir Marjanović[2]
[1]Verlab d.o.o., Sarajevo, Bosnia and Herzegovina
[2]International Burch University, Sarajevo, Bosnia and Herzegovina
[3]University of Sarajevo, Faculty of Electrical Engineering Sarajevo, Bosnia and Herzegovina
[4]University of Sarajevo, Faculty of Medicine Sarajevo, Bosnia and Herzegovina
[5]University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
almir@verlab.ba

Abstract - This paper presents a system for classification of asthma based on fuzzy rules.
Fuzzy rules are defined according to Global Initiative for Asthma (GINA) guidelines, as well as through consultations with long-term experience of pulmologists. Our fuzzy system for classification of asthma is based on a combination of spirometry (SPIR) and Impulse Oscillometry System (IOS) test results, which are inputs to fuzzy system. Additionally, the use of bronchodilatation and bronhoprovocation enabled a complete patient's dynamic assessment rather than a simple static assessment.
The system was retroactively tested with 1250 Medical Reports established by pulmologists, out of which 728 were diagnosed with asthma and 522 were healthy subjects. Sensitivity and specificity were assessed, on this dataset, which were 91.89% and 95.01%, respectively.

I. INTRODUCTION

Asthma is an inflammatory impairment of airways, which as result becomes hyperreactive and generates increased mucus, mucosal swelling and contraction of smooth airway muscles, all of which contribute to airway obstruction. In addition, the prevalence of asthma has been increasing in the recent decades, as well as the costs for asthma care [1-3]. One of the problems related to asthma management is that a number of patients with asthma are either improperly diagnosed or misdiagnosed for having other respiratory diseases such as a common cold, acute bronchitis or chronic obstructive pulmonary disease (COPD) [4, 5].

All presented asthma problems justify the need for the Expert System, which can help physicians diagnose this disease much more successfully. The most commonly used pulmonary function tests in asthma detection are Spirometry (SPIR) and Impulse Oscillometry System (IOS) [7][8].

Several Fuzzy Logic Expert Systems for the detection of different types of diseases have been proposed. In the [9], the authors presented a systematic approach for design and identification of tuberculosis using fuzzy based decision support system. Relation between different input attributes and its symptoms were discussed and it was concluded that fuzzy based expert system can be used for diagnosis of all other respiratory diseases. In the other study [10], the authors classified respiratory abnormalities using adaptive neuro fuzzy inference system based only on SPIR or IOS measurements. In [11], the authors performed the task of separation of lung sounds, obtained from patients with pulmonary pathology using recurrent neuro-fuzzy filter. The authors implemented a filter throw a pipelined Takagi-Sugeno-Kang recurrent fuzzy network, consisting of a number of modules interconnected in a cascaded form. Those studies focused on static assessment of the patient using combination of spirometry and/or IOS measuring test results.

Conversely, dynamic assessment of the patient is necessary process according to Global Initiative for Asthma (GINA) guidelines [12]. In order to get dynamic assessment of the patient, it is necessary to take the patient's symptoms and allergies in consideration, to perform auscultation of the patient, and to apply the bronchial dilation (BDT) and/or bronchial provocation (BPT) test. After BDT and/or BPT treatment and after the second/third measurement of lung function, potential changes of pulmonary parameter values may be present, from which physicians can get accurate information on the specifics of the disease. In previous works [13, 18-22] we presented different neuro-fuzzy systems for full dynamic assessment of the patient and those paper results were a good baseline for improving fuzzy logic rules in this research.

This paper presents an Expert System based on fuzzy rules for diagnosis of asthma with full dynamic assessment of the patient results.

II. METHODS AND MATERIALS

Input data of fuzzy rules system were test results of spirometry (SPIR) and IOS. The outputs of the fuzzy rules were classifications of asthma and healthy subjects. Symptoms based on GINA guidelines together with fuzzy logic system classification results provided final proposed diagnoses. A block diagram of the Fuzzy Logic System

for asthma classification was presented in Fig. 1.

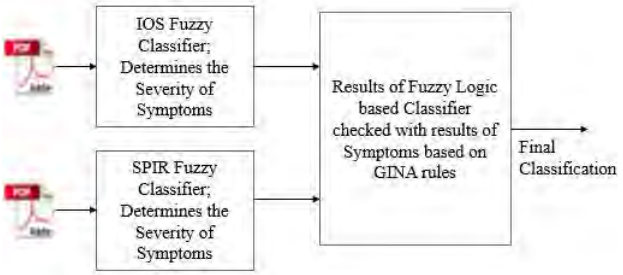


Fig. 1. Block diagram of Expert System for asthma classification

GINA rules for Asthma symptoms assessment considered: history of recurrent wheezing, cough, chest tightness or shortness of breath, in accordance to allergic symptoms, which could have a seasonal pattern or might exacerbate at night or by common exposure to allergen, exercise or smoking. Patients with positive symptoms and Resistance at 5 Hz (R5) > 150% pred, Resistance at 20 Hz (R20) > 150% pred, Reactance at 5 Hz (X5pred - X5) > 0.15 kPa/(L/s), Ratio of Forced Expiratory Volume in 1 second and Forced Vital Capacity (FEV1/FVC) < 0.8 were diagnosed with asthma. Patients with nonspecific test results were dedicated for bronchial hyper reactivity (BHR), first with BDT, and then with BPT. Patients were diagnosed with asthma when their R5 decreases by more than 25% (ΔR5), the absolute value of X5 decreases by more than 20% (ΔX5) and FEV1 decreases by more than 12% of baseline value and 200 ml. In cases where BDT did not achieve any improvement for final diagnose, patients were examined with methacholine BPT. The doses of methacholine, serially increased, were administered, and FEV1 was measured after each methacholine dose until a 20% decrease in FEV1. After this procedure, the system was able to make the final classification of asthma.

The fuzzy logic system presented in this paper has been designed according to GINA guidelines and experience of other researchers [14,15]. The review of disease classification based on measured values of spirometry was reported in Table I.

TABLE I
REVIEW OF DISEASE CLASSIFICATION BASED ON MEASURED VALUES OF SPIROMETRY

	Stage of disease			
Parameters	**Normal**	**Mild**	**Moderate**	**Severe**
FVC	> 80%	80 – 60%	60-40%	< 40%
FEV1	> 80%	80 – 60%	60-40%	< 40%
FEV1/FVC	> 0.7	0.7 – 0.6	0.6-0.5	< 0.5
PEF	> 80%	80 – 60%	60-40%	< 40%

Input variables for the implemented fuzzy system for the case of spirometry were values of Forced Vital Capacity (FVC), Forced Expiratory Volume in 1 second (FEV1), Peak Expiratory Flow (PEF) and ratio FEV1/FVC, which were obtained from spirometry test report, after pulmonary function testing.

Their fuzzy values were defined by the equations from 1 to 12.

$$\mu_{normal}(FVC) = \begin{cases} 0, & 0 < FVC < 75 \\ \frac{FVC-75}{5}, & 75 < FVC \leq 80 \\ 1, & FVC > 80 \end{cases} \quad (1)$$

$$\mu_{mild}(FVC) = \begin{cases} 0, & 0 < FVC < 55 \\ \frac{FVC-55}{15}, & 55 \leq FVC \leq 70 \\ \frac{85-FVC}{5}, & 70 < FVC \leq 85 \\ 0, & FVC > 85 \end{cases} \quad (2)$$

$$\mu_{moderate}(FVC) = \begin{cases} 0, & 0 < FVC < 40 \\ \frac{FVC-40}{10}, & 40 \leq FVC \leq 50 \\ \frac{60-FVC}{10}, & 50 < FVC \leq 60 \\ 0, & FVC > 60 \end{cases} \quad (3)$$

$$\mu_{severe}(FVC) = \begin{cases} 1, & 0 < FVC < 40 \\ \frac{45-FVC}{5}, & 40 \leq FVC \leq 45 \\ 0, & FVC > 45 \end{cases} \quad (4)$$

where: *μnormal*(FVC) was membership function of reference values of FVC from SPIR, *μmild*(FVC) was membership function of mild stage of disease based on FVC from SPIR, *μmoderate*(FVC) was membership function of moderate stage of disease based on FVC from SPIR, *μsevere*(FVC) was membership function of severe stage of disease based on FVC from SPIR.

$$\mu_{normal}(FEV1) = \begin{cases} 0, & 0 < FEV1 < 75 \\ \frac{FEV1-75}{5}, & 75 \leq FEV1 \leq 80 \\ 1, & FEV1 > 80 \end{cases} \quad (5)$$

$$\mu_{mild}(FEV1) = \begin{cases} 0, & 0 < FEV1 < 58 \\ \frac{FEV1-58}{8}, & 58 \leq FEV1 \leq 66 \\ \frac{80-FEV1}{14}, & 66 < FEV1 \leq 80 \\ 0, & FEV1 > 80 \end{cases} \quad (6)$$

$$\mu_{moderate}(FEV1) = \begin{cases} 0, & 0 < FEV1 < 40 \\ \frac{FEV1-40}{10}, & 40 \leq FEV1 \leq 50 \\ \frac{60-FEV1}{10}, & 50 < FEV1 \leq 60 \\ 0, & FEV1 > 60 \end{cases} \quad (7)$$

$$\mu_{severe}(FEV1) = \begin{cases} 1, & 0 < FEV1 < 40 \\ \frac{45-FEV1}{5}, & 40 \leq FEV1 \leq 45 \\ 0, & FEV1 > 45 \end{cases} \quad (8)$$

Where: *μnormal*(FEV1) was membership function of reference values of FEV1 from SPIR, *μmild*(FEV1) was membership function of mild stage of disease based on FEV1 from SPIR, *μmoderate*(FEV1) was membership function of moderate stage of disease based on FEV1 from SPIR, *μsevere*(FEV1) was membership function of severe stage of disease based on FEV1 from SPIR.

$$\mu_{normal}(PEF)=\begin{cases}0, & 0<PEF<79\\ PEF-79, & 79\le PEF\le 80\\ 1, & PEF>80\end{cases} \quad (9)$$

$$\mu_{mild}(PEF)=\begin{cases}0, & 0<PEF<59\\ \frac{PEF-59}{11}, & 59\le PEF\le 70\\ \frac{81-PEF}{11}, & 70<PEF\le 81\\ 0, & PEF>81\end{cases} \quad (10)$$

$$\mu_{moderate}(PEF)=\begin{cases}0, & 0<PEF<38\\ \frac{PEF-38}{12}, & 38\le PEF\le 50\\ \frac{62-PEF}{12}, & 50<PEF\le 62\\ 0, & PEF>62\end{cases} \quad (11)$$

$$\mu_{severe}(PEF)=\begin{cases}1, & 0<PEF<40\\ \frac{42-PEF}{2}, & 40\le PEF\le 42\\ 0, & PEF>42\end{cases} \quad (12)$$

where: *μnormal*(PEF) was membership function of reference values of PEF from SPIR, *μmild*(PEF) was membership function of mild stage of disease based on PEF from SPIR, *μmoderate*(PEF) was membership function of moderate stage of disease based on PEF from SPIR, *μsevere*(PEF) was membership function of severe stage of disease based on PEF from SPIR.

Input values can be calculated by using fuzzy T-norm operator [16].

Based on the following rules, we defined the outputs for the diagnosis that can be classified by using SPIR report:

1. If (FVC was normal) and (FEV1 was normal) and (FEV1/FVC was normal) and (PEF was normal) THAN (output was NORMAL);
2. If (FVC was mild) and (FEV1 was mild) and (FEV1/FVC was mild) and (PEF was mild) THAN (output was MILD);
3. If (FVC was moderate) and (FEV1 was moderate) and (FEV1/FVC was moderate) and (PEF was moderate) THAN (output was MODERATE);
4. If (FVC was severe) and (FEV1 was severe) and (FEV1/FVC was severe) and (PEF was severe) THAN (output was SEVERE);

where: NORMAL, MILD, MODERATE and SEVERE were stages of disease, while output was suggestion of diagnosis based on spirometry measurement.

Input variables for the implemented system of IOS fuzzy decisions were values of respiratory resistance R5 at 5 Hz, proximal resistance R20 at 20 Hz and lung reactance X5 at 5 Hz obtained after measurements of lung function. Their values of fuzzy decision-making were defined with equations from 13 to 20.

$$\mu_{normal}(R5)=\begin{cases}1, & 0<R5<150\\ \frac{R5-145}{5}, & 150\le R5\le 155\\ 0, & R5>155\end{cases} \quad (13)$$

$$\mu_{high}(R5)=\begin{cases}0, & 0<R5<150\\ \frac{R5-150}{5}, & 150\le R5\le 155\\ 1, & R5>155\end{cases} \quad (14)$$

$$\mu_{normal}(X5)=\begin{cases}1, & X5<0{,}15\\ \frac{0{,}155-X5}{0{,}005}, & 0{,}15\le X5\le 0{,}155\\ 0, & X5>0{,}155\end{cases} \quad (15)$$

$$\mu_{high}(X5)=\begin{cases}0, & X5<0{,}15\\ \frac{X5-0{,}15}{0{,}005}, & 0{,}15\le X5\le 0{,}155\\ 1, & X5>0{,}155\end{cases} \quad (16)$$

$$\mu_{none}(\Delta R)=\begin{cases}1, & \Delta R<0\\ 1-\frac{\Delta R}{0{,}05}, & 0\le \Delta R\le 0{,}05\\ 0, & \Delta R>0{,}05\end{cases} \quad (17)$$

$$\mu_{min}(\Delta R)=\begin{cases}0, & \Delta R<0{,}05\\ \frac{\Delta R-0{,}05}{0{,}075}, & 0{,}05\le \Delta R\le 0{,}125\\ \frac{0{,}125-\Delta R}{0{,}075}, & 0{,}125\le \Delta R\le 0{,}2\\ 1, & \Delta R>0{,}2\end{cases} \quad (18)$$

$$\mu_{sig}(\Delta R)=\begin{cases}0, & \Delta R<0{,}2\\ \frac{\Delta R-0{,}2}{0{,}15}, & 0{,}2\le \Delta R\le 0{,}35\\ 1, & \Delta R>0{,}35\end{cases} \quad (19)$$

whereby: μnormal(R5) was a reference measurement function R5 from IOS measurements, μhigh(R5) was a function of upper normal value (ULN) of R5, μnormal(X5) was a reference measurement function X5, μhigh(X5) was a function of upper normal value of X5, μnone(ΔR) was a function in case of no difference in value between R5 a d R20, μmin(ΔR) was a function in case when the difference in values between R5 and R20 was minimal, and μsig(ΔR) was a function in the case when the difference between R5 and R20 was significant. The values of fuzzy decision for R20 were the same as for R5. Input values were calculated using the T-norm operator fuzzy decision [16].

Based on the following rules, we defined the outputs for the diagnosis that can be classified by using IOS report:

1. IF (R5 was normal) and (R20 was normal) and (R5-R20=0) and (X5 was normal) THEN (output1is NSCO-IV);
2. IF (R5 was normal) and (R20is normal) and (R5-R20 was minimal) and (X5 was normal) THEN (output1 was NSCO-RV);
3. IF (R5 was normal) and (R20 was normal) and (R5-R20 was significant) and (X5 was normal) THEN (output1 was NSPO);
4. IF (R5 was ULN) and (R20 was ULN) and (R5-R20 = 0) and (X5 was normal) THEN (output1 was CO-IV);
5. IF (R5 was ULN) and (R20 was ULN) and (R5-R20 was minimal) or (X5 was normal) or (X5 was ULN) THEN (output1 was CO-RV);
6. IF (R5 was ULN) and (R20 was ULN) and (R5-R20 was significant) and (X5 was ULN) THEN (output1 was PO);

where NSCO-IV refers to Negative Sub- Central Obstruction – of Identical Values, NSCO-RV refers to Negative Sub-Central Obstruction - Differing Values, NSPO refers to Negative Sub-Peripheral Obstruction, CO-IV refers to Central Obstruction – of Identical Values, CO RV's Central Obstruction – of Differing Values, PO refers to Peripheral Obstruction and Out1 was one of the proposals of classification based on the findings of the IOS.

The results of the fuzzy classification based on the reports of IOS and SPIR obtaining final classification or suggesting additional tests. A prospective study was designed in order to test the proposed expert system in the Hospital Sarajevo Bosnia and Herzegovina. Ethics board approval for human subject from Hospital Sarajevo was obtained before starting the study. All the volunteers gave written informed consent. The patients were selected by the pulmonary specialists. The same medical reports were used in [23].

Finally, 1250 consecutive patients were enrolled in this study, 728 resulted diagnosed with asthma, while 522 patients resulted healthy. Basic information about patients involved in the study were reported in Table II.

TABLE II
BASIC INFORMATION ABOUT PATIENTS INVOLVED IN THIS STUDY

	Asthma	Healthy	p-value*
Number	728	522	-
N. of female	352	297	0.32
Other diseases	628	143	<0.01
Smoking	695	317	<0.01
Allergic subjects	224	184	0.41
Medications	544	348	1.47

*CHI-square

For the purposes of this study, baseline assessment for all patients was performed in order to capture information regarding symptoms and allergies, history and risk factors of asthma according to GINA guidelines.

In another step, lung function was measured incorporating SPIR and IOS test results. All pulmonary function test reports were obtained using the CareFusion Germany "Master Screen IOS" device, which allows performing spirometric and impulse oscillometric lung function tests.

Using spirometry, FVC and FEV1 were derived, while the ratio FEV1/FVC was calculated. The impulse oscillometry system (IOS) provided respiratory resistance R5 at 5 Hz, proximal resistance R20 at 20 Hz, lung reactance X5 at 5 Hz and resonant frequency Fres [17]. All measured results were compared with predicted (pred) values of the patient.

If the patients had nonspecific test results then it was necessary to do a bronchial dilation test (BDT) and/or bronchial provocation test (BPT). After this procedure, the Fuzzy Logic System was used to suggest diagnosis. This diagnosis was blinded to the medical doctors and then disclosed after clinicians formalized their diagnosis. Suggested diagnoses were finally compared with clinician ones and standard metrics to assess fuzzy logic system performance where used, including accuracy, specificity and precision (percentage of true classification).

III. RESULTS

Results of this study showed a high efficiency of integrated software suite in the classification of asthma disease (91.89%) and healthy patients (95.01%). In Table III are expressed percentages of hits and misses of the integrated software suite in the classification of asthmatics and healthy patients who were involved in this study.

TABLE III. EFFECTIVENESS OF THE IMPLEMENTED SYSTEM IN THE CLASSIFICATION OF ASTHMA PATIENTS AND HEALTHY PATIENTS

Patients with asthma		Healthy patients	
728		522	
Hits by software	Misses by software	Hits by software	Misses by software
669 (91.89%)	59 (8.11%)	496 (95.01%)	26 (5.00%)

From table III we can see that in the case of patients affected with asthma the system showed a small percentage of false positive results (8.11%) as compared to physician diagnosis. In these 59 cases, the system classified those patients with other forms of respiratory diseases, not as asthma. In the case of healthy patients, the software also showed a small percentage of false positive results (5.00%) as compared to physician's diagnosis. In these 26 cases, the system classified those patients as not healthy. Testing the efficiency of system at each single step of classification of asthma patients and healthy controls were obtained from the results of this study. These results are presented in Table IV. Based on the study results shown in Table IV it's clear that the implemented system classifies the most patients with asthma in a second step, i.e. after bronchial dilation test (68.90%). In these cases of patients R5 increased by more than 25%, X5 increased by more than 20% and FEV1 decreased by more than 12%.

In the first step, i.e. after the results of IOS and spirometry, the system has classified 9.42% of patients with asthma, while in the third step, i.e. after bronchial provocative test, classified 21.68% of patients with asthma. These patients have done methacholine bronchial provocation test and R5 increased by more than 35% and Fres increased by more than 30%.

TABLE IV. HITS AND MISSES OF IMPLEMENTED SYSTEM AT EACH SINGLE STEP OF CLASSIFICATION IN ASTHMATICS AND HEALTHY CONTROLS

Steps for classification of asthma	Criteria for classification of asthma	Patients with asthma (Hits by software) Σ = 669	Patients with asthma (Misses by software) Σ = 59	Healthy controls (Hits by software) Σ = 496	Healthy controls (Misses by software) Σ = 26
IOS and spirometry test results	R5 > 150% R20>150% X5 > 0.20 FEV1/FVC < 0.8	63 (9.42%)	0 (0%)	465 (93.67%)	17 (65.38%)
Bronchial dilation test results	R5 > 25% Fres > 20% FEV1 < 12	461 (68.90%)	18 (30.50%)	21 (4.23%)	6 (23.07%)
Bronchial provocation test results	R5 > 35% Fres > 30%	145 (21.68%)	41 (69.50%)	10 (2.10%)	3 (11.55%)

R5 - Total respiratory resistance
R20 - Proximal resistance
X5 - Distal capacitive reactance
FEV1 - Forced expiratory volume in one second
Fres - Resonant frequency

In the case of healthy controls, after the first step software classified 93.67% of patients as healthy controls. These patients had normal IOS and spirometry test results (also and symptoms). In the second and third step, where the test results of the patients contained minor deviations from normal predictive values, the percentage were 4.23% and 2.10%, respectively.

IV. CONCLUSION

In this work we present results of classification of asthma using implemented fuzzy system. Based on the total number of 1250 patients involved in this study, the software correctly classified asthma in 91.89% of cases, whereas in the case of healthy subjects correctly performed the classification in 95% of cases. Also, it is shown that for classification of diseases implemented fuzzy system takes into account the complete dynamic picture of patients. In this paper we also presented that the clinical evaluation of lung function is a more comprehensive with developed software.

The novelty of the presented paper for asthma classification and previous work of the researchers [18-22] is that this solution is implemented to make diagnosis of asthma without any help of artificial neural network. Because of its simplicity this solution can be very good additional tool for all pulmonary measurement commercial software's, what authors demonstrated during 6 months clinical testing period.

REFERENCES

[1] G.J. Tortora and S.R. Grabowski. *Principles of Anatomy and Physiology*. 10th ed. New York, USA: John Wiley & Sons, 2003.

[2] Mayo Foundation for Medical Education and Research, Mayo Clinic: COPD Definition. Available: http://www.mayoclinic.com/health/copd/DS00916

[3] N. Anthonisen. "Chronic obstructive lung disease," in *Cecil Medicine*, 24th ed. L. Goldman, D. Auseillo, Ed. Philadelphia, USA, 2007, ch. 88.

[4] European Lung Function: COPD Burden in Europe. Available http://www.european-lung-foundation.org/index.php?id=63

[5] J. Ferlay. "Cancer Incidence, Mortality and Prevalence Worldwide," *IARC Cancer Base*. IARC Press, Lyon, France, vol. 5., 2004.

[6] C.J. Murray, A.D. Lopez. "Alternative projections of mortality and disability by cause 1990-2020," World Health Organization, Global Burden of Disease Study 1997, pp. 1498-504, 1997.

[7] K.G. Nielsen, H. Bisgaard. "The Effect of Inhaled Budesonide on Symptoms, Lung Function, and Cold Air and Methacholine Responsiveness in 2- to 5-year-old Asthmatic Children," *American Journal of Respiratory and Critical Care Medicine*., vol. 162, pp. 1500-1506, 2005.

[8] T.W. Song. "Correlation between spirometry and impulse oscillometry in children with asthma," *ActaPediatrica*, vol. 97, pp. 51-54, 2008.

[9] N. Walia, S.K. Tiwari, R. Malhotra. "Design and Identification of Tuberculosis using Fuzzy Based Decision Support System," *Advances in Computer Science and Information Technology*, vol. 2, pp. 57-62, 2015.

[10] M. Asaithambi, S.C. Manoharan, S. Subramanian. "Classification of Respiratory Abnormalities Using Adaptive Neuro Fuzzy Inference System," *Int. Inf. and Database SysLecture Notes in Computer Science,* vol. 7198, pp 65-73, 2012.

[11] D. Stavrakoudis, P. Mastorocostas, J. Theocaris "A Pipelined Recurrent Fuzzy Neural Filter for the separation of Lung Sounds," *IEEE International Fuzzy Systems Conference*, pp 1-6, 2007.

[12] Global Initiative for Asthma: Pocket guide to Asthma Diagnosis, Management and Prevention. http://www.ginasthma.org/local/uploads/files/GINA_Report_2014_Aug12.pdf, Accessed 05 March 2015

[13] Badnjevic, A., Cifrek, M., Koruga, D., & Osmankovic, D. (2015). Neuro-fuzzy classification of asthma and chronic obstructive pulmonary disease. BMC Med Inform Decis Mak BMC Medical Informatics and Decision Making, 15(Suppl 3).

[14] S.K. Anand, R. Kalpana, S. Vijayalakshmi. "Design and Implementation of a Fuzzy Expert System for Detecting and Estimating the Level of Asthma and Chronic Obstructive Pulmonary Disease," Middle-East Journal of Scientific Research, vol.14, pp 1435-1444, 2013.

[15] N. Mishra, D.B.V. Singh, M.K. Bandil, P. Sharma. "Decision Support System for Asthma," International Journal of Information and Computation Technology, vol. 3, pp. 549-554, 2013.

[16] J. Fodor J. Left-continuous t-norms in fuzzy logic. An overview. Acta Polytechnica Hungarica 2, 2004.

[17] U. Uncu, E. Koklukaya, A. Gencsoy. "A fuzzy rule-base model for classification of spirometric FVC graphs in chronical obstructive pulmonary diseases," in 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Buenos Aires, Argentina, 2010, pp 3866 – 3869.

[18] Badnjevic, A, Koruga, D, Cifrek, M, Smith, H.J., Bego, T. „Interpretation of pulmonary function test results in relation to asthma classification using integrated software suite", IEEE 36. International convention on information and communication technology, electronics and microelectronics (MIPRO), pp: 140 - 144, 21.-25. May 2013. Opatija, Croatia.

[19] Badnjevic, A., Cifrek, M., Koruga, D. Classification of Chronic Obstructive Pulmonary Disease (COPD) Using Integrated Software Suite. XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013, Vol. 41, pp 911-914.

[20] Badnjevic, A., Cifrek, M. Classification of asthma utilizing integrated software suite, 6th European Conference of the International Federation for Medical and Biological Engineering (MBEC), 07-11. September 2014., Dubrovnik, Croatia.

[21] Badnjevic, A., Cifrek, M., Koruga, D. Classification of Chronic Obstructive Pulmonary Disease (COPD) using integrated software suite, IFMBE XIII Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON), 25-28. September 2013., Sevilla, Spain.

[22] Badnjevic, A., Cifrek, M., Koruga, D. Integrated software suite for diagnosis of respiratory diseases, IEEE International conference on computer as tool (EUROCON), pp: 564-568, 01-04. July 2013., Zagreb, Croatia

[23] Badnjevic, A., Gurbeta, L., Cifrek, M., Marjanovic, D., Classification of Asthma Using Artifical Neural Network. Submitted to IEEE 39. International convention on information and communication technology, electronics and microelectronics (MIPRO), May 2016. Opatija, Croatia.

Robust beat detection on noisy differential ECG

Primož Lavrič, Matjaž Depolli
Jožef Stefan Insitute, Department of Communication Systems, Ljubljana, Slovenia
pl9506@student.uni-lj.si, matjaz.depolli@ijs.si

Abstract – The differential ECG measurement is performed by a small gadget attached to a person's chest. In contrast to the standard 12-channel ECG, which is most often short and measured on a resting subject, the small form factor of the gadget allows for several days long measurement of active subjects. The resulting measurements are not only novel in their time and activity coverage, but they also require novel approach to their analysis. First, ECGs are noisier because the electrodes are placed close together and because measurements are made on physically active subjects. Second, the gadget is not always placed on by a trained professional and can be miss-oriented, producing a wide range of ECG orientations. Finally, the ECGs are measured with low sample ratio which helps conserve battery life of the gadget. In this work, we try to develop the first stage of such differential ECG analysis - algorithms for noise estimation and for beat detection. These novel algorithms have to be able to deal with noisy and sparsely sampled ECGs and with various beat shapes. Furthermore, these algorithms should be able to run in real-time applied to the ongoing measurement and on a very computationally weak portable devices, while maintaining power efficiency. Two algorithms for noise estimation and one for beat detection that fulfill those constraints are presented in this work.

I. Introduction

In this article, an algorithm for heartbeat detection accompanied by two noise estimation algorithms is presented. The main goal of these algorithms is to design a robust and efficient heartbeat detection on noisy differential ECG measurements, which were made with a small wireless gadget. Unlike the other heartbeat detectors that work on the standard 12-channel ECG, which usually monitors a resting subject, the proposed detector must work on a single channel measurement recorded on an active subject. Noise is difficult to filter out of such measurements because of the absence of multiple channels that would provide multiple viewpoints on the signal. The gadget itself may also often be used by untrained users, which may not attach it on optimal positions. That is why a wide spectrum of device orientations and noise levels must be taken in consideration when developing the algorithms for measurement analysis.

The rest of the paper is structured as follows. In section II, an overview of the state of the art beat detection algorithms is made. In section III, the procedure of differential ECG measurement by a wearable gadget follows. Then the heart beat detection procedures and the two noise estimators are presented in III. Section IV contains the description of tests and the results, separately for the beat detection and noise estimation. Finally, section V concludes the paper.

II. Beat detection algorithms

Most of the ECG beat detection algorithms focus on detection of QRS complex, since it is the most pronounced feature of the ECG. Within the QRS complex, the amplitude of the first order signal differential is relatively large compared to the rest of the signal. The extremes of this differential can therefore be located and then passed through a threshold filter to get estimates of the beat locations. Since the signal-to-noise ratio is relatively low, the detection is relatively easy to perform and is also quite accurate.

Other common ways of determining QRS location are methods using digital filters, wavelet transform, filter-bank methods and neural network approaches [1]. Digital filters work on a bandpass filtered signal. Nonlinear transformation is performed on the filtered signal which leads to the relative suppression of the low amplitudes and slight smoothing of the peaks. The output of this transformation is then thresholded to extract the beats.

Most wavelet-based QRS detection methods derive from the proposed algorithm by Mallat and Hwang [2] where the R wave peaks are found by scanning for a simultaneous modulus maxima (singularities) in the relevant scales of the wavelet transform. Wavelet-based methods are also one of the most precise in detecting beats even in arrhythmia [3].

With the filtered bank methods the N-band filter bank is used to generate the down sampled sub-band signals. It is assumed that the QRS complex is characterized by a simultaneous occurrence of the ECG frequency components in the sub-bands, which leads to derivation of frequency components into the features which are then used for the actual QRS detection.

Most common neural network approaches for ECG signal processing consist of the multilayer perception (MLP), radial basis function (RBF) networks and learing vector quantization (LVQ).

III. Methodology

A. Differential ECG measurements

Measurements are recorded using the small gadget which is placed on the subject's torso [4]. This gadget is designed to be very energy efficient, thus the measurements are recorded with low sample rate. In the current stage of the development, the sole purpose of this device is to record the ECG signal and transmit it over the Bluetooth Smart wireless protocol to a mobile device running Android.

The mobile device provides a user interface towards the wearable gadget, management of the measurements, basic visualization of the raw ECG signal, storage of the mea-

surements, and the option of synchronising all data to a computer server for further analysis and management. Currently the gadget offers very little in terms of ECG processing, and this is where the newly developed algorithms will come into play. On the mobile device, the analysis should be performed in real time, which makes it absolutely crucial for the proposed heartbeat detection algorithm to perform very efficiently.

After the measurement is finished, the user or an authorized medical specialist can analyze it with special-purpose software. Locally developed software for general ECG analysis is currently used for this purpose. Eventually, specialized software for the analysis of noisy single channel ECG will replace it. ECG analysis is mainly targeted for arrhythmia and various palpitations, while one of its first usages included the detection of sleep apnea [5].

B. Proposed beat detection algorithm

The proposed heartbeat detection algorithm is in many aspects similar to the known detection algorithms, but it is specifically designed to work well with low sample rate and low signal-to-noise ratio. In addition, the algorithm is not affected by the orientation of the ECG features and is able to detect beats even when the electrode orientation is inverted, producing negative R waves in the ECG.

The main principle of the algorithm is to detect QRS complex by using as little processing power and memory as possible. The QRS complex is chosen because it is the most prominent feature of almost any ECG and is also less sensitive to morphology changes than the other ECG waves are [6]. The main feature of the QRS complex, which appears regardless of the gadget placement and orientation, is at least one large spike. Spike is a rapid change in the signal amplitude – either up or down, followed by a rapid change in the reverse direction. Therefore, the algorithm is designed for locating spikes in the signal and deciding weather each spike indicates a new QRS, is a part of the already discovered QRS, or is just noise.

The algorithm comprises four steps: filtering, extrema extraction, spike likelihood estimation, and peak detection.

1) Filtering

In the first step, the signal is filtered with a low-pass filter. This filter should eliminate as much of the high frequency noise from the signal as possible. High frequency noise in the measurement is a result of many factors, such as loose contact between the electrodes and skin, muscle noise, power line and other device interference, and general subject movement. Because the ECG signal resides in the low frequency spectrum, everything above about 45 Hz threshold can be safely filtered out. The band-pass filter could be used for even better effect, with the lower cutoff frequency set to 0.05 Hz without disrupting the ECG at all [7].

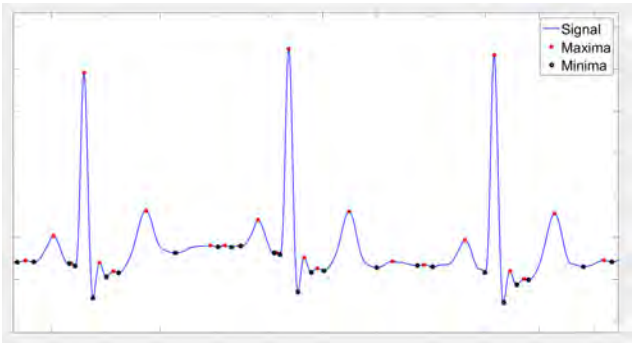


Figure 1. Extrema (red and black markers) of a sample signal (blue line).

Since the filtered signal is not used for anything else than beat detection, even narrower band could be used, that would fit the QRS frequency band better, e.g. 10-40 Hz. For the sake of simpler and more efficient implementation, however, low-pass filter is used instead of band-pass. The most efficient low-pass filter is used, the exponential smoothing, also called the exponential moving average (EMA) with $O(1)$ time and space complexities. The absolute value of the input signal is not used much in the following steps which means that the unfiltered low frequency noise should not pose much problems.

2) Extrema extraction

The extrema of the input signal are extracted to lower the amount of data that has to be processed in the following steps and thus to simplify further processing. Extrema are the samples in which a change occurs in the signal direction, or, more formally, in which the signal derivative changes sign. The derivative is calculated using the first order numerical derivation: $dx_n = x_{n+1} - x_n$.

Local extrema appear in an oscillating fashion, with each minimum followed by a maximum and vice-versa. Figure 1 shows the extrema on a short signal. From the monotonic interval between each pair of consecutive signal extrema one extreme in the derivative is extracted. Based on the derivative sign, the extracted extreme is either minimum or maximum. For example, from the interval between a minimum and maximum in the signal, a maximum in derivative is extracted, since the derivative on that interval is positive. The output of extrema extraction is a stream of signal extrema interleaved with a stream of derivative extrema. The algorithm for extrema extraction as defined above can be written to handle each sample in $O(1)$ time and space complexities.

3) Spike likelihood estimation

Likelihood estimation takes extrema as input. For each signal extreme, the likelihood is calculated as: $a \cdot b \cdot c \cdot d$, where a is the difference between the given and the previous extreme, b is the difference between the given and the next extreme, c and d are the derivative extremes to the left and to the right of the given extreme (see Figure 2).

Again, the time and space complexity of this step is $O(1)$. Furthermore, the estimation function works on the stream of extrema, which is usually sparser than the input signal. Figure 3 shows an example of the likelihood function of an ECG signal. By observing the graph, one can immediately notice that peaks in likelihood correspond very nicely to the QRS complexes of the input signal.

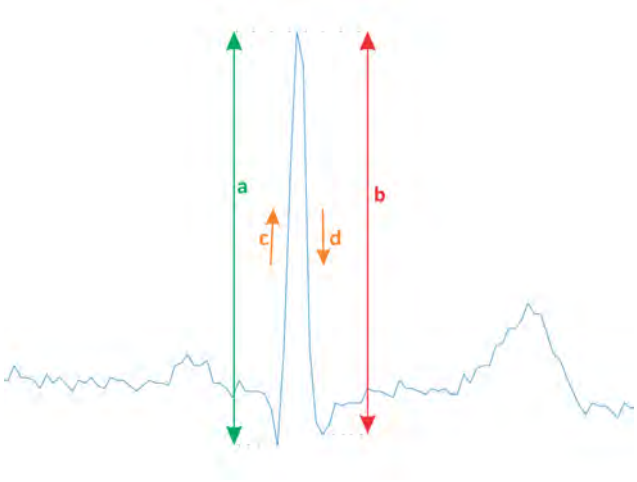


Figure 2. Likelihood estimation scheme. For every triplet of extrema from the input signal (plotted in blue), the following values are taken: the difference between second and first extreme a, the difference between the third and second extreme b, maximal differential of the signal c where the signal is increasing and minimal differential of the signal d where the signal is decreasing.

Likelihood is a relative measure, though, which means that its absolute value is not meaningful. Therefore, to extract peaks, an algorithm must observe any particular likelihood value relative to other likelihood values.

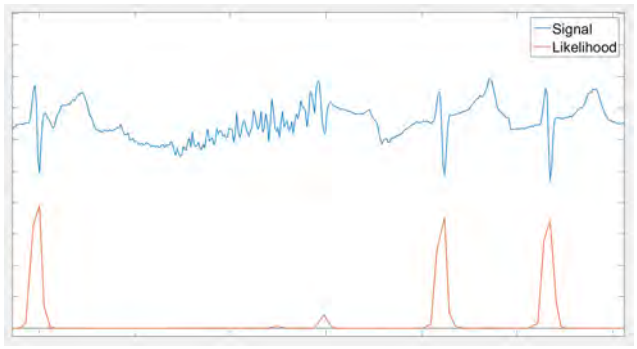


Figure 3. Input signal (top) and the likelihood of spikes in that signal (bottom). Axes are not marked since absolute values are irrelevant.

4) Peak detection

The peak detection starts by extracting extrema from the likelihood in a similar fashion as the extrema were extracted from the input signal, but without extracting the extrema of the derivative this time. Then, the actual peak detection, which is based on Brown's simple exponential smoothing [8], is applied on the likelihood extrema.

Brown's exponential smoothing calculates the smoothed value $\mathbf{s} = [s_i]$ of a sampled signal $\mathbf{x} = (x_i)$ as $s_1 = x_1$ and $s_n = \alpha \cdot x_n + (1-\alpha) \cdot s_{n-1}$ for $n \geq 2$. α is a smoothing factor, which can be chosen from the interval $[0, 1]$. This is the basic low-pass filter, which for example, is used in the first step of the beat-detection algorithm. For smoothing of irregularly sampled signals, i.e. signals for which the time step $t_n - t_{n-1}$ is not equal for $\forall n$, such as is the case of input to the peak detection, a constant α cannot be used. Instead, α is derived from the time constant τ for each sample. Time constant defines the time in which the smoothed response to the unit step input would reach value $1 - 1/e$. The relationship between τ and α can be written as: $\alpha = 1 - e^{\frac{t_{n-1}-t_n}{\tau}}$.

To extend this smoothing into peak detection, two additional parameters are defined. Parameter f defines the *relative peak factor* and a defines the *absolute minimal value* of the maximum to be even considered as a peak. Then, the smoothed response is no longer used as an output but rather as an internal variable, and a new output variable $\mathbf{p} = p(t)$ is defined. The formula for peak detection is then written as:

$$\begin{aligned} s(t_1) &= x(t_1) \\ \alpha &= 1 - e^{\frac{t_{n-1}-t_n}{\tau}} \\ p(t_n) &= \big(x(t_n) > a\big) \text{ and } \big(x(t_n) f > s(t_{n-1})\big) \qquad (1) \\ s(t_n) &= \begin{cases} \alpha x(t_n) + (1-\alpha) s(t_{n-1}), & \text{if not } p(t_n) \\ x(t_n), & \text{if } p(t_n). \end{cases} \end{aligned}$$

The peak detection output $p(t-n)$ is defined as a logical *and* operation over the two threshold checks, one of which uses an absolute value while the other depends on the smoothed response. The smoothed response differs from the one obtained by basic exponential smoothing only when large peaks appear on input. For each peak in the input $x(t)$, which satisfies the $p(t)$ condition, the smoothed response $s(t)$ is set to the current input value. Otherwise the smoothed response is calculated using the regular exponential smoothing. The output $p(t_n)$ is true if the input value $x(t_n)$ is higher than a and at least f times higher than the previous value of the smoothed response $s(t_{n-1})$. This type of peak detection is fully adaptable to the input signal and only detects the highest peaks on a time scale that is controlled with parameter τ.

The time and space complexities of peak detection are composed of the extreme extraction complexity and the modified smoothing complexity. Both are again $O(1)$, as is the total peak detection complexity. The output of peak detection is also the result of the beat detection algorithm and corresponds to the locations of the QRS complexes in the input signal.

C. Signal to noise ratio estimation

Although in the proposed beat detection algorithm, the input signal is first passed through a frequency filter to eliminate noise, the noise in the QRS frequency band cannot be filtered without corrupting the signal and has to be accounted with. Since the QRS complex can sometimes be falsely interpreted as noise or vice-versa – the noisy segments can be interpreted as QRS, a noise estimator is used to establish a trust factor in the individual beat detection. The requirements for this estimator are the same as the requirements for the whole algorithm – to be fast and resource efficient.

First, it should be noted that very low frequency noise, such as the baseline wander, does not influence peak detection much, since peak detection works with relative values and derivatives only. Therefore, the low frequency noise estimate is not required to establish the trust factor. Two noise estimators are proposed, which both estimate mostly high frequency noise.

1) Median filter

The first noise estimator estimates the signal-to-noise ratio of every sample by filtering the derivative of the unfiltered input signal with median filter. This detector estimates the signal to noise ratio throughout the whole signal and can be used prior to the heartbeat detection. Figure 4 shows the response of the median filter noise estimator to a noisy input signal. While the median window size M should equal approximately the length of an average heartbeat, the actual size may vary, since the beat rate is not necessarily known prior to filtering.

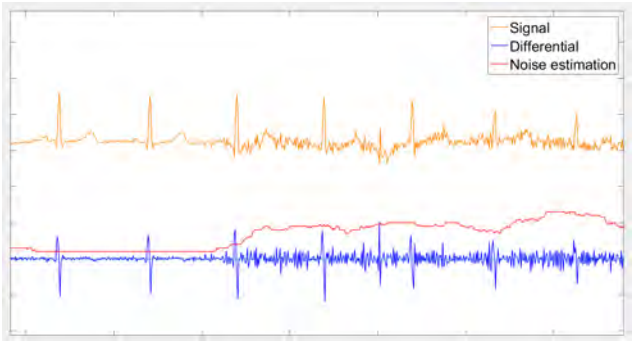


Figure 4. Unfiltered input signal to the median noise estimator (top), its derivative (bottom blue) and the estimator response (bottom red).

Median calculation with the selection algorithm [1] has time and space complexity $O(M)$. This can be optimised by implementing the filter as a moving median window with samples stored in two binary heaps, where moving the window by a single sample results in $O(log(M))$ time complexity and retrieval of the median in $O(1)$ time complexity. Space complexity does not change with this implementation.

2) Frequency of similar extremes estimator

Second noise estimator requires the beats to be detected before the actual noise estimation, and it only calculates the noise level at each beat. The largest first derivative amplitude of the located beat is used to define the frequency of nearby extremes with similar amplitude. Window of this estimator may be predefined but the implementation works better if the window interval is dynamically adjusted. A dynamical window adjustment may use heart rate to determine the window size in samples. For example, if H represents the estimated heart rate, and f represents the sampling frequency, the window size W is defined as $W = \frac{H}{60} * f$.

The amplitude of the first derivative extremes in this predefined window is used to derive the number of events that could be picked up by the peak detection algorithm. The high frequency estimator works as follows. The extrema of the derivative are used as an input. Frequency is then determined by counting all the extrema with amplitude higher than $T = \beta * B$, where B is the amplitude of the detected beat extrema and β is the threshold parameter. The result of noise estimation using this estimator is shown on Figure 5.

According to first tests, the best value for β lies in the interval $[0.25, 0.67]$ and depends on the amount noise expected in the signal. For this estimator to work reliably, the highest absolute derivative found within the QRS complex should be taken as B even if the beat detector marks the beat elsewhere. The output of the estimator is the number of extrema that have higher amplitude than the threshold value T. This estimator has $O(W)$ time and space complexity, where window size W is expressed in number of samples.

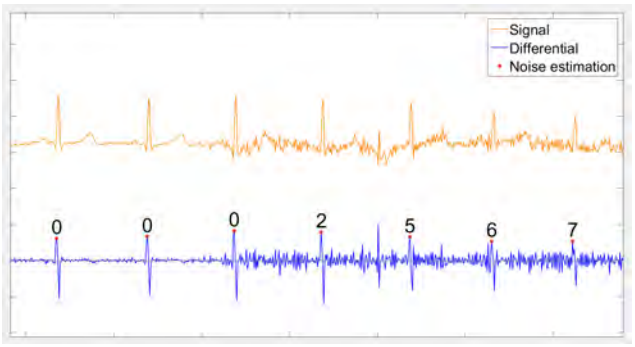


Figure 5. Input signal to the frequency of similar extremes noise estimator (top), and its derivative (bottom) with the noise estimator response in labels.

IV. Results and Discussion

A. Noise estimators

The noise estimators were tested using an artificial measurement. A noise-free recording of a single heart beat, sampled at 1800 Hz on a regular ECG device was taken. Measurement was then created by chaining 200 beats one after the other, then low-pass filtered and re-sampled down to 125 Hz. Finally, varying level of Gaussian white noise was added to the signal.

The tested signal to noise ratio SNR ranged from -10 to 25 dB. The noise was added to the signal based on the signal power SP, where

$$SP = 10 * log_{10}(\sum_{i=1}^{n} x^2_i).$$

Based on the derived NP, the additive $noise$ for each sample is defined as:

$$\begin{aligned} NP &= SP - SNR \\ noise_i &= \sqrt{10^{\frac{NP}{10}} * randn}, \end{aligned}$$

where $randn$ represents the random generated number based on the Gaussian distribution. Beat detection was applied on the provided signal, to get the relative frequency of errors in beat detection – a sum of false positives and false negatives normalized by the total number of beats. The relative frequency of errors should help determine weather the noise estimators can reliably discern the noise-free segments of measurements with low possibility for miss-detection of beats from noisy segments with higher possibility for miss-detection of beats.

1) Median filter

The response of the median filter is observed per beat, with mean and standard deviation calculated from the 200 beats of the artificial measurement. Figure 6 shows the response of the estimator with a window size of $0.8\ s$ as a function of input noise. Overall, the estimator response grows linearly with decreasing SNR, diverging from the linearity at about 10 dB SNR and more. Relative frequency of errors in the beat detector is also plotted on the same figure.

It shows that using the median filter makes it hard to differentiate between the noise levels where a large number of errors is likely to occur from those where no errors is likely to occur in the beat detection.

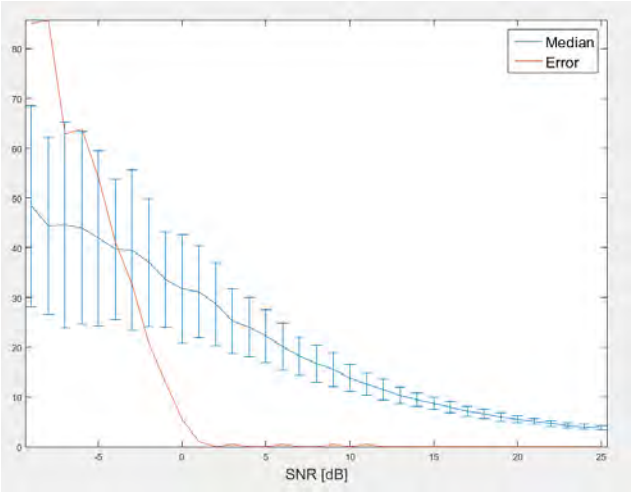


Figure 6. Median filter noise estimator response and relative frequency of errors made by the beat detector as functions of SNR. Mean estimator response ± one standard deviation are plotted in blue. Relative frequency of errors (in percent) is plotted in red.

2) Frequency of similar extremes

The response of the estimator based on the frequency of events similar to beats is also observed per beat with mean and standard deviation calculated from the 200 beats of the artificial measurement. Figure 7 shows the response of the estimator with a window size of $0.8\ s$ and threshold $\beta = 0.25$, as a function of input noise. This estimator's response, in contrast to the previous one's, is much less linear, but varies much more between the areas where there are possible errors in beat detection and areas where there are not.

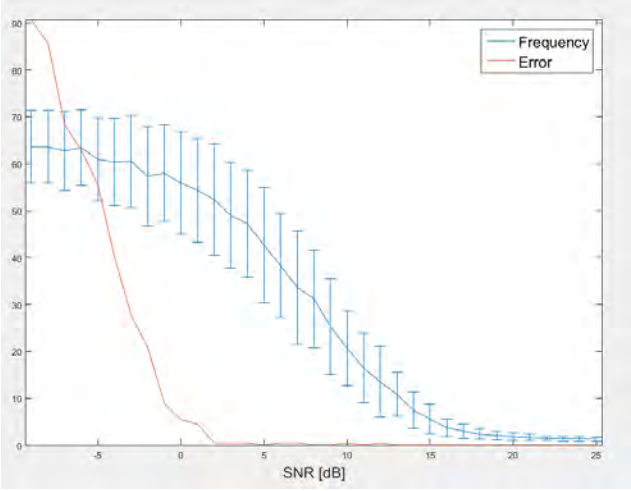


Figure 7. Frequency of similar peaks response and relative frequency of errors made by the beat detector as functions of SNR. Mean estimator response ± one standard deviation are plotted in blue. Relative frequency of errors (in percent) is plotted in red.

B. Beat detector

The proposed algorithm was tested on the ECG samples from the PhysioNet's MIT–BIH Arrhythmia database [9], using MLII channel of each measurement.

TABLE 1. STATISTICS OF THE BEAT DETECTOR ON THE MIT–BIH ARRHYTHMIA DATABASE

File	tp	fp	fn	Precision	Recall
100	2271	0	2	100.0%	99.9%
101	1862	6	3	99.7%	99.8%
102	2184	0	3	100.0%	99.9%
103	2082	0	2	100.0%	99.9%
104	2198	68	31	97.0%	98.6%
105	2552	69	20	97.4%	99.2%
106	2014	8	13	99.6%	99.4%
107	2127	4	10	99.8%	99.5%
108	1605	30	158	98.2%	91.0%
109	2527	1	5	100.0%	99.8%
111	2104	1	20	100.0%	99.1%
112	2537	1	2	100.0%	99.9%
113	1793	0	2	100.0%	99.9%
114	1874	5	5	99.7%	99.7%
115	1951	0	2	100.0%	99.9%
116	2391	2	21	99.9%	99.1%
117	1533	2	2	99.9%	99.9%
118	2255	24	23	98.9%	99.0%
119	1985	48	2	97.6%	99.9%
121	1856	1	7	99.9%	99.6%
122	2474	1	2	100.0%	99.9%
123	1516	0	2	100.0%	99.9%
124	1617	3	2	99.8%	99.9%
200	2587	76	14	97.1%	99.5%
201	1884	1	79	99.9%	96.0%
202	2128	4	8	99.8%	99.6%
203	2860	140	120	95.3%	96.0%
205	2641	0	15	100.0%	99.4%
207	1898	39	434	98.0%	81.4%
208	2915	18	40	99.4%	98.6%
209	3002	8	3	99.7%	99.9%
210	2607	16	43	99.4%	98.4%
212	2746	6	2	99.8%	99.9%
213	3248	10	3	99.7%	99.9%
214	2253	6	9	99.7%	99.6%
215	3356	4	7	99.9%	99.8%
217	2205	2	3	99.9%	99.9%
219	2149	29	5	98.7%	99.8%
220	2045	0	3	100.0%	99.9%
221	2421	4	6	99.8%	99.8%
222	2476	1	7	100.0%	99.7%
223	2601	2	4	99.9%	99.8%
228	2036	44	17	97.9%	99.2%
230	2254	3	2	99.9%	99.9%
231	1569	0	2	100.0%	99.9%
232	1773	3	7	99.8%	99.6%
233	3073	2	6	99.9%	99.8%
234	2751	1	2	100.0%	99.9%
$\sum$	108786	693	1180	99.4%	98.9%

The detected heart beats were compared to the heart beats annotations from the MIT–BIH database hosted by Physionet. The following parameters were used: input low-pass filter time constant $\tau_{input} = 0.004$, peak detection in likelihood time constant $\tau = 0.05$, relative threshold $f = 8$

and absolute threshold $a = 0.0013$. The results are presented in Table 1 as true positive values (tp: the number of correctly detected beats), false positives (fp: the number of events misidentified as beats) and false negatives (fn: number of beats that were not detected). Precision and recall, defined as $tp/(tp+fp)$ and $tp/(tp+fn)$, respectively, are also calculated and presented. Note that precision and recall are above 95% for all given measurements, but the measurement numbered 207, where recall is only 81%, this measurement includes ventricular fluttering, which the presented algorithm is unable to detect. The results summed over the whole database seem quite good with 99.4% precision and 98.9% recall, although the state-of-the-art QRS detection algorithms achieve even better results (see overview in [10]).

V. Conclusion

The differential ECG measurements are being recorded that require a novel way of automatic data processing. In this work, algorithms for coarse beat detection and noise estimation are proposed. All of the algorithms have very low time and space complexities and can therefore be implemented in an efficient manner, even on very basic hardware, such as the wearable ECG monitoring gadgets for example.

Beat detection was tested on MIT–BIH arrhythmia database to great success. Since it was designed for noisier measurements than the ones available in MIT–BIH database, its true value will be demonstrated in the future, once the database of such measurements becomes available. Furthermore, its efficiency was only estimated so far and should be tested by implementing it to run on hardware with limited capabilities.

The two noise estimators were tested on a signal with artificial noise added, and also produced encouraging results. The variability of the results was not as low as desired, and noise power was not perfectly estimated, but at least an indication of noise was achieved. The estimators should be further tested with implementation on hardware with limited capabilities to verify their efficiency.

A very important task was left for future work – the integration of the beat detector and the noise estimator. The noise estimation could be fed back into the beat detection, or the detected beats could be subsequently modified, based on the noise estimates. For now, though, the detected beats can be annotated with the noise estimate, which is still much better than having only the beat times available.

Acknowledgment

The authors acknowledge the financial support from the state budget by the Slovenian Research Agency under Grant P2-0095

References

[1] B. U. Köhler, C. Hennig and R. Orglmeister. The principles of software QRS detection. IEEE Engineering in Medicine and Biology Magazine, 2002, 21(1), pp. 42-57.

[2] S. Mallat and W. L. Hwang, Singularity detection and processing with wavelets, IEEE Trans. Inform. Theory, 1992, vol. 38(2), pp. 617-643.

[3] J. P. Martínez, R. Almeida, S. Olmos, A. P. Rocha and P. Laguna, . A wavelet-based ECG delineator: evaluation on standard databases. Biomedical Engineering, IEEE Transactions on, 2004, 51(4), pp. 570-581.

[4] A. Rashkovska, I. Tomašić, R. Trobec. A telemedicine application : ECG data from wireless body sensors on a smartphone. In proceedings of: MIPRO 2011, 34th International Convention, May 23 - 27, 2011, Opatija, Croatia. P. Biljanović, K. Skala (ed.), pp. 293-296.

[5] V. Avbelj. Extranodal pacemaker activity during sleep : a case study of wireless ECG sensor data. Proceedings of MIPRO 2015, 38th International Convention, May 25-29, 2015, Opatija, Croatia. P. Biljanović (ed.), pp. 389-391.

[6] V. Avbelj, R. Trobec, B. Geršak. Beat-to-beat repolarisation variability in body surface electrocardiograms. Medical & biological engineering & computing, 2003, 41(5), pp. 556-560.

[7] F. Buendía-Fuentes, M. A. Arnau-Vives, A. Arnau-Vives, et al., High-Bandpass Filters in Electrocardiography: Source of Error in the Interpretation of the ST Segment, ISRN Cardiology, 2012, 2012(1), 10 pages.

[8] Brown, R. G. Exponential smoothing for predicting demand. In Operations Research, 1957, 5(1), pp. 145-145.

[9] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals, Circulation 101(23), 2000, pp. 215–220. Database and tools available at: http://www.physionet.org/.

[10] E. J. D. S. Luz, W. R. Schwartz, G. Cámara-Chávez, and D. Menotti, ECG-based heartbeat classification for arrhythmia detection: A survey. Computer Methods and Programs in Biomedicine, 2015.

Classification of Asthma Using Artificial Neural Network

A. Badnjević[1,2,3,4], L. Gurbeta [1,2], M. Cifrek[5], D. Marjanovic[1]
[1] International Burch University, Sarajevo, Bosnia and Herzegovina
[2] Verlab Ltd Sarajevo, Bosnia and Herzegovina
[3] Faculty of Electrical Engineering, University of Sarajevo, Bosnia and Herzegovina
[4] Faculty of Medicine, University of Sarajevo, Bosnia and Herzegovina
[5] University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia

Abstract - **This paper presents a system for classification of asthma based on artificial neural network. A total of 1800 Medical Reports were used for neural network training.**
The system was subsequently tested through the use of 1250 Medical Reports established by physicians from hospital Sarajevo. Out of the aforementioned Medical Reports, 728 were diagnoses of asthma, while 522 were healthy subjects. Out of the 728 asthmatics, 97.11% were correctly classified, and the healthy subjects were classified with an accuracy of 98.85%. Sensitivity and specificity were assessed, as well, which were 97.11% and 98.85%, respectively.
Our system for classification of asthma is based on a combination of spirometry (SPIR) and Impulse Oscillometry System (IOS) test results, whose measurement results were inputs to artificial neural network. Artificial neural network is implemented to obtain both static and dynamic assessment of the patient's respiratory system.

I. INTORODUCTION

For some people, asthma is a minor health issue, but for others, it can be a major problem that interferes with daily activities and may lead to a life-threatening asthma attacks. Asthma cannot be cured, but early diagnosis is crucial in controlling this disease. Asthma is a paroxysmal, often allergic disorder of respiration, characterized by bronchospasm, wheezing, and difficulty in expiration, often accompanied by coughing and a feeling of constriction in the chest [1]. WHO recognizes that asthma is of major public health importance and works on developing different strategies in asthma management [2]. According to WHO estimates, 235 million people suffer from asthma and it is creating a substantial burden to individuals and families and possibly restricting individuals' activities for a lifetime [2]. One of greatest obstacles in asthma management is proper diagnosis of patients, so number of patients with asthma are improperly diagnosed or misdiagnosed for having other respiratory diseases [3,4]. Early confirmation or exclusion of the diagnosis of asthma may avoid needless trails or therapy and other medical tests. Beside physical exam and a medical and family history, medical professionals, most commonly, use spirometry (SPIR) and Impulse Oscillometry (IOS) test results [5,6] for static assessment of patients. Dynamic assessment of the patient is necessary process, as well, according to Global Initiative for Asthma (GINA) guidelines [7]. In order to get dynamic assessment of the patient, allergies and patient's symptoms must be taken into account and bronchial dilation (BDT) and bronchial provocation tests (BPT) also must be conducted. The prevalence of asthma has been increasing in the recent decades, as well as the costs for asthma care [8-11]. Significant efforts are made to develop systems which will help medical professionals in correct diagnosis of patients with asthma. According to that, researchers have proposed different solutions in disease detection and classification, making a breakthrough in using the informational technologies in classification of disease.

In [12], the authors used the Multi-Layer Neural Network (MLNN) with one hidden layer and they used BP training algorithm for diagnosis of Chronic Obstructive Pulmonary Disease (COPD). In [13], the authors classified respiratory abnormalities using adaptive neuro fuzzy inference system based only on SPIR or IOS measurements. In the other study [14], MLNNs with one and two hidden layers is used and BP momentum as the training algorithm for diagnosis of COPD disease. The authors utilized a 10-fold cross-validation procedure to train the neural networks. In [15], the authors performed the task of separation of lung sounds, obtained from patients with pulmonary pathology using recurrent neuro-fuzzy filter. Those studies focused on static assessment of the patient using combination of spirometry and/or IOS measuring test results.

This paper presents the results of study developing Artificial Neural Network (ANN) for classification of asthma based on static and dynamic assessment of patient. Data which have to be entered are derived from GINA guidelines. Measuring parameters of IOS, spirometry, information on symptoms, allergies and auscultation results, are included in the ANN. This information allows the ANN to suggest proper classification of asthma or normal lung condition. In cases where it is not possible to determine reliable results, the ANN uses BDT and/or BPT in order to complete the assessment.

II. METHODS

Spirometry (SPIR) and IOS measuring test results are input vector of the neural network. In cases where it is not possible to determine reliable results, the ANN uses BDT and/or BPT in order to complete the assessment. Allergies

and patient's symptoms must be taken into account in disease classification according to the GINA guidelines, so this is also one of input parameters to implemented the ANN. A block diagram showing data flow is presented in Figure 1.

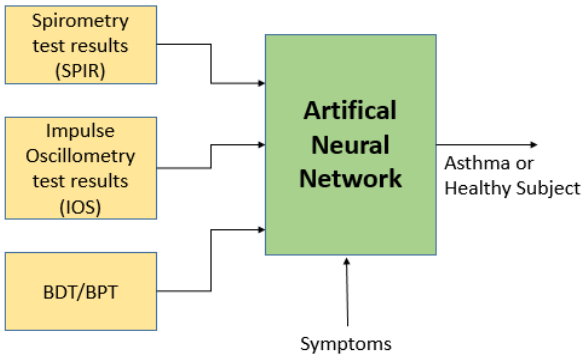


Figure 1. Architecture for disease classification

Feedforward network with two layers has been used. According to the application experts, the linear feed forward (FF) neural network are mostly used for classification task [16-20]. The implemented network is divided into layers. The architecture of implemented neural network is shown in Figure 2.

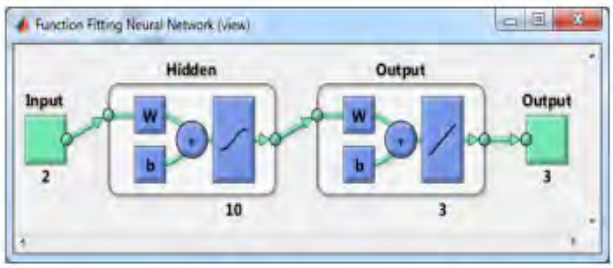


Figure 2. Architecture of implemented neural network

The input layer is made of network inputs. Input layer is followed by a hidden layer which consists of 10 neurons connected parallel. Each neuron performs a weighted summation of the inputs, which is then passed to some non-linear activation function. Since, this classification is basically highly non-linear data classification, we choose tansig activation function for hidden layer which is equivalent to the hyperbolic tangent function. This function is used in highly non-linear data classification [21]. Weighted summation of the outputs of the neurons in the hidden layer makes the output of neural network. Possible outputs are healthy, asthma or BDT/BPT. Linear activation function, commonly used in regression problems [22], is used for the output layer. Implemented neural network classifies input data into one of three classes: asthma, healthy subject or inconclusive where additional BDT or BPT inputs are needed for proper classification. The ANN is trained using estimation data and validated using validation data. The training data set is divided using 80-20 division to get estimation and validation dat. The training is performed with 1800 Medical Reports. 1440 Medical Reports were used for estimation and 360 for validation. The training algorithm we used is Levenberg-Marquardt (LMA), which is common training algorithm in data classification [23]. Estimation and validation data for NN are shown in Table I.

TABLE I. ESTIMATION AND VALIDATION DATA FOR NEURAL NETWORK WITH 10 NEURONS IN HIDDEN LAYER

Number of reports	1800			
Estimation set size	1440			
Validation set size	360			
ESTIMATION SET				
	No. of reports	True diagnosis	False diagnosis	% of true diagnosis
Asthma	1228	1219	9	99.26%
Healthy	212	210	2	99.68%
			Average:	99.47%
VALIDATION SET				
	No. of reports	True diagnosis	False diagnosis	% of true diagnosis
Asthma	298	297	2	98.66%
Healthy	62	62	0	100%
			Average:	99.33%

Also, the effect of change in number of neurons in hidden layer was evaluated during the validation process of the neural network.

III. RESULTS

The accuracy of implemented neural network with different numbers of neurons are presented in Figure 3. The ANN was tested with 1, 2, 5, 10 and 20 neurons.

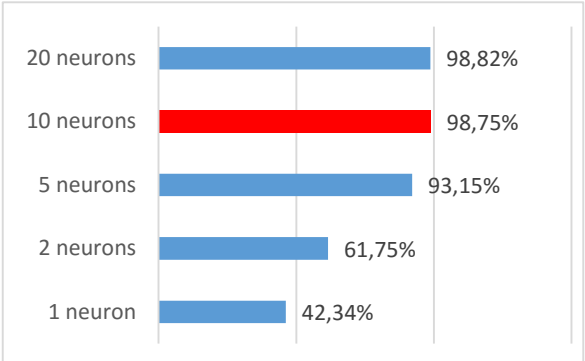


Figure 3. Effect of change in number of neurons in hidden layer during validation process

During the validation process the best result were achieved with 20 neurons in the hidden layer. Even though, the results with 10 neurons are negligibly with lower accuracy, they were chosen in order to lower the computational complexity of the network. The system was subsequently tested through the use of totally different 1250 Medical Reports established by physicians from hospital Sarajevo. The same medical reports were used in [31].

TABLE II. PERFORMANCE OF THE IMPLEMENTED NEURAL NETWORK

Total population	Σ 1250	COPD	Healthy subjects	**Average Accuracy**
Asthma	Σ 728	707	21	97.11 %
Healthy subjects	Σ 522	516	6	98,85 %

Out of the aforementioned Medical Reports, 728 were diagnoses of asthma, while 522 were healthy subjects. Out of the 728 asthmatics, 97.11% were correctly classified, and the healthy subjects were classified with an accuracy of 98.85%. The performance of the implemented neural network is shown in Table II. Sensitivity and specificity were assessed, as well, which were 97.11% and 98.85%, respectively.

Our neural network for classification of asthma is based on a combination of spirometry (SPIR) and Impulse Oscillometry System (IOS) test results. Artificial neural network is implemented to classify input data based on static and dynamic assessment of the patient's respiratory system. For implementation of Artificial Neural Network Matlab software package was used.

For more accurate results, implemented neural network can be combined with fuzzy rules implemented in accordance with international guidance to implement neuro-fuzzy system for asthma classification [31].

IV. CONCLUSION

In this study we have presented Artificial Neural Network for classification of asthma. The network is implemented in Matlab. Input parameters to Artificial Neural Network are based upon Global Initiative for Asthma (GINA) guidelines.

A total of 1800 Medical Reports were used for neural network training and subsequent testing was done using totally new 1250 Medical Reports established by physicians from hospital Sarajevo. Sensitivity and specificity were assessed, as well, which were 97.11% and 98.85%, respectively.

The novelty in this implementation of the ANN for asthma classification and previous work of the researchers [25-30] is that this solution is implemented to make classification directly from the measured data of the patient.

LITERATURE

[1] Collins English Dictionary - Complete & Unabridged 10th Edition. Retrieved February 20, 2016 from Dictionary.com website: http://dictionary.reference.com/browse/asthma

[2] Worl Helath Organisation, available at: http://www.who.int/respiratory/asthma/activities/en/

[3] Helms, P. J. Issues and unmet needs in pediatric asthma. Pediatr Pulmonol 2005; 30:159-65.

[4] Enright, P. L., McClelland, R. L., Newman, A. B., Gottlieb, D.J., Lebowitz, M. D. Under diagnosis and under treatment of asthma in the elderly. Cardiovascular health study research group. Chest 1999;116:603-13.

[5] Song, T.W. Correlation between spirometry and impulse oscillometry in children with asthma. Acta Pediatr. 2008., Vol. 97, pp 51-54.

[6] Larsen, G. L. Impulse oscillometry versus spirometry in a long-termstudy of controller therapy for pediatric asthma. J Allergy Clin Immunol, 2009; Vol. 123, pp 861-867.

[7] Global Initiative for Asthma: Pocket guide to Asthma Diagnosis, Management and Prevention, available at: http://www.ginasthma.org/local/uploads/files/GINA_Report_2014_Aug12.pdf , Accessed 05 March 2015

[8] Tortora. G.J., Grabowski, S.R. Principles of Anatomy and Physiology. 10th ed. New York, USA: John Wiley & Sons, 2003.

[9] Mayo Foundation for Medical Education and Research, Mayo Clinic: COPD Definition. Available: http://www.mayoclinic.com/health/copd/DS00916

[10] Anthonisen, N. Chronic obstructive lung disease in Cecil Medicine, 24th ed. L. Goldman, D. Auseillo, Ed. Philadelphia, USA, 2007, ch. 88.

[11] European Lung Function: COPD Burden in Europe. Available http://www.european-lung-foundation.org/index.php?id=63

[12] Ashizawa, K., Ishida, T., MacMahon, H., Vyborny, C.J., Katsurgawa, S., Doi, K. Artificial neural networks in chest radiography. Application to the differential diagnosis of interstitial lung disease, Academic radiology, 11, 2005, pp 29-37

[13] Asaithambi, M., Manoharan, S.C., Subramanian, S. Classification of Respiratory Abnormalities Using Adaptive Neuro Fuzzy Inference System. Int. Inf. and Database SysLecture Notes in Computer Science, 202, vol. 7198, pp 65-73.

[14] Coppini, G., Miniati, M., Paterni, M., Monti, S., Ferdeghini, E.M. Computer aided diagnosis of emphysema in COPD patients: Neural network based analysis of lung shape in digital chest radiographs. Medical Engineering and Physics, 2007, pp 76-86

[15] Stavrakoudis, D., Mastorocostas, P., Theocaris, J.A. Pipelined Recurrent Fuzzy Neural Filter for the separation of Lung Sounds. IEEE International Fuzzy Systems Conference, pp 1-6, 2007.

[16] Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Meltzer, P.S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine 2001, 7(6):673-679.

[17] Cevenini, G., Severi, F. M., Bocchi, C., Petraglia, F., Barbini, P. An informative probability model enhancing real time echobiometry to improve fetal weight estimation accuracy. Med Biol Eng Comput 2008, 46(2):109-120.

[18] Güler, N.F., Übeyli, E.D., Güler, I. Recurrent neural networks employingLyapunov exponents for EEG signals classification. Expert Systems with Applications 2005, 29(3):506-514.

[19] Byvatov, E., Fechner, U., Sadowski, J., Schneider, G. Comparison of supportbvector machine and artificial neural network systems for drug/nondrug classification. Journal of Chem Inf and Comput Sci 2003, 43(6):1882-1889.

[20] Kermani, B.G., Schiffman, S.S., Nagle, H.T. Performance of the LevenbergMarquardt neural network training method in electronic nose applications. Sensors and Actuators B: Chemical 2005, Vol. 110, pp. 13-22.

[21] Azemi, A., Sabzevari, V.R., Khademi, M., Gholizade, H., Kiani, A., Dastgheib, Z.S. Intelligent arrhythmia detection and classification using ICA. In Engineering in Medicine and Biology Society, 28th Annual International Conference of the IEEE 2006, pp 2163-2166.

[22] Wang, G.C.S., Jain, C.L. Regression Analysis Modelling & Forecasting. Graceway Publishing Company Inc; 2003.

[23] Hao, Y., Wilamowski, M. Levenberg-Marquardt Training.

[24] Young. The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[25] Badnjevic, A, Koruga, D, Cifrek, M, Smith, H.J., Bego, T. „Interpretation of pulmonary function test results in relation to asthma classification using integrated software suite“, IEEE 36. International convention on information and communication technology, electronics and microelectronics (MIPRO), pp: 140 - 144, 21.-25. May 2013. Opatija, Croatia.

[26] Badnjevic, A., Cifrek, M., Koruga, D. Classification of Chronic Obstructive Pulmonary Disease (COPD) Using Integrated Software Suite. XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013, Vol. 41, pp 911-914.

[27] Badnjevic, A., Cifrek, M. Classification of asthma utilizing integrated software suite, 6th European Conference of the International Federation for Medical and Biological Engineering (MBEC), 07-11. September 2014., Dubrovnik, Croatia.

[28] Badnjevic, A., Cifrek, M., Koruga, D. Classification of Chronic Obstructive Pulmonary Disease (COPD) using integrated software suite, IFMBE XIII Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON), 25-28. September 2013., Sevilla, Spain.

[29] Badnjevic, A., Cifrek, M., Koruga, D. Integrated software suite for diagnosis of respiratory diseases, IEEE International conference on computer as tool (EUROCON), pp: 564-568, 01-04. July 2013., Zagreb, Croatia

[30] Badnjevic, A., Cifrek, M., Koruga, D., Osmankovic, D. (2015). Neuro-fuzzy classification of asthma and chronic obstructive pulmonary disease. BMC Med Inform Decis Mak BMC Medical Informatics and Decision Making, 15(Suppl 3).

[31] Badnjevic, A., Gurbeta, L., Cifrek, M., Marjanovic, D., Diagnostic of Asthma Using Fuzzy Rules Implemented in Accordance with International Guidelines and Physicians Experience. Submitted to IEEE 39. International convention on information and communication technology, electronics and microelectronics (MIPRO), May 2016. Opatija, Croatia.

Brain-computer Interface Based on Steady-state Visual Evoked Potentials

K. Friganović*, M. Medved* and M. Cifrek*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
kresimir.friganovic@fer.hr

Abstract - Brain computer interface (BCI) can establish communication between human brain and a computer which is independent from normal neuromuscular pathways. This allows giving instructions to the computer without the use of standard communication channels, such as a mouse or keyboard. This paper describes the development of a steady-state visual evoked potential (SSVEP) based BCI system. EEG amplifier with one bipolar channel is designed for the acquisition of raw EEG data from the posterior region of the head over the occipital lobe. The three white LED chessboards with programmable flicker frequencies are used as stimulation to induce different SSVEPs. For feature extraction, the Fourier transform of autocorrelation of the EEG signal is used. In MATLAB, a graphical user interface (GUI) application is implemented, showing the EEG signal in time and frequency domain, both in real time. Simple game of turning on and off the three light bulbs, while looking at different LED chessboard, is also implemented in GUI.

Keywords: brain-computer interface (BCI), steady-state visual evoked potentials (SSVEP), EEG amplifier

I. INTRODUCTION

Brain computer interfaces (BCI) can establish communication between human and computers. This communication is independent from the normal neuromuscular pathways and allows the subject to give instructions to computer by paying a bit of attention or thinking a single thought [1].

Bioelectrical brain activity is measured and used to detect specific patterns in activity. Brain activity can be measured invasive, when electrodes are placed on the surface of the brain (electrocorticography, ECoG). Signals measured with this method have greater amplitude and better signal to noise ratio (SNR) than signals measured from the surface of the scalp (electroencephalography, EEG) do [2]. However, due to the non-invasive procedure of EEG recording, the EEG method is mostly used in the brain-computer interfaces.

The EEG is used to measure and detect specific patterns in the bioelectrical brain activity, called event-related potentials (ERPs). The three most used ERPs in BCI systems are: P300, SSVEP and motor imagery potentials. In this paper SSVEP are used to establish communication.

SSVEPs are manifested as a periodic response of the bioelectrical brain activity over the occipital lobe. They occur when subject is focusing on the continuous flickering visual stimulus. The frequency of SSVEP is equal to the flicker frequency of the stimulus. Fig. 1 shows a diagram of the BCI system based on the SSVEP.

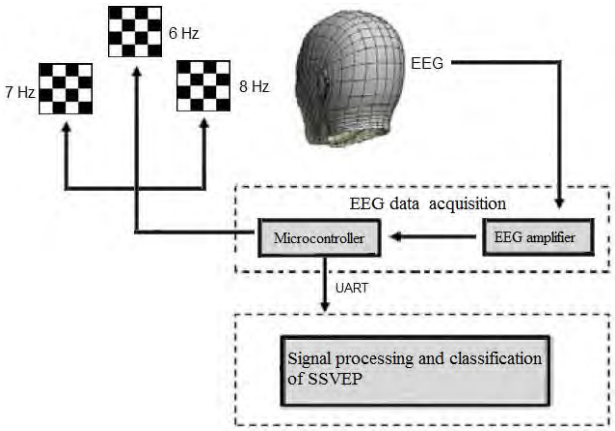


Figure 1. A Block diagram of SSVEP based BCI system

Algorithms for extraction and classification of SSVEPs can be found in related works ([3] - [5]). Design of EEG amplifier, including preamplifier and noise reduction circuitry is represented in the relevant works ([6] - [9]).

In our research, we have developed and designed both hardware and software components of SSVEP BCI system, using only one bipolar EEG channel recording from the posterior region of the head over the occipital lobe (electrodes are placed over position O1 from 10-20 system) and three LED stimuli.

II. HARDWARE

A. EEG Amplifier

Hardware design is an important part of the measurement system because all digital signal processing methods depend on the quality of the analog preprocessing. An interference affecting the EEG signal determines the design of an EEG amplifier. The main interferences are from the power lines (50 Hz common mode interference) and the human body signals that don't belong to the EEG signal spectrum. Additional factors that affect the quality of the EEG signal are the impedance and voltage offsets, created by inadequate preparation of the skin-electrode interface. Key parameters in design of EEG signal measuring system are:

- The input impedance of the preamplifier must be high, so that the interference could be rejected.

- A differential amplifier with the high common mode rejection ratio (CMRR) is required to suppress common mode interference.
- The amplifier must have a low noise with gain higher than 10^4, due to the small EEG signals amplitude.
- The frequency spectrum of a desired EEG signal is lower than 40 Hz, and, therefore, higher frequencies need to be suppressed.

The amplifier consists of the several stages in a cascade. A block diagram of the whole EEG amplifier is shown at Fig. 2. The first stage begins with a RFI filter that suppresses unnecessary high frequency signals from the input. An instrumentation amplifier is a preamplifier with high input impedance, low voltage drift, low current consumption, low noise and high CMRR. A shield driving circuit is used to additionally eliminate common mode interferences that are engaged in the measuring circuit. The third cascade stage is a high pass that is needed to suppress the DC frequency component so that non-inverting amplifier does not amplify unneeded signal component. A low pass filter is realized as the fourth order by using operational amplifier circuit sections to eliminate high frequencies before the final amplification. At the end of the cascade is a programmable gain amplifier (PGA). This amplifier is an advantage, since it allows for the signal to be aligned to operating range of different analog-digital converters (ADC).

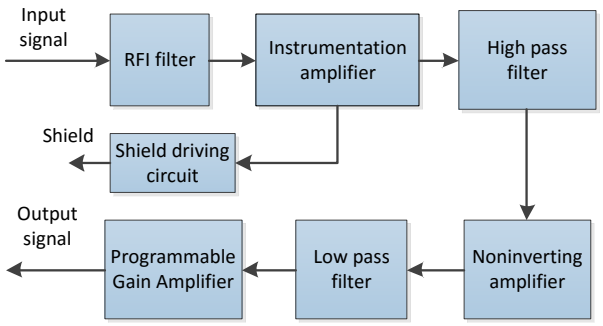


Figure 2. Block diagram of EEG amplifier

The RFI low pass filter is implemented by two resistors and three capacitors as it is shown at Fig. 3. The filter is designed to bypass differential signals lower than 419 Hz, and common signals lower than 39.8 kHz. Most of unwanted high frequencies are eliminated by this filter.

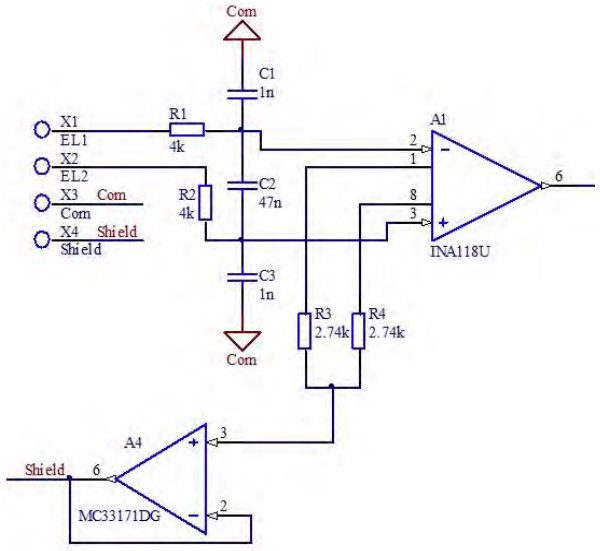


Figure 3. Schematic of EEG amplifier input stage

An instrumentation amplifier INA118U is used as the first stage amplifier. The gain is set to 10 by resistors $R_3 = R_4 = 2.74$ kΩ and it can be calculated by (1).

$$G = 1 + 50\text{k} / (R3 + R4) \quad (1)$$

Main features of this amplifier are high CMRR (110 dB), low offset voltage ±50 µV, 1 nA input bias current, offset drift around ±2.1 µV/°C and input impedance about 3 GΩ [10]. The gain of differential signal shouldn't be too high because offset voltage is also amplified, as well as unwanted, especially in this application.

To suppress the common signal, which is coupled to the amplifier circuit, a shield drive circuit is used. The composition of the circuit is made by the operation amplifier MC33171DG. Therefore, common signal is passed to cable shields and its influence on the measured signal is reduced.

The next cascade section is a high pass filter to eliminate the DC component. The cut-off frequency is set to 0.8 Hz by selecting capacitor 470 nF and resistor 422 kΩ. Consequently, the signal is more stable and accurate so it can be amplified again.

The non-inverting amplifier is implemented using the operation amplifier OPA2277UA which has low power consumption (0.9 mA) and low input voltage offset (±50 µV). Signal may still contain unwanted components at higher frequencies. Therefore, gain is set to 40 in order to decrease amplification of signal.

The low pass filter is used as antialiasing filter which has the cut-off frequency 37.2 Hz. It is implemented as the Butterworth 4th order filter using two Sallen-Key sections. This enables sampling frequencies higher than 100 samples per second and the suppression of 50 Hz signal for at least 9 dB.

The final stage of the EEG amplifier is a programmable gain amplifier. This section is shown at Fig 4. Different gains are realized by jumpers because the speed of switching is not important. Special topology is chosen where the jumper properties do not affect the gain and resistors do not load amplifier output. Gains implemented by this configuration are 1.001, 10.01, 101.

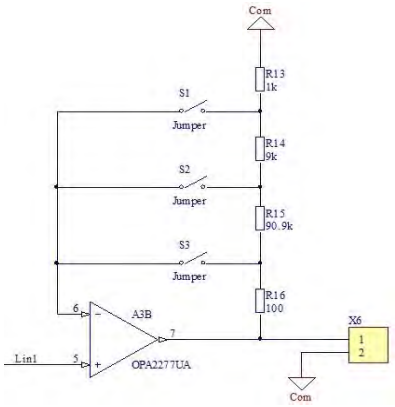


Figure 4. Programmable Gain Amplifier

B. *Signal adjustment and data acquisition*

The EEG amplifier is made ts one printed circuit board (PCB) and connected at bipolar battery supply ±5V. Before the signal is brought to ADC, it needs to be adjusted by special circuitry to fit the ADC unipolar voltage range 0-5 V. Circuitry is shown on Fig 4. When using PGA gain of 100 times, the final gain of the whole cascade is 40600. Since the highest expected input signal amplitude is 60 μV, then in total gain the signal is amplified to the value of around 2.5 V. The amplified signal in range of ±2.5 V is shifted by 2.5 V to meet the ADC bipolar supply requirements.

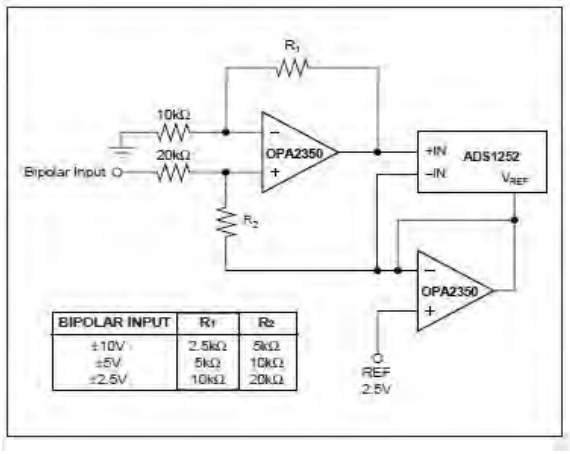


Figure 5. Signal adjusting circuitry [11]

As its shown at Fig 5., ADC used is ADS1252, which is chosen because of 24-bit resolution. Selected sampling rate is 200 S/s which is more than enough for EEG signals.

A microcontroller collects data by serial peripheral interface (SPI) and transmits them to PC by UART. Signals are received and processed in real time using MATLAB.

C. *Stimuli*

Flickering stimuli for SSVEP based BCI system must have stable frequency to insure less SSVEP jitter in frequency domain of EEG data. Three LED chessboards (Fig 6.) are designed for this purpose. They are powered by +5 V output from the microcontroller. Flickering frequencies are controlled by microcontroller PWM signals and can be changed within MATLAB GUI application.



Figure 6. LED chessboard stimuli

III. SOFTWARE

A. *Signal processing*

After obtaining the raw EEG signal from the occipital lobe, three steps of signal processing were done to extract relevant information. The first step is the preprocessing of the signal, the second is analyzing it and the third is the feature extraction and classification of the features. Those three steps are done in MATLAB.

During the real time measurement, EEG is constantly recorded. In the preprocessing step, a 4 seconds long block of raw EEG data is acquired every 0.5 seconds. This block is then filtered with the bandpass Butterworth filter of the 4th order with cutoff frequencies at 4 and 30 Hz. Filtering is needed to reduce noise from artefacts. The filtered block is then analyzed.

Spectrum components of the EEG signal containing SSVEPs are first processed with autocorrelation (2). Autocorrelation of the periodic signal is a periodic function of the same frequency and autocorrelation of aperiodic signal tends to zero [12]. Therefore, the periodic SSVEP signals are enhanced and background EEG activity is reduced.

$$R_{ff}(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t) f(t-\tau) dt \qquad (2)$$

For spectrum analysis of auto-correlated EEG data, a fast Fourier transform is used, due to the fast execution performance necessary for real time systems.

B. *Feature extraction and classification*

To determine on which chessboard the subject is focusing on, the amplitude of the EEG spectrum on flickering frequencies (f_1, f_2, f_3) and their first and second

harmonics are examined. The maximum sum of amplitudes is then chosen for classification feature F:

$$F(f1, f2, f3) = \max_{f=f_1,f_2,f_3} \{X(f) + X(f \cdot 2) + X(f \cdot 3)\} \quad (3)$$

If the maximum F is larger than threshold T, F is classified to classes LED1, LED2 or LED3 (corresponding to one of the three chessboard stimuli). The threshold T is determinate as a mean of the calculated EEG spectrum.

C. *Application*

To demonstrate the usage of SSVEP based BCI system, we simulated a real life application of turning lights on and off in the MATLAB GUI application (Fig 7).

The subject's task was to turn on one of the three light bulbs (by personal choice), while focusing on the corresponding flickering chessboard stimuli. Also, the EEG signal is shown in real time, both in time and frequency domain.

IV. EXPERIMENT

The experimental task was to turn on every light bulb in the application. Two healthy subjects with normal to corrected vision, averagely age 23, were tested. Subjects were in a room with dim lights, comfortably seated 1 m from the computer screen and the LED chessboards stimuli (Fig 8).

Three electrodes were placed on the subject's scalp. Two electrodes for the bipolar channel were set over position O1 and one referent ground electrode was stick to the forehead (position Fpz). Flickering frequencies were set to 6, 7 and 8 Hz. The subjects tried to turn each light bulb 10 times. Each successful task is marked as TP (true positive detection). False positive (FP) detections are also marked. Fig 9. shows results for both subjects.

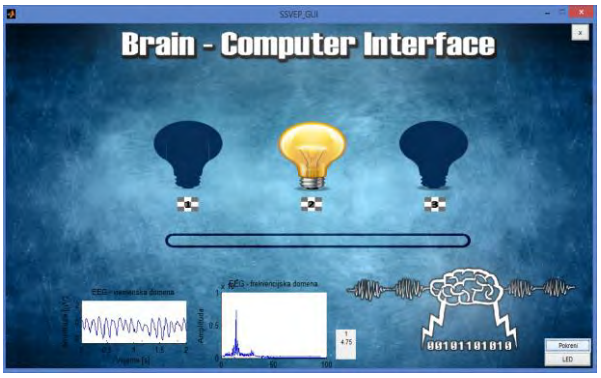


Figure 7. MATLAB GUI application; down left: time domain EEG signal – alpha waves; down middle: spectrum

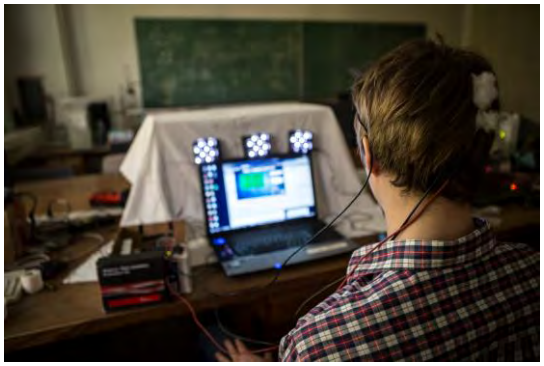
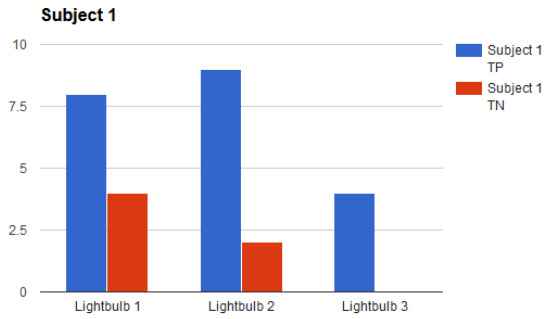


Figure 8. The subject during experimental task



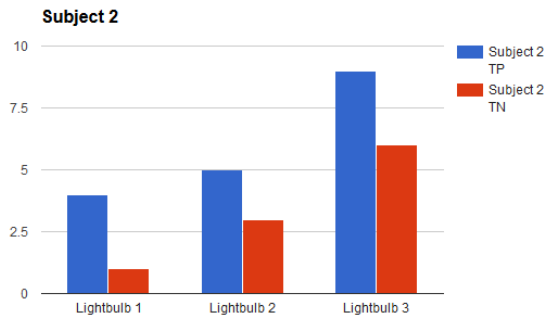


Figure 9. True positive and true negative detections for each stimulus

V. CONCLUSION

In this paper we have developed a BCI system based on SSVEPs. We propose a small, portable system that uses only one bipolar channel. Three LED chessboards are used to control simple application of turning light bulbs on and off, which can be easily implemented in to homes. The results show good detection for specific stimulus frequency. The methodology was tested for future work. Improvements of system focuses on frequency calibration for each subject and machine learning classification to improve robustness and accuracy.

REFERENCES

[1] Sanei S.,Chambers J.A. „EEG Signal Processing“, John Wiley & Sons Ltd., Sveučilište Cardiff, 2007.

[2] Šefer A.B. „Obrada elektroencefalografskog signala u svrhu ostvarivanja sučelja mozak-računalo“, Postgraduate doctoral study coursework, Univeristy of Zagreb, 2011.

[3] Hsiang-Chih C. Et al. Real-Time Control of an SSVEP-Actuated Remote-Controlled Car, SICE Annual Conference, 2010, pp. 1884-1887.

[4] Chun-Shu W. Et al. Detection of Steady-state Visual-evoked Potential Using Differential Canonical Correlation Analysis, 6th Annual International IEEE EMBS Conference on Neural Engineering, 2013, pp 57-60.

[5] Guangyu B. Et al. The SSVEP Topographic Scalp Maps by Canonical Correlation Analysis, 30th Annual International IEEE EMBS Conference, 2008, pp. 3759-3762.

[6] I. Krois, “Measurement System for Neurophysiological Functions Monitoring,” *Doctoral Thesis*, University of Zagreb, 2001.

[7] Whitchurch A. K., Abraham J. K., Lonkar M.A., Varadan V. K., „Design of a compact amplifier and signal conditioning module for wireless EEG monitoring “, IEEE Region 5 Tehnical Conference, 2007., pp. 153 156.

[8] Zhang L., Guo X., Wu X, Zhou B. „Low-cost circuit design of EEG signal acquisition for the brain computer interface system “, IEEE BMEI 6th International Conference, 2013., pp. 245-250.

[9] Tan Y. Q., Ibrahim F., Moghavvemi, „Two Electrodes Low Voltage Input EEG Amplifier for Brain-Computer-Interface “, ICIAS 2007, pp. 315-320.

[10] Blurr-Brown Application Guide, „Precision, Low Power Instrumentation Amplifier “

[11] Blurr-Brown Application Guide, „24-Bit, 40kHz Analog-To-Digital Converter, ADS1252“

[12] Lončarić, S., „Random processes in systems – lectures “, Faculty of Electrical Engineering and Computing, Zagreb, 2008.

Comparison of Wireless Electrocardiographic Monitoring and Standard ECG in Dogs

A. Krvavica*, Š. Likar*, M. Brložnik**, A. Domanjko-Petrič***, V. Avbelj****

* Student of Veterinary faculty, University of Ljubljana, Slovenia
** PRVA-K, Small Animal Clinic, Gorkičeva 6, Ljubljana, Slovenia
*** Clinic for small animals and surgery, Veterinary faculty, University of Ljubljana, Slovenia
**** Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia

Correspondence: viktor.avbelj@ijs.si

Abstract – **Electrocardiographic (ECG) data obtained by a wireless body electrode attached to the skin and connected to a smart device via low power Bluetooth technology were compared with a standard ECG in 8 dogs. The ECG data were gained from dogs with suspected arrhythmias due to cardiac or systemic diseases. A 2-minute standard ECG has been compared to a 15-minute recording obtained with wireless body electrodes. It has been established that this wireless electrocardiographic monitoring is a sensitive and specific method for identification of heart rates, duration of ECG waves and arrhythmia. When compared to a standard ECG, equivalent results were obtained for the heart rate and duration of different waves. Due to longer recording time the wireless device was more sensitive for documenting arrhythmias. With the wireless body electrodes, the ECG data were obtained while the dogs were lying down, standing or walking. The wireless electrode proved to be reliable and simple to use. This device enables a good option of long-term monitoring of canine cardiac rhythm in real-world environment.**

I. INTRODUCTION

In veterinary medicine the standard electrocardiogram (ECG) is used in every day practice to document many cardiac diagnoses, especially arrhythmias [1]. With the development of electronics over the past several years, many new approaches have been proposed to replace conventional methods of ECG monitoring [2 %23]. Wireless ECG monitoring that allows a long%term evaluation is more suitable, since in veterinary hospitals many dogs have heart rhythm disturbances that may only occur intermittently. In addition to the long%term measurements wireless ECG is easier to use and more comfortable to wear. Electrodes are suitably small, attached directly to the skin with a patch carrier and waterproof [1, 6]. Dogs are often disrupted by wires, which prevent them from moving and this causes additional stress that can lead to misinterpretation of results and inadequate diagnosis. Furthermore, the characteristics of ECG signals are highly dependent on the patient's physical activity; it is therefore very important to monitor ECG during physical activity, also [4].

The wireless ECG monitor is a very promising method by which diagnostic accuracy of the ECG measurements could be obtained. The main goal of our measurements was to compare a standard ECG with the wireless ECG monitoring to determine the usefulness of the latter in canine electrocardiography.

II. PRESENTATION

A. Materials and Methods

Wireless ECG sensor consisted of an electronic module with battery and two self%adhesive electrodes with a distance of 9 cm (Fig. 1). If necessary to prevent detachment, the ECG sensor was bandaged.

The ECG sensor recorded one bipolar lead. Wireless communication with the sensor via low power Bluetooth technology allowed the display of the ECG signal on a smart device. This device records real%time data from the electrodes and has already been described in the literature [2, 16, 23]. Evaluation of the ECG recordings was performed with NevroEKG software (Jožef Stefan Institute, Clinical Centre Ljubljana). The software can also extract the respiration rate based on the amplitude changes of QRS complexes [15].

The ECG data were obtained from 8 hospitalized dogs with suspected arrhythmias due to cardiac or systemic disease (5 and 3 dogs, respectively). The diagnoses of the dogs with cardiac disease were myxomatous mitral valve disease, tricuspid valve disease, pericarditis, pericardial effusion, and dilated cardiomyopathy (DCM). The systemic diseases included electrolyte disturbances, gastric dilatation and volvulus (GDV), pyelonephritis, and splenic tumor. During measurements, animals were lying in cages or moving freely in the room. The duration of wireless recording of the cardiac rhythm was 15 to 30 minutes.

All data were compared to the recordings with the 9%lead ECG, which included 6 limb leads and 3 precordial leads (V1, V2, V4) (Fig. 2).



Fig. 1: Wireless body ECG sensor placed on the left side of the thorax, negative electrode near atria and positive electrode at the apex of the left ventricle.



Fig. 2: Simultaneous recording of the 9% lead and wireless ECG.

B. Results

In eight patients, heart rate and rhythm from the wireless ECG data were compared with the data from the standard ECG.

Six dogs were placed in cages during the measurements and could not move away from the smart device. Two dogs were moving freely in the room. Not many movement artifacts or loss of the signal were noticed. When skin under the sensor was folded, the ECG waves were small, because the distance between electrodes was shorter. When the negative electrode was placed at the left base of the heart, i.e. near atria, the P wave was clearly seen. This position also caused more artifacts due to the moving of the animal's left front leg moving.

Instantaneous and average heart rates measured with the two devices were identical in all cases.

In four dogs (50%) the results of the wireless ECG data were equivalent to the results of standard ECG for detection of type of arrhythmia. Ventricular tachycardia, ventricular premature complexes (VPCs), sinus tachycardia, and normal sinus rhythm were all detected with both ECG devices.

In three dogs (37.5%) the extended ECG monitoring time increased the diagnostic yield of arrhythmias, the wireless device detected more arrhythmia than the standard ECG. The patient diagnosed with a DCM had a normal sinus rhythm on the standard ECG. However, the wireless sensor detected several VPCs (Fig. 3). In both recordings, P waves were clearly seen. In the patient, diagnosed with a splenic tumor, sinus bradycardia was detected in both measurements, but an episode of idioventricular rhythm and frequent VPCs were observed additionally with the wireless sensor (Fig. 4). The patient with tricuspid regurgitation and pleural effusion had atrial fibrillation on the standard ECG. The wireless sensor documented atrial fibrillation with several VPCs and a short asystole (Fig. 5 and Fig. 6).

In one case (12.5%), which was a patient monitored after surgery of GDV, the recording of the wireless device was diagnostically less useful than in other cases due to respiratory artifact caused by panting. In the recording T waves could not be followed, but arrhythmia and QRS intervals could still be observed.

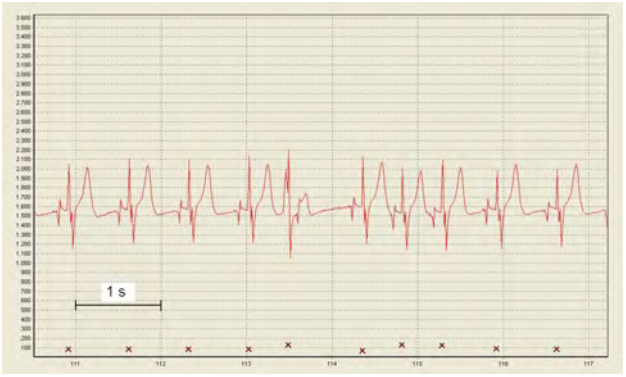


Fig. 3: Display of the recording with NevroEKG software of a dog, heart rate is 85 beats/min, P waves are seen and the 5th beat is a ventricular premature complex (VPC).

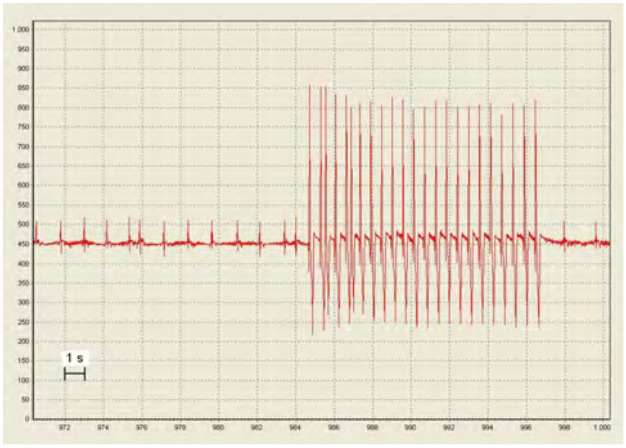


Fig. 4: Display of the recording with NevroEKG software, sinus bradycardia rate was 52 beats/min before and after idioventricular rhythm with the rate of 112 beats/min.

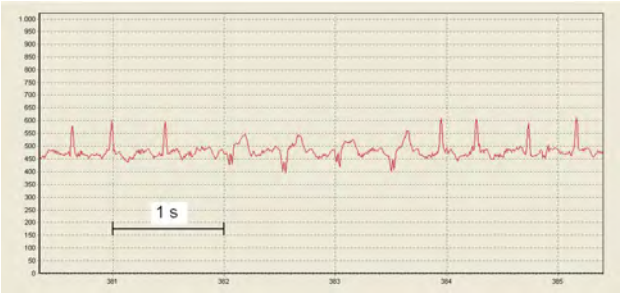


Fig. 5: Display of the recording with NevroEKG software, heart rate 140 beats/min, atrial fibrillation with four ventricular beats.

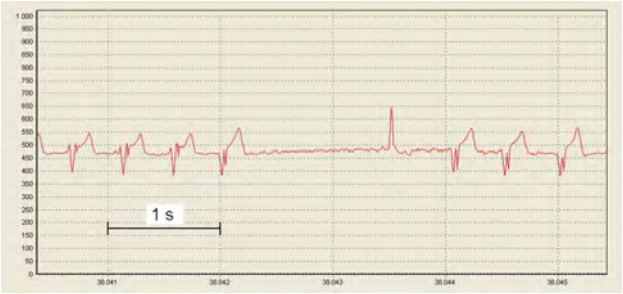


Fig. 6: Display of the recording with NevroEKG software, four ventricular beats of a rate 140/min, asystole, one atrial beat and three ventricular beats.

C. Discussion

There is an increasing demand by veterinarians and dog owners to continuously monitor dogs' vital signs, including ECG. In some breeds with high risk of cardiomyopathies (Boxers and Dobermans for example), the arrhythmias can be early indicators of an occult disease and can cause sudden death [2, 24]. Therefore the ECG sensor could be a useful tool for early detection and diagnostics. In veterinary medicine there are reports of other wearable sensors for monitoring the vital signs, but these devices are more invasive or have wires attached to the body [5, 19].

In our series of patients, the wireless ECG monitoring proved to have many advantages in monitoring heart rate and rhythm over standard short-time ECG recording. One major advantage is unlimited length of recordings that can be reviewed subsequently. Another advantage is a long autonomy due to wearable comfort, and a long battery life of the sensor. Additionally, it can be used also in a moving patient. These advantages have been reported previously [2, 6, 8, 9, 11]. The system used in this report has no wires, smaller number of electrodes, and the ability to synthesize the multi-lead ECG [20, 21, 23]. The NevroEKG software enables quick and easy analysis of instantaneous heart rate and rhythm, as well as average heart rate during sampled periods. Additionally, the NevroEKG enables good overview of the recording so that any changes of the rhythm, heart rate or duration of the waves can be promptly noticed, even if the arrhythmias are intermittent [2, 16]. A small-sized wireless sensor enables its use in various breeds of dogs. Due to longer recording time the wireless device was more sensitive for documenting arrhythmias in 37.5% of the dogs in our study. A higher diagnostic yield with prolonged duration monitoring for detection of arrhythmia events has been reported previously [8, 10, 12, 24]. All arrhythmias documented with standard ECG were also documented with wireless sensor. Instantaneous and average heart rates measured with the two devices were identical in all cases. High diagnostic accuracy of the wireless devices has been noted in previous studies, as well [6, 8-14, 18].

There were problems encountered in the wireless ECG monitoring, also. For example, ECG signal changes with the position of the body, and in case of rapid movements of the thorax (jumping, running, panting, etc.) the evaluation of the ECG signal can be more difficult due to multitude of artifacts and changes in the shape and size of the signal. Of course, excessive motion artifacts are also common in ECG monitoring with Holter and Event monitors, which are currently used for long-term monitoring in veterinary medicine [25, 26]. Another problem were electrode patches, which did not attach firmly. The firmness of the attachment depended on the quality of the electrode patches and on the thoroughness of shaving the hair and dry-cleaning of the skin. Bandaging the sensor decreased the number of the artifacts due to detachment and excessive motion. Bandaging and special coats are used to prevent detachment and excessive motion in Holter and Event monitors, as well.

Our pilot study proved that wireless ECG monitoring can give satisfactory ECG recordings regardless of the position, motion or size of the dog.

III. CONCLUSIONS

The wireless electrocardiographic monitoring described in this paper is highly reliable, simple to use, reusable, lightweight, small and comfortable to wear. It can be used during physical activities and offers long-term monitoring, therefore enhancing diagnostic yield. Such devices could become an important tool for clinical use in veterinary medicine.

REFERENCES

[1] Côté E, "Electrocardiography and cardiac arrhythmias," in Textbook of veterinay internal medicine, 7th ed, vol. 3, S. J. Ettinger, E. C. Feldman, eds. St.Louis: Elsevier Saunders, 2010: pp. 1159–1186.

[2] M. Brložnik, V. Avbelj, "Wireless electrocardiographic monitoring in veterinary medicine," Proceedings of MIPRO 2015, Opatija, Croatia, 2015, pp. 356–359.

[3] M. S. Kheertana, A. E. Manjunath, "A survey on wearable ECG monitoring using wireless transmission of data," International journal of advanced research in computer and communication engineering, vol. 4, 2015, pp. 277–279.

[4] F. Miao , Y. Cheng, Y. He, H. Qingyung, Y. Li, "A wearable context-aware ECG monitoring system integrated with built-in kinematic sensors of the smartphone," Sensors, vol.15, 2015, pp. 11465–11484.

[5] H. Mongue-Din, A. Salmon, M. Y. Fiszman, Y. Fromes, " Non-invasive restrained ECG recording in conscious small rodents: a new tool for cardiac electrical activity investigation," Eur. J. Physiol., Vol. 454, 2007, pp. 165–171.

[6] S. S. Lobodzinski, “ECG patch monitors for assessment of cardiac rhythm abnormalities,” Prog. Cardiovasc. Dis., vol. 56, 2013, pp. 224–229.

[7] I. Tomašić, R. Trobec, “Electrocardiographic systems with reduced numbers of leads%synthesis of the 12%lead ECG,” IEEE Rev. Biomed. Eng., vol. 7, 2014, pp. 126–142.

[8] P. M. Barret, R. Komatireddy, S. Haaser, S. Topol, J. Sheard, J. Encinas, A. J. Fought, E. J. Topol, “Comparison of 24%hour Holter monitoring with 14%day novel adhesive patch electrocardiographic monitoring,” Am. J. Med., vol. 127, 2014, pp. 95 e11–7.

[9] M. P. Turakhia, D. D. Hoang, P. Zimetbaum, J. D. Miller, V. F. Froelicher, U. N. Kumar, X. Xu, F. Yang, P. A. Heidenreich. “Diagnostic utility of a novel leadless arrhythmia monitoring device,” Am. J. Cardiol, vol. 112, 2013, pp. 520–524.

[10] D. Schreiber, A. Sattar, D. Drigalla, S. Higgins, “Ambulatory cardiac monitoring for discharged emergency department patients with possible cardiac arrhythmias,” West. J. Emerg. Med., vol. 15, 2014, pp. 194–198.

[11] S. L. Higgins, “A novel patch for heart rhythm monitoring: is the Holter monitor obsolete?,” Future Cardiol., vol. 9, 2013, pp. 325–333.

[12] M. Rosenberg, M. Samuel, A. Thosani, P. Zimetbaum, “Use of a noninvasive continuous monitoring device in the management of atrial fibrillation: a pilot study,” Pacing. Clin. Electrophysiol., vol. 36, 2013, pp. 328–333.

[13] G. Hindricks, E. Pokushalov, L. Urban, M. Taborsky, K. H. Kuck, D. Lebedev, G. Rieger, H. Pürerfellner, XPECT trial investigators, “Performance of a new leadless implantable cardiac monitor in detecting and quantifying atrial fibrillation: Results of the XPECT trial,” Circ. Arrhythm. Electrophysiol., vol. 3, 2010, pp. 141–147.

[14] M. Schrivastav, S. Padte, V. Arora, M. Biffi, “Pilot evaluation of an integrated monitor%adhesive patch for long%term cardiac arrhythmia detection in India,” Expert. Rev. Cardiovasc. Ther., vol. 12, 2014, pp. 25–35.

[15] R. Trobec, A. Rashkovska, V. Avbelj, “Two proximal skin electrodes – a respiration rate body sensor,” Sensors (Basel), vol. 12, 2012, pp. 13813–13828.

[16] A. Rashkovska, I. Tomašić, R. Trobec, “A telemedicine application: ECG data from wireless body sensors on a smartphone,” Proceedings of MIPRO 2011, Opatija, Croatia, 2011, pp. 293–296

[17] R. Trobec, M. Depolli, V. Avbelj, “Wireless network of bipolar body electrodes,” Proceedings of the 7th International Conference on Wireless On%demand Network Systems and Services, WONS 2010, pp. 145–149.

[18] M. S. Kraus, F. C. Brewer, M. Rishniw, A. R. Gelzer, “Comparison of the AliveCor® ECG device for the iPhone with a reference standard electrocardiogram,” ACVIM Seattle 2013: abstract.

[19] R. Brugaloras, J. Dieffenderfer, K. Walker, A. Wagner, B. Sherman, D. Roberts, A. Bozkurt, “Wearable Wireless Biophotonic and Biopotential Sensors for Canine Health Monitoring,” IEEE 2014.

[20] I. Tomašić, R. Trobec, V. Avbelj, “Multivariate linear regression based synthesis of 12%lead ECG from three bipolar leads,” Proceeding of the 3th Int. Conf. Health Informatics, HealthInf 2010, pp. 216–221.

[21] R. Trobec, I. Tomašić, “Synthesis of the 12%lead electrocardiogram from differential leads,” IEEE Trans. Inf. Technol. Biomed. vol. 15, 2011, pp. 615–621.

[22] I. Tomašić, S. Frljak, R. Trobec, “Estimating the Universal Positions of Wireless Body Electrodes for Measuring Cardiac Electrical Activity,” IEEE Trans Biomed Eng 2013, vol. 60, pp. 3368–3374.

[23] A. Rashkovska, I. Tomašić, K. Bregar, R. Trobec, “Remote monitoring of vital functions % Proof%of%concept system,” Proceedings of MIPRO 2012, Opatija, Croatia, 2012, pp. 463–467.

[24] G. Wess, A. Schulze, N. Geraghty, K Hartmann, “Ability of a 5% minute electrocardiography (ECG) for predicting arrhythmias in Doberman Pinschers with cardiomyopathy in comparison with a 24%hour ambulatory ECG,” J. Vet. Intern. Med., vol. 24, 2010, pp. 367–371.

[25] Miller RH, Lehmkuhl LB, Bonagura JD, et al. “Retrospective analysis of the clinical utility of ambulatory electrocardiographic (Holter) recordings in syncopal dogs: 44 cases (1991–1995),” J Vet Intern Med 1999, vol. 13, pp. 111–122.

[26] J. M. Bright, J. V. Cali, “Clinical usefulness of cardiac event recording in dogs and cats examined because of syncope, episodic collapse, or intermittent weakness: 60 cases (1997–1999),” J Am Vet Med Assoc 1999, vol. 216, pp. 1111–1114.

A Medical Cloud

Jurij Tasic
Univerza v Ljubljani, Slovenia
{jurij.tasic}@fer.ulj.si

Marjan Gusev, Sasko Ristov
Ss. Cyril and Methodius University, Skopje, Macedonia
{marjan.gushev, sashko.ristov}@finki.ukim.mk

***Abstract*—Recent trends and development in the production of various wearable biosensors enable a lot of medical and environment information to be available for each human being. Processing of data coming from these sensors, extracting valuable information and analyzing the electronic health record with a sufficient expertise is a complex processing task, which requires more resources than an ordinary mobile device or personal computer can perform with the available technology today.**

In this paper, we propose a cloud-based solution to deal with these challenges. The design includes use of wearable biosensors, personalized medical devices as transmitters of sensor data and the cloud that delivers various healthcare services over the Internet. A medical cloud hosts a specially developed application, which communicates with medical devices and sensors from one side and caregivers, such as a medical institution. The services include analysis by an expert ICT system, and also by a medical expert in case of an alert for detected problems analyzing the current state by received sensor data.

***Index Terms*—cloud computing, mobile healthcare, biosensors, cloudlet, fog computing**

I. Introduction

Although cloud computing [1] is not a new concept to the technological advancement, its potential use in healthcare is rising on the horizon. One can witness the benefits of this technology that recently is taking into practice, especially in cases when the personalized mobile devices, such as smartphones, tablets or laptop computers can not manage with complex computations and storage needs.

Recently, a variety of biosensors [2] is produced as wearables on a human body [3], such as ECG monitors, heart rate monitors, glucose monitors, pulse oximeters, body temperature, humidity and blood pressure monitors, SpO2 and various hemodynamic parameters of the patient. Additional information is gathered by various motion and position sensors, including GPS, accelerometer, gyroscope, step meter, etc. They can form a Medical Body Area Network (MBAN) which is a small-scale wireless communication network [4] used as a technology to provide medical-related raw data from attached sensors.

The trends show that in near future all these sensors will be realized on a smaller scale to provide continuos medical diagnosis, including micro and nano chemical smart sensors [5].

A lot of research is conducted aiming at provision of mobile healthcare systems, and most of these designs are using clouds for computation offload. For example, Doukas et al. [6] present an implementation of a mobile system based on Android that uses a cloud for electronic healthcare data storage, update, and retrieval.

Many services have to be realized in order to deal with multimodal and heterogeneous physiological signals and provide persistent personalized services. They require complex computations and need offload to a cloud. However, most of the current designs refer to a cloud as a storage system for Electronic Health Records (EHRs). All current solutions are based on static information and their design involves a data processing server. In this paper, we introduce a medical cloud that is capable of dealing with streaming and static structured data from EHRs.

The rest of the paper is organized as follows. Section II describes the architecture and functional description for building a medical cloud. Several design aspects of such a medical cloud are presented in Section III along with an analysis of essential application characteristics, such as elasticity and scalability, security and privacy, quality of service, etc. Section IV discusses the applicability, benefits, pricing and other non-functional issues and challenges. Related work of various approaches for dealing with personalized healthcare and a comparison of other approaches to the proposed medical cloud are given in Section V. The conclusions are presented in Section VI.

II. Organization of a Medical Cloud

In this section, we will discuss the architecture of the proposed solution, data communication issues, an organization of data processing and an overview of functions of the medical cloud.

A. Architecture design

The three-tier architecture of a medical cloud is presented in Figure 1. Various biosensors attached to a person can stream raw data to the nearby mobile device. Most of the sensors form the MBAN, they can also be independent sensors made as a very small wearable units that can stream data wirelessly to a receiver, which is positioned at a very close distance. This saves energy and the battery life of the biosensors, that are not easily detachable and can be re-charged. It will make them last for a longer period without any interruption in streaming data.

B. Data communications

If we analyze a given sensor as a transmitter in our design, then the receiver needs to be at a very close distance. Typically, the solutions offered by various researchers in the literature

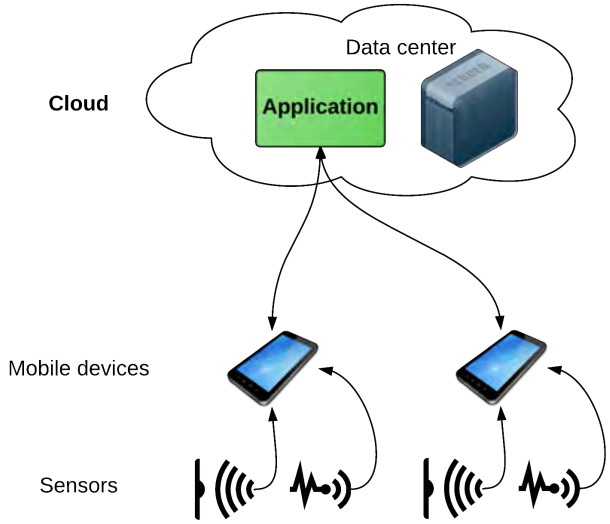


Fig. 1: Three tier architecture of a medical cloud

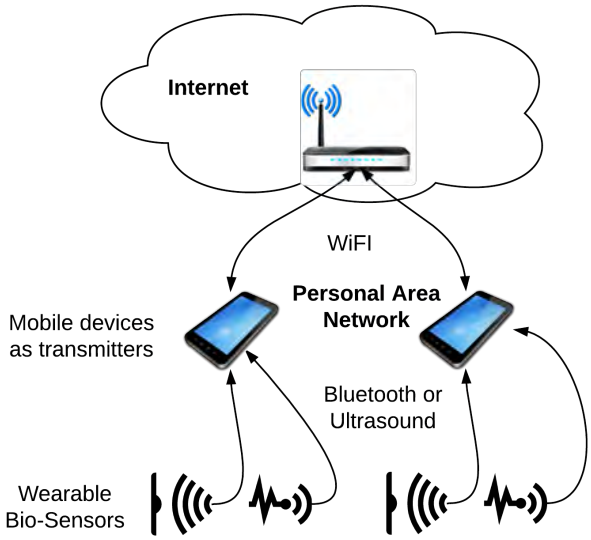


Fig. 2: Signal description of the proposed design

include gateways or similar devices, which are positioned in a home environment. These solutions need sensors that transmit more powerful wireless signals at an average distance of 5-10m.

Analyzing the new trends in production biosensors on a micro scale, that are very small and can transmit to distances that are at most 1m, and on average at a 0.5 m. Therefore, our solution uses a nearby device capable of receiving these signals and then forwarding them with the use of WiFi integrated in the smart mobile device to the cloud. This concept of using a mobile device just as an intermediator that receive weak signals and transmits them to the cloud via Internet is new and is not used in the solutions proposed as prototypes or a limited number of realized commercial products.

This idea is presented in Figure 2. Instead of using WBAN or WSN, where the sensors are powerful enough to transmit WiFi over 5m range, this design is based on the use of wearable biosensors that can transmit signals with a very low emitted power at 10 times fewer distances. For example, these wearable biosensors can communicate with ultrasound or low-power Bluetooth to the personalized devices in the Personal Area Network (PAN). These mobile devices can transmit received signals to the Internet using the WiFi router.

This design also solves the mobility problem. In case the person that uses wearable biosensors and PAN is not in an area that uses known WiFi network, then the mobile device can use its connectivity to the mobile network, such as 4G or similar. It is advised to avoid these situations as much as possible since the achieved speeds usually are higher on WiFi connected to an optical cable operator instead on Internet connectivity through mobile operator.

The mobile device can also contribute to the determination of other environmental data using the built-in sensors, such as GPS, accelerometer, gyroscope, step meter, motion sensors etc. These environmental sensors can provide sufficient information, such as the position of the person, its motion, and physical activity.

C. Data processing

Although this design does not include any processing on the mobile device, there are at least two occasions when this can be realized. The first case happens when the mobile device can not establish a communication to any of the WiFi or mobile operator. For example, it is happens in planes or other situations when there is no service by a mobile operator or WiFi. In this case, the raw data storage and initial data processing should be realized on the mobile device itself.

The second occasion is initiated by the necessity to react as fast as possible. This includes initial stream processing with the determination of abnormal situations that need further alerting. For example, in the case when received values of sensor data exceed the extreme values set as a normal physiological condition, then there is a need for an alert, as soon as possible. This processing is usually done on-the-fly, whenever the stream data arrives at the mobile device, with a small number of instructions performed on each received data item, to keep the battery life of the mobile device as long as possible. Usually, this processing is based on a comparison of the incoming data item with pre-defined low and high levels. Alerting is a consequence of detecting an abnormal physical condition and can include sending messages to the user, and also to its caregivers, such as medical institutions or close relative or friend subscribed to it.

Although the solution initially proposed by Chen et al. [7] is a three-tier architecture, it is different from our approach since it is using a smart gateway as an intermediator between

the sensors and the server, instead of using a mobile device to transmit relevant information from sensors to the cloud.

Fog computing concept also includes a three-tier architecture, but in this case, all communication and processing goes to an infrastructure offered by the mobile operator. However, these solutions are almost always vendor lock-in situations, when the user can not transfer its data and processing to another provider.

D. Functions of the medical cloud

The medical cloud can support several information-based services used for delivery of personalized healthcare:

- receiving streaming raw data from sensors,
- storing unstructured received sensor data,
- pre-processing for abnormal physical condition,
- initiating alerts and messaging to caregivers,
- processing the raw data to extract relevant information,
- storing the structured information in EHR, and
- processing of latest physical state by an expert system.

The medical cloud is capable of receiving multiple streams of sensor data and storing them in a sequential record structure. As a part of the stream processing when each data item arrives, the system is capable of determining if its meaning is in a range of predefined values, or if it in combination with another relevant information exceeds the predefined normal values. In the case of detecting an abnormal physical condition, then a proper alert is initiated along with further actions that deliver this message to the person itself or to the caregiver (medical institutions or close relatives or friends) subscribed to receive such messages. Note that this pre-processing and storage of unstructured raw data can be done in a more efficient way, as presented in the next Section.

Extracting the relevant information from the received raw data is another important function of the medical cloud. Raw sensor data needs to be processed and only the significant values should be transferred for further determination of the physical condition of the person. This structured data needs to be stored in a database as EHR. For example, if a sensor is transmitting ECG values from attached ECG wearable sensor, then the processed information may include heart rate, any abnormalities, such as arrhythmia. If this is the first occurrence of such a diagnosis, then the complete ECG is stored in the EHR with the detected diagnosis. If this is already detected, then only small relevant information that detects a change in a physical condition should be updated in the EHR.

Note that the current physical state is always matched to other environmental conditions received by other sensor data. For example, in the case of a physical exercise detected by the motion sensors and step meters, it is normal to expect increased heart rate.

In addition to these functionalities from ICT aspects, the medical cloud is capable of realizing a healthcare monitoring service. This means it is capable of presenting the relevant information from the EHR, including the latest information or the history of previously recorded values. Besides presenting values, it is also capable of presenting graphical medical images, such as any ECG data, MRI, X-ray or other images. Data can be presented only to subscribed caregivers.

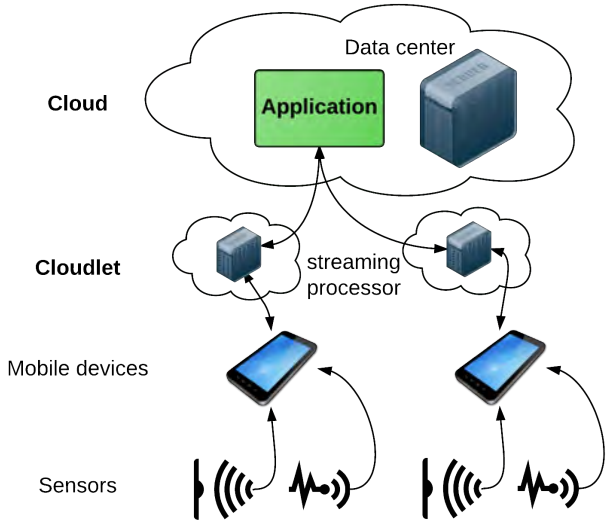


Fig. 3: Upgraded four tier architecture of a medical cloud

III. ICT Challenges in the Design of a Medical Cloud

In this section, we will analyze the several design aspects of the medical cloud, such as the elasticity and scalability, security and privacy, quality of service, availability and disaster recoverability.

A. Improved design

One huge challenge is to solve the communication demands in case of streaming data from multiple sensors. The initial problem, in this case, is the huge amount of received data, that needs to be received and processed, which creates a communication bottleneck or computation overload, which cause an inability to accept and process all received data.

In such a case the architecture of the medical cloud needs an upgrade by inserting another cloudlet tier, as presented in Figure 3. Cloudlets are local nearby servers that can process of streaming data and send processed information to the medical cloud. This design distributes the pre-processing of sensor data streams to several smaller and nearby servers situated in cloudlets.

Some IoT-based health monitoring solutions use smart gateways and local data center as an intermediate to transfer and process sensor data prior to their final storage and processing at the remote healthcare center, such as the solution proposed by Rahmani et al. [8]. Although, this is also a 4 -tier solution, it differs from ours solution. Instead of using mobile devices for transmission of sensor data to the cloudlet and afterward to the cloud, they use specialized gateways and local data centers.

B. Elasticity and Scalability

Scalability and elasticity are one of the main features of the medical cloud. The idea is not to be used only for storage of EHRs, but also to offer various services that diagnose the medical records and setup a diagnosis.

The improved design presented in Figure 3 is a highly scalable solution. When a given cloudlet is servicing numerous streaming sensors it may be a victim of a communication bottleneck or can face more processing demand than it can support. In these cases, scalability is solved by starting a new cloudlet instance and transferring some sensors from the existing cloudlet to the new one.

The need for porting a sensor from one to another cloudlet can be solved by conventional or specialized solutions that transfer all relevant data from the sensor to another location. In the case of having a system monitor that checks the state of existing computing and communication capabilities of a given cloudlet, this can be realized by an automatic procedure that leads to an elastic design. Of course, this process has also a downgrade component, meaning that to enable sufficient efficiency of the cloudlet, one has to optimize the number of tasks a cloudlet is performing. It means that the elasticity can both decrease and increase the workload of a given server in the cloudlet.

C. Security and Privacy

Security and privacy are very important for the existence of the medical cloud. Personal medical records have to be secured and available only to authorized users, such as responsible medical staff. The medical records are processed and stored in the cloud with sufficient security. This processing is also a matter of several legislations in various countries over the world, and the user of the medical cloud has to sign a contract to whom should the medical records will be available with transferring all reliability and trust to a medical institution or staff. The provider of the medical cloud should also guarantee that all known measures will be taken for data protection according to the available legislation and there will be no abuse of data.

Analyzing the overall solution, our proposal provides sufficient security and data privacy.

When a sensor transmits the streaming data to the nearby mobile device, the raw data is not encrypted or protected. Raw data is usually a stream of bytes without any identification of their meaning. The range of signal communication is very close and only devices that are very close the person can detect those signals. Only those devices, which are positioned very close to the person can damage these signals and initiate various intruder attacks. These devices are easily discoverable and can, therefore, be eliminated to prevent such situations.

When the mobile device transmits data to the cloudlet or cloud, it sends sensor id, and raw data, along with the date and location stamp. This data is still not encrypted to prevent any additional pre-processing on the mobile device and save its battery life. However, the established communication can be realized as a secured transaction, such as https. This prevents any data intrusion attacks and disables immediate capture of sensible data. Even in the case of capturing data, it is still not known for whom this data correlates, or what kind of raw data is sent, it is just a sequence of data bytes.

Data in the EHR stored in the cloud is protected according to the well-known data storage techniques. Data identification of the person with a corresponding EHR is kept on the level of the medical cloud and its provision of monitoring services.

Only authenticated and authorized users classified as caregivers can access the monitoring service offered by the medical cloud.

IV. Other challenges

This section discusses the applicability, benefits, pricing and other non-functional issues and challenges

A. Quality of Service

The quality of service is another important challenge. The provider of the medical cloud should guarantee that there will be no interruption in the provision of availability of offered services, and all measures will be taken to provide an alternative communication and processing server if there is an incident with the current server. The cloud provider has to guarantee that a proper disaster recoverability is enabled to keep the records of all interactions with the customer.

The improved design allows high performance since in case of increased utilization of a given server in the cloudlet over its capabilities, the system allows distribution of the workload to another server or cloudlet.

B. Benefits, pricing and other non-functional issues

The overall benefit of using a medical cloud are numerous. The users can obtain a personalized healthcare system, which is based on sensing some physiological parameters that can determine the overall physical condition. A huge benefit is using an expert ICT system that recognizes the sensing health related data and processes it by analyzing all related environmental and physical conditions. These systems can understand whenever some parameter is out of a predefined range of normal behavior and can initially alert in case of detection of an abnormal situation. In these cases, a medical expert from a caregiver medical institution can react and establish a more precise diagnosis. This concept gives the user healthcare protection.

The pricing strategy of using a medical cloud is based on the concept of provision of a Personalized Healthcare As A Service (PHAAS). This means that the users will pay a monthly subscription fee to have a constant monitoring of relevant wearable biosensors. The physical condition is matched upon other environmental information captured by the sensors built on the mobile device. An expertise system checks if the scanned data is in a predefined range assuming the complete physical condition of the user. If there is abnormality detected, such as exceeding a maximal value of some measured or calculated parameter then an alert is raised and several actions are initiated. Messages are sent to authorized caregivers in a

form of an alerting service. A typical monitoring and check are started if an expert from a medical institution registered as a caregiver receives such an alerting signal. This expert has the task to analyze what caused the alert and monitor the overall scanned physical state provided by the sensors. The expert can additionally get more information by establishing a direct mobile phone conversation with the user and in a case of emergency to call an ambulance or give and healthcare advice.

This solution solves a lot of challenges in establishing ubiquitous healthcare systems such as energy efficiency, scalability, and reliability issues. Our approach is a highly energy efficient solution, saving the battery life of both the sensors and mobile devices. Scalability and elasticity are solved by the improved design with distributing the stream processing to the nearby server in the cloudlet tier. The overall goal of the improved design is the provision of relevant Quality of Service and keeping the performance at high levels.

This solution also offers easy attaching of different wearable biosensors, by keeping the independence of any platform or operating system. Interoperability issues are also important in the design of new sensors and solutions to support the available technology and communication standards.

V. Related work

The technology used is based on concepts of cloudlets and fog computing.

Cloudlets have been proposed as a solution for various problems, such as face recognition at airports [9], based on early designs presented by Satyanarayanan et al. [10]. More cloudlet challenges are discussed in [11]. Fog computing [12] is also an alternative solution to this design by extending cloud services to the edge of the network [13].

Next, we will compare our design of a medical cloud with the existing approaches.

There are several patents that address a cloud-based solution for a healthcare provided by a mobile device. For example, the Patent US2014275928 by Acquista et al. describes a system and method for monitoring and diagnosing the patient condition based on wireless sensor monitoring data [14]. However, the main intention of the patent invention relates to one or more wireless sensors and a network of wireless sensors for monitoring of vital signs. Additionally, it provides improved diagnosis, monitoring, and treatment of medical conditions, based on the integration of wireless sensors with an electronic medical record and management system for patient healthcare. The cloud in this context is used as a monitoring device, or as a device that connects to the electronic medical record or patient health record. This invention does not cover our definition and architecture of a medical cloud.

He et al. [15] designs a novel efficient cloud platform for ubiquitous healthcare services. Their architecture consists of six layers according to the specific requirements. This platform utilizes message queue as a cloud engine aiming at a prototype cloud platform, with robust, stable, and efficient features, that can satisfy high concurrent requests from ubiquitous healthcare services.

Medcloud [16] is a healthcare cloud computing system, designed to enable scalability of patient's data storage. Moreover, the authors provide an architectural design of a personal health record system that utilizes and integrates services from Hadoops ecosystem. Their design presents data layer, server management layer (coordinator manager, query manager, concurrency manager, data storage manager, jobs manager) and application layer (authenticator, authorizor, request receiver, data integrator, NPP registry, authorizers registry, service registry, disclosure tracker).

Chen et al. [7] design a smart gateway for health care system using a wireless sensor network (WSN). They analyze the challenges of transmitting big quantities of sensor data to a remote server or to a cloud for data processing, healthcare decision making and sending emergency messages. To avoid communication delays and bottlenecks they have designed a prototype of a smart gateway dedicated to WSN health care systems at home environment. Their smart gateway consists of a central control unit, database (DB), WSN module, WLAN AP, and GSM module.

Rahmani et al. [8] also analyze a Smart e-Health Gateway gateway solution, where the smart gateways are considered as devices with beneficial knowledge and constructive control over both the sensor network and the data to be transmitted through the Internet, and also as devices that offer several higher-level services such as local storage, real-time local data processing, embedded data mining, etc.

Instead of having a specialized gateway, our solution is based on using available nearby computing and storage infrastructure capable of receiving and analyzing streaming data.

Suciu et al. analyze several new trends in IT, such as Big Data, Internet, and Cloud convergence, and propose an architecture for secure e-Health applications [17]. Their design is a typical Internet-based solution, where all relevant actors communicate via Internet. Sensors are connected to Internet via a remote telecommunication unit (RTU) that is usually connected with an assistance of a mobile network provider. This solution is a typical solution of the edge/fog computing model.

Fernandez and Pallis discuss the system engineering space [18], analyzing the business model, nonfunctional requirements, application context, physical environment and characteristics, and other relevant aspects. Kanth et al. [19] also present technological development towards personalized and pervasive healthcare systems. Their solution also addresses sensors, access points, and cloud to deal with patient-oriented services that transfer medical checks and healthcare services from hospitals to a home environment.

Ferrucci et al. [20] discuss how the IBM Watson technology can be applied to health care and describe a vision for an evidence-based clinical decision support system. It is based on the DeepQA technology, that affords exploration of a broad range of hypotheses and their associated evidence, as well as uncovers missing information that can be used in mixed-initiative dialog. It has been efficiently used in oncology decision making [21], [22].

VI. Conclusion

We designed a medical cloud that collects data from personalized biosensors for a human being and other related environmental sensors, communicates with personalized mobile devices, processes the available information, analyzes complex algorithms to determine the health state of the analyzed person and eventually set a medical diagnosis. The cloud communicates also with a medical service for extended consultation about the diagnosis and in a case of detecting a malfunction or a health problem, alerts the medical expert.

The functions that provide a personalized healthcare are rather complex to be realized by a personalized mobile device, so the suggested design to use a medical cloud is the only alternative. A three-tier architecture that consists of a medical cloud at the top level, mobile devices at middle tier and various wearable or environmental sensors on the lower level can be upgraded by a four-tier architecture, setting a cloudlet between the cloud and mobile devices. In this case, the cloudlet can be used for stream processing functions, since transmitting stream data on a distant cloud can be inefficient and cause communication bottleneck.

Our design is based on personalized mobile devices as an intermediator in the PAN with tasks to communicate with the wearable biosensors from one side and WiFi from the other side. Typical solutions found in the research literature or prototypes in labs use WBAN or sensors that use WiFi to communicate directly to the cloud or use a specially designed smart gateway that not only accepts the sensor data, but also processes it locally. These designs lack mobility if the gateway is situated in a home environment. If the concept of fog computing is used, then everything is set on the mobile operator infrastructure and there is no flexibility of the user to choose a solution, nor availability to avoid lock-in situations. Therefore, our design offers more benefits compared to other solutions.

Presented design of a medical cloud as an IoT-based health monitoring system offers enhanced overall system energy efficiency, performance, interoperability, security, and reliability. Following the advances in the realization of expert diagnosis and recommender systems, one can expect that in near future the medical cloud will offer full healthcare services, even without the expert help of a medical institution.

References

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica *et al.*, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[2] A. P. Turner, "Biosensors: sense and sensibility," *Chemical Society Reviews*, vol. 42, no. 8, pp. 3184–3196, 2013.

[3] J. R. Windmiller and J. Wang, "Wearable electrochemical sensors and biosensors: a review," *Electroanalysis*, vol. 25, no. 1, pp. 29–46, 2013.

[4] H.-B. Li, K.-i. Takizawa, B. Zhen, and R. Kohno, "Body area network and its standardization at ieee 802.15. mban," in *Mobile and Wireless Communications Summit, 2007. 16th IST*. IEEE, 2007, pp. 1–5.

[5] P. Rai, S. Oh, P. Shyamkumar, M. Ramasamy, R. E. Harbaugh, and V. K. Varadan, "Nano-bio-textile sensors with mobile wireless platform for wearable health monitoring of neurological and cardiovascular disorders," *Journal of The Electrochemical Society*, vol. 161, no. 2, pp. B3116–B3150, 2014.

[6] C. Doukas, T. Pliakas, and I. Maglogiannis, "Mobile healthcare information management utilizing cloud computing and android os," in *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*. IEEE, 2010, pp. 1037–1040.

[7] Y. Chen, W. Shen, H. Huo, and Y. Xu, "A smart gateway for health care system using wireless sensor network," in *Sensor Technologies and Applications (SENSORCOMM), 2010 Fourth International Conference on*. IEEE, 2010, pp. 545–550.

[8] A.-M. Rahmani, N. K. Thanigaivelan, T. N. Gia, J. Granados, B. Negash, P. Liljeberg, and H. Tenhunen, "Smart e-health gateway: Bringing intelligence to internet-of-things based ubiquitous healthcare systems," in *Consumer Communications and Networking Conference (CCNC), 2015 12th Annual IEEE*. IEEE, 2015, pp. 826–834.

[9] T. Soyata, R. Muraleedharan, C. Funai, M. Kwon, and W. Heinzelman, "Cloud-vision: Real-time face recognition using a mobile-cloudlet-cloud acceleration architecture," in *Computers and Communications (ISCC), 2012 IEEE Symposium on*. IEEE, 2012, pp. 000 059–000 066.

[10] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for vm-based cloudlets in mobile computing," *Pervasive Computing, IEEE*, vol. 8, no. 4, pp. 14–23, 2009.

[11] A. Bahtovski and M. Gusev, "Cloudlet challenges," *Procedia Engineering*, vol. 69, pp. 704–711, 2014.

[12] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*. ACM, 2012, pp. 13–16.

[13] I. Stojmenovic and S. Wen, "The fog computing paradigm: Scenarios and security issues," in *Computer Science and Information Systems (FedCSIS), 2014 Federated Conference on*. IEEE, 2014, pp. 1–8.

[14] A. J. Acquista, A. Kometz, L.-H. Ma, and J. Shambroom, "System and method for monitoring and diagnosing patient condition based on wireless sensor monitoring data," Mar. 17 2014, uS Patent App. 14/216,174.

[15] C. He, X. Fan, and Y. Li, "Toward ubiquitous healthcare services with a novel efficient cloud platform," *Biomedical Engineering, IEEE Transactions on*, vol. 60, no. 1, pp. 230–234, 2013.

[16] D. Sobhy, Y. El-Sonbaty, and M. Abou Elnasr, "Medcloud: healthcare cloud computing system," in *Internet Technology And Secured Transactions, 2012 International Conference for*. IEEE, 2012, pp. 161–166.

[17] G. Suciu, V. Suciu, A. Martian, R. Craciunescu, A. Vulpe, I. Marcu, S. Halunga, and O. Fratu, "Big data, internet of things and cloud convergence–an architecture for secure e-health applications," *Journal of medical systems*, vol. 39, no. 11, pp. 1–8, 2015.

[18] F. Fernandez and G. Pallis, "Opportunities and challenges of the internet of things for healthcare: Systems engineering perspective," in *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*, Nov 2014, pp. 263–266.

[19] R. K. Kanth, P. Liljeberg, T. Westerlund, H. Kumar, H. Tenhunen, Q. Wan, G. Yang, and L.-R. Zheng, "Information and communication system technology's impacts on personalized and pervasive healthcare: A technological survey," in *Norbert Wiener in the 21st Century (21CW), 2014 IEEE Conference on*. IEEE, 2014, pp. 1–5.

[20] D. A. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller, "Watson: beyond jeopardy!" *Artif. Intell.*, vol. 199, pp. 93–105, 2013.

[21] C. Schmidt, "Cancer: Reshaping the cancer clinic," *Nature*, vol. 527, no. 7576, pp. S10–S11, 2015.

[22] P. Bach, M. G. Zauderer, A. Gucalp, A. S. Epstein, L. Norton, A. D. Seidman, A. Caroline, A. Grigorenko, A. Bartashnik, I. Wagner *et al.*, "Beyond jeopardy!: Harnessing IBM's watson to improve oncology decision making," in *Journal of Clinical Oncology*, vol. 31, no. 15, 2013.

A Hospital Cloud-Based Archival Information System for the Efficient Management of HL7 Big Data

Antonio Celesti, Maria Fazio, Agata Romano, Massimo Villari
Deparmtent of Engeneering, University of Messina
Contrada di Dio - 98166 Messina, Italy
Email: {acelesti,mfazio,mvillari}@unime.it
romano.agata16@gmail.com

***Abstract*—Nowadays, the Open Archive Information System (OAIS) model is widely adopted in hospitals to manage data related to both doctors and patients. However, the Archival Storage systems of hospitals are typically based on old Relational DBMS that makes hard the management of patients' data especially in HL7 format. In fact, data have to be continuously parsed in order to be stored in relational databases and sent to other hospital systems. Considering also interoperable scenarios where HL7 data continuously grow, the management of patients' information can become very hard. In this paper, we discuss an OAIS system able to manage HL7 Big Data. In particular, considering a case of study of HL7 glucose observations in JSON format, we demonstrate that in a scalable scenario an archival storage for big data processing is more convenient for hospitals than traditional archival storage systems.**

Keywords*—*Cloud Computing, OAIS, HL7, Big Data, Hospital Information System.

I. Introduction

Future healthcare systems will need to efficiently respond to new requirements arising from current changes in demographic, organizational and resourcing factors, and due to the increasing proliferation of Information and Communication Technology (ICT) solutions. In this direction, also the European Commission is strongly pushing on the sustainability of health systems and on the necessary efforts towards economic recovery and innovation growth, thus investing about 500 millions of euros in cooperation projects and actions, as outlined in the Europe 2020 strategy [1]. Hospitals, clinics and care centers need to adapt their information management towards electronic systems where data management and interoperability will be key issues. In order to solve such issues, HL7 (Health Level 7) was introduced in 80's. It is a comprehensive framework with related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services [2]. Despite its faraway definition, only in recent years HL7 is having a rapid diffusion. This "second life" is mainly due to the recent evolution of ICT technologies and, in particular, to the adoption of XML (in HL7 v3) and JSON (in recent HL7 non-standard initiatives) formats.

Due to the huge amount of health data that need to be stored, transmitted and processed in real life applications (i.e., the well-known Big data problem), innovative distributed and scalable technologies have been developed to make traditional healthcare systems more efficient [3]. Thus, hospitals are looking at Cloud computing as fruitful solutions to implement their information systems and to share data with many different care facilities[4]. Currently, most of Hospital Information System (HIS) architectures have adopted centralized SQL-like Relational Database Management Systems (RDBMS) yet. However, we think that such solutions will not be shortly so adequate for the following reasons and/or trends: i) movement from cookbook to evidence-based medicine; ii) giving everyone a chance to participate in health processes; iii) building applications that make electronic health record (EHR) "smart"; iv) "domestication" of data for better public health reporting and research; v) healthcare IT vendors are looking at Software as a Service (SaaS) strategies; vi) usage of free public health data for informed strategic planning; vii) growing interest toward inter-operable HIS(s) according to a Cloud computing fashion.

In this paper, we present a new cloud architecture able to manage HL7-compliant big data, which is designed according to the directives for an Open Archival Information System (OAIS). OAIS is one of the major models adopted in HIS(s) and is aimed at preserving information and making it available for a designated community. Our cloud architecture is mainly composed by three elements, that are Ingest, Archival Storage, and Access. They are respectively responsible to receive and prepare data for the storage, store data, and prepare data for the transmission. More specifically, in order to make the OAIS able to manage big data and inter-operable, we present an inpmlementation of the architecture that includes a HL7 XML-to-JSON/JSON-to-XML parsers and the No-SQL distributed parallel database MongoDB. The parsing task from HL7 XML to HL7 JSON and viceversa is necessary to guarantee the interoperability between current versions of the HL7 standard, whereas MongoDB is one of the major NoSQL solutions for the management of big data and stores data in a JSON-like format (i.e., BSON). We tested the proposed solution performing some experiments on a dataset of HL7 observations on the level of glucose in the blood. Moreover, since many Hospitals currently use storage system based on SQL-Like solutions, we compare the performance of two different implementations of our architecture, that are based on MySQL and on MongoDB. According to the behaviour of the two solutions, we discuss

when the implementation with MongoDB outperform the one with MySQL.

The rest of the paper is organized as follows. Section II discusses related works. An overview of OAIS is provided in Section III, whereas a Hospital OAIS architecture based on HL7 and MongoDB is discussed is Section IV. An overview on HL7 v3 along with an example of observation on the level of glucose in the blood is discussed in V. Experiments are discussed in Section VI. Section VII concludes the paper.

II. Related Work

Cloud computing is a crucial technology to accomplish healthcare IT modernization, but it introduces new challenges for data integration and processing, information sharing, data mining and security. In [5], the authors propose a cloud-based approach for the design of interoperable Electronic Health Record (EHR) systems. The framework enables data interoperability by using the HBase database running on top of the Hadoop Distributed File System (HDFS), and the distributed coordination service Zookeeper. Supported standards include HL7, CDA, CCR, DICOM, X12, CCD, XML, NCPDP, EDI, Delimited Text, and raw ASCII. A solution specific for health data processing is proposed in [6], where a task-level adaptive MapReduce framework is introduced. The framework extends the generic MapReduce architecture addressing the varied arrival rate of big-data splits. The cloud architecture presented in [7] is a Virtualization PaaS (Platform as a Service) that could be usefully adopted for a city-wide healthcare appointment system. The architecture is based on a Virtual SQL engine, which supports SQL relational database with extended capabilities to handle cloud data virtualization, and a Data Grid distributed computing infrastructure, for distributed processing. To enable the exchange of healthcare data among different interested parties in an healtcare scenario (e.g. doctors, researchers and policy makers), the authors of [8] propose a publish/subscribe based system moderated through a web service. An important function of the web service is to provide the right level of abstraction to the publishers, and the right level of expressiveness to the subscribers, thus to make the healthcare data exchange possible in different medical contexts. Issues related to the secure management of 3D medical images are discusses in [9], where the architecture of a SaaS (Software as a Service) system is proposed to perform a lossless dynamic and adaptive compression of 3D medical images. A cyber-physical system for patient-centric healthcare applications and services, called Health-CPS, is presented in [10]. It is built on cloud and big data analytics technologies and consists of a data collection layer with a unified standard, a data management layer for distributed storage and parallel computing, and a data-oriented application layer. A privacy-preserving framework to transit insensitive healthcare data from trusted private to commercial public clouds is presented in [11]. The framework provides privacy protection features and defend against potential collusion between public cloud service providers and data users. Related to security issues in health data management, authors in [12] present an architecture for inter-organization data sharing, which provides a security and privacy features for patient data management in semi-trusted cloud computing environments. The architecture exploits attribute-based encryption techniques for selective access authorization and cryptographic secret sharing in order to disperse data across multiple clouds.

III. OAIS Overview

The Open Archival Information System (OAIS) functional model is the major standard for the digitization of European and national health systems. It became an ISO standard in 2003 and it was update in 2012. In particular the OAIS model defines the concepts, models, and functionalities inherent to digital archives, digital preservation and therefore long-term digital storage preservation. The pieces of information that are stored in digital form should be long-term preserved for users that can be far in terms of time, space, and cultural context. The objective of OAIS is to lead governments in the de-materialization of documents. OAIS is the reference model for the development of open "Digital Preservation's Centers" or "Deposits of Digital Archives" which have the primary purpose of preserving the long-term digital storage. The term "open" means that pieces of data have to be available for any interested subject or organization. The objective of the standard isn't to preserve a document, but rather to preserve the bit string and further metadata required to represent the document itself.

In the following we describe the main OAIS actors and components. *Producer* and *Consumers* are legal people, public or private subjects, who/which produces/retrieves information objects. They have to respect the requirements and procedures previously agreed with the administrator of the archive. The producer sends pieces of data to OAIS that, through an acquisition process (i.e., ingest process), accepts and prepares them for the memorization, whereas the consumer queries pieces of data to the OAIS that, through an access process retrieves and prepare data outputs. In a Hospital Information System (HIS) producers and consumers can be doctors, patients, medical equipments, or other HIS(s). The OAIS functional model is based on six macro-components: Ingest, Data Management, Archival Storage, Administration, Preservation Planning, Access. *Ingest* includes the set of processes responsible to accept digital resources sent by producers and to store them in the *Archival Storage* that is the part of the storage system that manages the memorization, recovery, and long-term availability of data. *Data Management* is responsible to control the Archival Storage by means of a set descriptive metadata associated to archived information. *Preservation Planning* provides services responsible to implement adaptation strategies when upgrate related to users, system, and technology occur. In addition, it guarantees the interoperability between legacy systems. *Access* enables users to locate and retrieve information about the stored information. In the end, *Administration* is responsible to support the management of daily operations on OAIS. Moreover, it coordinates the activities of the other macro-components and the interactions with Producers. In addition, there are three different types of information packages: entry or Submission Information Package (SIP), storage or Archival Information Package (AIP) and distribution or Dissemination Information Package (DIP). SIP is sent by the Producer to the Ingest macro-component that codifies it for the memorization in the Archival Storage. The resulting package is called AIP. When a particular consumer requests pieces of information, he/she/it contacts the Access macro-component that retrieves them from the Archival Storage in form of AIP. Subsequently, the Access macro-component transforms the AIP in DIP and

deliver it to the consumer. The DIP is codified according to the pieces of information that can be accessed by the requiring consumer. Examples of OAIS will be shown in the next Section.

IV. OAIS Based on HL7 and MongoDB

In this Section, firstly we discuss an OAIS architecture based on a "traditional" SQL-like DBMS, i.e., MySQL, acting as Archival Storage, and then, an alternative OAIS architecture based a NoSQL DBMS, i.e., MongoDB acting as Archival Storage. In order to guarantee standard compliance and interoperability between different HIS(s), in both cases we codified SIP(s) by using HL7 v3. In fact, we assumed that all messages between producers and consumers are exchanged in HL7 v3 format (e.g., the data collected from sensors placed on patients, messages exchanged between the hospital staff, etc). The choice of using the HL7 standard is due to the fact that it represents the main reference worldwide standard to share hospital data. Further details regarding HL7 are discussed in Section V.

Figure 1 shows an OAIS architecture based on HL7 and MySQL. In this scenario, all clinical reports are sent by

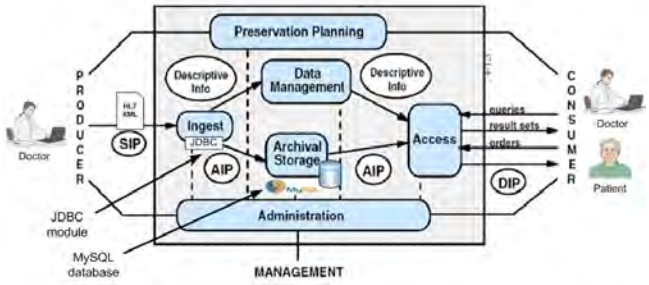


Fig. 1: The functional model of an OAIS system based on HL7 and MySQL.

doctors through agents (which act as producers). An agent is a software front-end with a human-readable graphic interface. A SIP is sent from the producer's agent to the Ingest in form of HL7 v3 (in XML format). The Ingest extracts data and through Java Database Connectivity (JDBC) Application Program Interface (API) produces an AIP that is sent to MySQL to store data. Consumers of this scenario can be both other doctors or patients. However, even though this scenario is interoprable thanks to the adoption of HL7 v3, it is not so scalable due to the centralized nature of MySQL. According to the aforementioned motivations, we propose an alternative OAIS architecture able to manage big data based on HL7 and MongoDB. Figure 2 shows an OAIS architecture based on HL7 and MongoDB. In this scenario, all clinical reports are sent by doctors through agents (which act as producers). A SIP is sent from the producer's agent to the Ingest in form of HL7 v3 (in XML format). Since MongoDB stores data in BSON format, in a JSON fashion, the Ingest macro-component has to parse each SIP from the HL7 XML format to the HL7 JSON format. In particular, we used a XML2JSON sub-component responsible to parse HL7 messages from XML to JSON format and to store data in MongoDB by using the official Java MongoDB driver. In our OAIS architecture, we chose to not use a native HL7

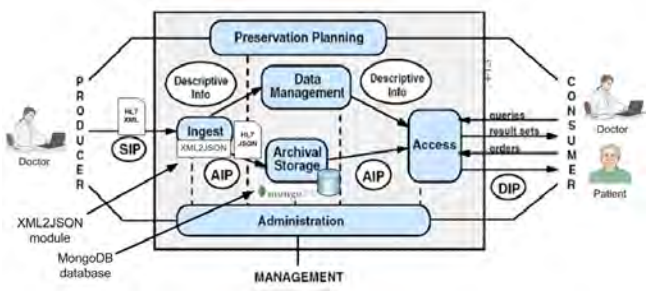


Fig. 2: The functional model of an OAIS system based on HL7 and MongoDB.

JSON format in order to maintain the compatibility among different health legacy systems based on the HL7 v3 standard that support message exchange in XML format. Figure 3 depicts the message flow between Ingest, Archival Storage and Access macro-components in an OAIS architecture based on HL7 and MongoDB.

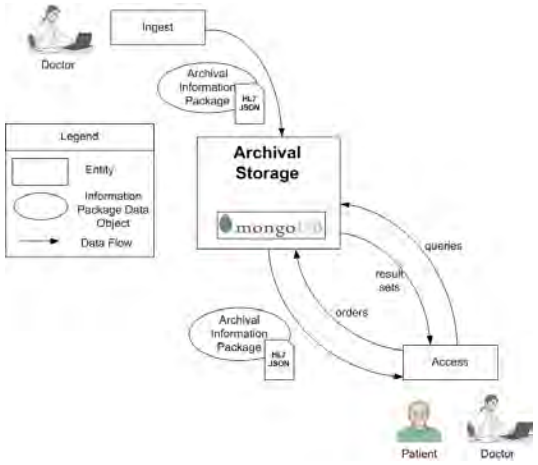


Fig. 3: Exchange flow of messages in the OAIS architecture based on HL7 and MongoDB.

V. Representation of Clinical Reports

In this Section, firstly we briefly discuss the main features of HL7 v3 [13] and then we analyze an example of HL7 document related to the observation of glucose level in the blood.

A. HL7 Overview

HL7 (Health Level 7) is a no-profit organization founded in the US in 1987 with the aim of promoting international standards in the field of information systems in the health care business, such as HIS, LIS (Laboratory Information System) and RIS (Radiology Information System), for healthcare. The organization aims to prove a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services. "Level Seven" refers to the seventh level of

the International Organization for Standardization (ISO) seven-layer communications model for Open Systems Interconnection (OSI) - the application level. The areas of application HL7 proposed by the Association concern various sectors including electronic clinical document structure, EPR (Electronic Patient Record) data format etc. that can be summarized as follows:

- development of conceptual standard, i.e., Reference Information Model (RIM);
- development of documents standard, i.e., Clinical Document Architecture (CDA);
- development of applications standard, i.e., Clinical Context Object Workgroup (CCOW);
- development of messaging standards (HL7 versions 2.x and 3.0).

```
<observationEvent>
<id root="2.16.840.1.113883.19.1122.4"
    extension="1045813"
    assigningAuthorityName="GHH LAB Filler Orders"/>
    <code code="1554-5" codeSystemName="LN"
          codeSystem="2.16.840.1.113883.6.1"
          displayName="GLUCOSE^POST 12H
              CFST:MCNC:PT:SER/PLAS:QN"/>
    <statusCode code="completed"/>
    <effectiveTime value="200202150730"/>
    <priorityCode code="R"/>
    <confidentialityCode code="N"
        codeSystem="2.16.840.1.113883.5.25"/>
        <value xsi:type="PQ" value="120" unit="mg/dL
            "/>
        <interpretationCode code="H"/>
        <referenceRange>
            <interpretationRange>
                <value xsi:type="IVL_PQ">
                    <low value="70" unit="mg/dL"/>
                    <high value="105" unit="mg/dL"/>
                </value>
            </interpretationRange>
        </referenceRange>
        <interpretationCode code="N"/>
</observationEvent>
```

Listing 1: Example of Content Domain.

HL7 CDA Release 2.0 (Clinical Document Architecture of Health Level 7) is a standard that specifies the structure and semantics of clinical documents written in XML, for the exchange within the healthcare domain. The purpose of HL7 CDA is therefore the standardization of clinical documents for the exchange and interoperability of medical data. An XML document constructed according with the standard CDA, is constituted by a series of elements, organized in a tree structure defined by the reference pattern. A CDA document represents a complete information object that may contain text, images, sounds and other multimedia content. It's composed of different information blocks that convey information about the patient, the doctor, the health facility, the author of the document, the signer of the document, the clinical events, the observations or medical procedures to which the document refers. Each document must be, as required the standard itself, "human readable" and therefore able to be viewed in an algorithmic by the receiver of the document without the need to know the specifics. The structure of a CDA document is formally derived by HL7 v3 RIM. The objectives of a CDA document are: give priority to the process of patient care; support the exchange of readable documents between people, also with different levels of technical knowledge; promote the maintenance over time of all the encoded pieces of information; compatibility with XML and HL7 RIM. In this paper, we specifically focus on HL7 v3. Its information structure is based on the object paradigm that is formalized by means of Unified Modeling Language (UML) diagrams.

```
{
    "observationEvent": {
        "id": {
            "_root": "2.16.840.1.113883.19.1122.4",
            "_extension": "1045813",
            "_assigningAuthorityName":
                            "GHH LAB Filler Orders"
        },
        "code": {
            "_code": "1554-5",
            "_codeSystemName": "LN",
            "_codeSystem": "2.16.840.1.113883.6.1",
            "_displayName": "GLUCOSE POST 12H
                CFST:MCNC:PT:SER/PLAS:QN"
        },
        "statusCode": {
            "_code": "completed"
        },
        "effectiveTime": {
            "_value": "200202150730"
        },
        "priorityCode": {
            "_code": "R"
        },
        "confidentialityCode": {
            "_code": "N",
            "_codeSystem": "2.16.840.1.113883.5.25"
        },
        "value": {
            "_xsi_type": "PQ",
            "_value": "120",
            "_unit": "mg/dL"
        },
        "interpretationCode": [
            {
                "_code": "H"
            },
            {
                "_code": "N"
            }
        ],
        "referenceRange": {
            "interpretationRange": {
                "value": {
                    "low": {
                        "_value": "70",
                        "_unit": "mg/dL"
                    },
                    "high": {
                        "_value": "105",
                        "_unit": "mg/dL"
                    },
                    "_xsi_type": "IVL_PQ"
                }
            }
        }
    }
}
```

Listing 2: Example of Content Domain in JSON format.

This means that any piece of information represents an object that has its own attributes and that maintains relationships with other classes (or other types of information). This data model is known in literature as HL7 Reference Information

Model (RIM) and it includes six classes: Entity, Act, Role, Participation, ActRelationship and RoleLink.

B. Case of Study: Observation on the Level of Glucose

In HL7 v3 the fundamental aspect of messaging, is given by iteration. An iteration is a unique association between types of messages (information transfer), a particular trigger event that starts the exchange (the cause of the transfer) and the responsibility of the recipient. *the transmission of the result of the glucose value in blood equal to 182 mg/dL, authorized by Dr. Howard H. Hippocrates, requered by Dr. Patricia Primary for the patient Eve E. Everywoman. The hospital is called Good Health Hospital (GHH)*. The HL7 v3 document has the classical structure of an XML file and it includes three basic parts: Transmission Wrapper, Trigger Event Control Act Wrapper, and Domain Content. The Transmission Wrapper includes pieces of information regarding the transmission of the HL7 document specifying data including sender and receiver, Trigger Event Control Act Wrapper specifies information regarding who triggered the event. Domain content describes the medical performance itself. Listings 1 shows an example of Domain Content regarding an observation on the Level of glucose in the blood codified in HL7 v3 XML format. The pieces of data contained within it describes the type of medical performance, its identification code, the performance status and the resulting value of the observation (i.e., 182) with the unit of measure (i.e., mg/dL). Other pieces of information regards the normal thresholds within the value should be included (i.e, high (H), low (L) or regular (R)). How described in the previous Section, in order to maintain the interoperability among HIS(s) based on HL7 v3 (that uses the XML format) and in order to use an Archival Storage implemented by means of MongoDB able to manage big data, a parsing operation of HL7 documents from XML to JSON is needed by the Ingest macro-component. Considering the aforementioned code example, Listings 2 shows the corresponding Domain Content parsed in JSON format. Listings 3 shows the Java code used to parse HL7 document form XML to JSON.

```
public String toJSON (String xmlText) {
int PRETTY_PRINT_INDENT_FACTOR = 4;
String TEST_XML_STRING = HL7XML;
try {
JSONObject xmlJSONObj = XML.toJSONObject(xmlString);
return xmlJSONObj.toString(
    PRETTY_PRINT_INDENT_FACTOR);
}
catch (JSONException je) {
System.out.println(je.toString());
return null;
}
}
```

Listing 3: Code to parse HL7 documents from XML to JSON format.

VI. Experiments

In this Section, we discuss the performances of two Archival Storage prototypes: one based on MySQL and on based on MongoDB. In particular, using the glucose observation case of study that we discussed in the previous Section, we simulated 6 datasets respectively including 2400, 4800, 9600, 192000, 384000, 769000 HL7 documents in XML format, each one including a random value of glucose belonging to the range [40, 500]. In fact, we created a skeleton XML HL7 v3 document and we generated random values belonging to the aforementioned range. For example, if we consider a hospital with roughly 200 glucose examinations every day, we obtain 769000 clinical reports in roughly 10 years. More specifically, we inserted the same datasets in both Archival Storages after proper parsing tasks. The two Archival Storage prototypes were deployed in blade with CPU Dual-Core AMD Opteron(tm) Processor 2218 HE, RAM 6GB, OS: ubuntu server 12.04.2 LTS 64 BIT. For reason of page limits, we specifically focus on the MongoDB-based Archival Storage prototype.

```
public void insertJSON(String strJSON) {
MongoClient mongo;
DB database;
DBCollection collection;
BasicDBObject obj;
try{
  mongo=new MongoClient();
  database=mongo.getDB("clinical_reports");
  collection=database.getCollection("
      glucose_observetions");
  obj=(BasicDBObject)= JSON.parse(strJSON);
  collection.insert(obj);
}
catch(UnknownHostException ex){
  ex.printStackTrace();
}
```

Listing 4: Example of Java code able to insert JSON HL7 data in MongoDB.

Listings 5 shows the Java code used to extract data considering a particular range of glucose values.

```
DBCollection collection =
        db.getCollection("Observation");
BasicDBObject andQuery = new BasicDBObject();
ArrayList<BasicDBObject> obj = new ArrayList<
    BasicDBObject>();
obj.add(new BasicDBObject("observationEvent.value",
    new BasicDBObject("$gte",min).append("$lte", max
    )));
BasicDBObject fields = new BasicDBObject();
fields.put("observationEvent.id.root", 1);
DBCursor cursor = collection.find(fields);
try{
  if(cursor.hasNext())
    System.out.println(cursor.next());
}
catch(Exception e){
  System.out.println("Error.");
}
finally{
  cursor.close();
}
}
```

Listing 5: Example of Java code able to retrieve HL7 data from MongoDB.

Listings 4 show the Java code used in the Ingest subcomponent used to parse XML HL7 v3 documents in JSON format to insert data in the MongoDB-based prototype. After that we populated the two databases with the same datasets, we executed query tests to evaluate the time taken to extract data considering the instant when we submitted the request and the instant when we get results. For each dataset, we

executed respectively the same query 30 consecutive times on MySQL and 31 times on MongoDB, so as to obtain mean execution times and corresponding confidence intervals at 95%. In the case of MongoDB, we run the same query 31 times because, for performance optimization, when the same query is executed on the same dataset, MongoDB processes all actual data only once, putting results in cache. In this way, if the same query is performed on the same dataset many times, in order to speed up the retrieval task, it will consider queries on cached data from the second execution on. Only after a data update MongoDB executes the queries again on actual data. Figure 4 shows the hinstogram comparing the execution times of queries performed on MongoDB and MySQL Archival Storages considering the same datasets and the same range of glucose values. In particular, for each

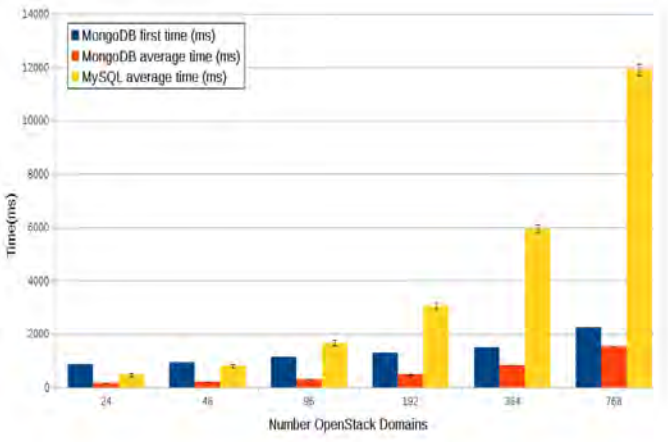


Fig. 4: Query execution time in MySQL and MongoDB.

dataset, we reported the execution time in milliseconds of the first query in MongoDB, the meam execution time of the repeated queries in MongoDB, and the meam execution time of queries in MySQL. How we can observe, for each dataset, the first query execution time in MongoDB is always grater than the mean query execution times of the following 30 repeated queries. In general, we can observe that, considering cached data, MongoDB has always better performances than MySQL. For small data, i.e., considering datasets of 24000 and 48000 HL7 documents, we can observe that MySQL has slighlty better query execution times than the first query execution time in MongoDB. For datasets of 96000, 192000, 384000 and 768000 HL7 documents, the query execution time in MySQL is always grater than the query execution time in MongoDB (considering both the first query execution time and the mean value on the following 30 ones) and this gap becomes more and more evident increaseing the numner of HL7 documents. Moreover, in MySQL we can observe confidence intervals larger than MongoDB. These experiments proved that MongoDB is a good choice for the implementation of Archival Storage sub-component of an OAIS-based HIS able to manage big data.

VII. Conclusion and Future Work

In this paper, we discussed a OAIS-based HIS able to manage big data in a Cloud computing context. In particular, considering the HL7 v3 standard we discussed two possible implementations of Archival Storage sub-components respectively based on MySQL and MongoDB. MySQL is one of the major DBMS currently used in HIS based on OAIS architectures, whereas MongoDB is one of the major emerging solutions of big data management systems. In order to highligt the data parsing tasks required by the Ingest sub-component, we considered a use case of HL7 clinical reports regarding glucose observations. Experiments proved that MongoDB is a good choice for the implementation of Archival Storage sub-component able to manage big data. In future work, we plan to focus on inter-operable federated OAIS-based HIS architectures.

Acknowledgment

This research work has received funding from the European Union's Seventh Framework Programme Project frontierCities under grant agreement no. 632853.

References

[1] Regulation (EU) No 282/2014 of the European Parliament and of the Council of 11 March 2014 on the establishment of a third Programme for the Union's action in the field of health (2014-2020), http://ec.europa.eu/health/programme.

[2] S. M. P. Chinniah, "Hl7 standard based hospital information system," *International Journal of BioSciences & Technology*, vol. 5, pp. 32–39, 2012.

[3] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, 2014.

[4] J. J. Seddon and W. L. Currie, "Cloud computing and trans-border health data: Unpacking u.s. and {EU} healthcare regulation and compliance," *Health Policy and Technology*, vol. 2, no. 4, pp. 229 – 241, 2013.

[5] A. Bahga and V. Madisetti, "Healthcare data integration and informatics in the cloud," *Computer*, vol. 48, no. 2, pp. 50–57, Feb 2015.

[6] F. Zhang, J. Cao, S. U. Khan, K. Li, and K. Hwang, "A task-level adaptive mapreduce framework for real-time streaming data in healthcare applications," *Future Generation Computer Systems*, vol. 4344, pp. 149 – 160, 2015.

[7] P. He, P. Wang, J. Gao, and B. Tang, "City-wide smart healthcare appointment systems based on cloud data virtualization paas," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 2, pp. 371–382, 2015.

[8] R. Wadhwa, A. Mehra, P. Singh, and M. Singh, "A pub/sub based architecture to support public healthcare data exchange," in *Communication Systems and Networks (COMSNETS), 2015 7th International Conference on*, Jan 2015.

[9] A. Castiglione, R. Pizzolante, A. D. Santis, B. Carpentieri, A. Castiglione, and F. Palmieri, "Cloud-based adaptive compression and secure management services for 3d healthcare data," *Future Generation Computer Systems*, vol. 4344, pp. 120 – 134, 2015.

[10] Y. Zhang, M. Qiu, C.-W. Tsai, M. Hassan, and A. Alamri, "Healthcps: Healthcare cyber-physical system assisted by cloud and big data," *Systems Journal, IEEE*, vol. PP, no. 99, pp. 1–8, 2015.

[11] W. Wang, L. Chen, and Q. Zhang, "Outsourcing high-dimensional healthcare data to cloud with personalized privacy preservation," *Computer Networks*, vol. 88, pp. 136–148, 2015.

[12] B. Fabian, T. Ermakova, and P. Junghanns, "Collaborative and secure sharing of healthcare data in multi-clouds," *Information Systems*, vol. 48, pp. 132 – 150, 2015.

[13] HL7 Version 3 Product Suite, http://www.hl7.org/implement/standards/product_brief.cfm?product_id=186.

Recognition and Adjustment for Strip Background Baseline in Fluorescence Immuno-chromatographic Detection System

GaoYueming[1] Lin Chuanyang[2] Pun Siohang[3] VaiMangi[4] Du Min[5]

[1,2,5]*College of Physics and Information Engineering, Fuzhou University, Fuzhou, China*

[1,2,5]*Key Laboratory of Medical Instrumentation and Pharmaceutical Technology of Fujian Province, Fuzhou, China*

[3,4]*Department of Electrical and Computer Engineering, Faculty of Science and Technology, University of Macau, Macao*

[1]*fzugym@163.com,* [2]*n131127011@fzu.edu.cn,* [3]*lodge@mail.eee.umac.mo,* [4]*fstmiv@umac.mo,* [5]*dm_dj90@163.com*

Abstract-Fluorescence immune-chromatography is a kind of immunoassay which is widely used in quantitative detection area. However, the background baseline drift with relatively large amplitude during test is a huge problem for the overall performance of the quantitative detection system, and the existing background baseline recognition and adjustment methods were based on online or offline data processing, which brings huge computational burden. To solve this problem, this paper proposed a background baseline recognition and adjustment method from the aspect of optical design. It integrated background acquisition, conditioning and control functions which could adjust the photo-electric signal of the background in real time until the sampling value of background baseline was within the threshold. Finally, the peak value of the test strip was detected. To examine the performance of the system, ten standard test strips with different concentrations were selected for background baseline detection. Besides, linearity and coefficient of variation (CV) of the system was also calculated. Results showed that the fluctuation range of baseline was within±4% of the threshold, whereas the CV of the test strip background was 2.57%. Furthermore, although the occupancy rate of background amplitude for A/D unit was decreased from 31.7% to 2.4%, CV of the system was still lower than 3%, which indicates that the system has much wider detection range than other congeneric detection devices.

***Keywords*：Immuno-fluorescence, background baseline, adjustment method.**

I. INTRODUCTION

Fluorescence immune-chromatography is a kind of immunoassay based on chromatographic and antigen-antibody specific reaction techniques [1-3]. It utilizes the fluorescence particles to capture the biomarkers in the tested samples, then, through irradiation of excitation light at certain wavelength, the biological signal can be rapidly detected. With some attracting features, e.g., high specificity, ease of operations, and single serving detection [4-5], fluorescence immune-chromatography is widely used in quantitative detection area. Furthermore, it is greatly suitable for the development trend of (Point-of-Care) POC test for its high sensitivity and wide detection range [6-7].

In general, a good recognition and adjustment for strip background baseline plays a significant role in the overall performance of the quantitative detection system. In literatures, [8] introduced a baseline fitting method to reduce and stabilize the background amplitude, and the fitted curve was selected as the baseline of the background. In [9], the (coefficient of variation) CV of the system was less than 5%. This work searched the position of the inflection point of the characteristic curve by second-order differential method, and regarded the value of the inflection point as background baseline value. In our previous researches on fluorescence immunoassay analyzer system [10, 11], background baseline drift with relatively large amplitude during test was such a huge problem. In [10], mean value of continuous sampling at the start position of the test strip was extracted as the background baseline. Whereas [11] implemented a wavelet transform algorithm for background baseline determination.

Certainly, aforementioned background baseline recognition and adjustment methods are based on online or offline data processing, which brings huge computational burden and challenges to test equipment on real time field. To solve this problem, this work proposed a background baseline adjustment method from the aspect of optical design. The developed baseline adjustment method integrated background acquisition, conditioning and control functions, which improved the stability of the background baseline and reduced the value of background sampling amplitudes. Furthermore, it expanded the linear detection range of the system and enhanced the accuracy of test results without compromising any detection sensitivity.

The rest of this paper is organized as follows. Section II presents the methodology of the proposed system followed by the background discrimination and

processing algorithm in Section III. Section IV explains the result and discussion. Finally, the conclusions are drawn in Section V.

II. METHODOLOGY

A. Fluorescence Immune-chromatographic Strip

Basically, a fluorescence immuno-chromatographic test strip consists of three sections including sample, reaction and absorbent pad. As shown in Fig. 1. In sample pad, tested samples with different concentrations are dropped. Beside sample pad, fluorescence particles are mobilized as sample markers in conjugate pad. The absorbent pad, which is located at the other end of the testing strip, creates capillary action for the tested samples. As the tested sample wick from sample pad to absorbent pad, there is reaction pad in between which is composed of T line (test line) and C line (control line), where reactions occur and test result is shown [12].

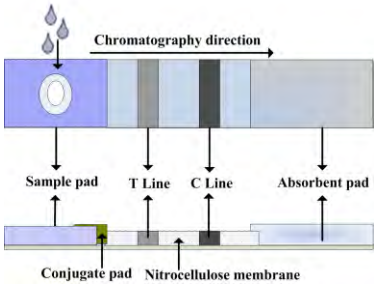


Fig. 1. Schematic diagrams the fluorescence immuno-chromatographic strip

B. Composition of Background Detecting System

The proposed background detecting system was mainly composed of three modules, including optical measurement module, background acquisition module and background adjustment module, which will be illustrated in details in this section.

1) Optical Measurement Module

To read test results from the test strip, an optical measurement module was designed. It included UV LED, several lenses, and filter, as shown in Fig. 2. There were two convex (Lens A and B), one cylindrical lens and one dichroic lens. When a test started, UV LED was opened to emit ultraviolet (UV) light for excitation. Through Lens A, the scattered UV light was adjusted to parallel light. The dichroic lens was placed in a certain angle to reflect the UV light. As a result, the UV light incidented into the cylindrical lens vertically, which was used to focus the light into a beam on the reaction pad of the test strip. Then, the fluorescence particles on the reaction pad reemitted light which was proportional to the sample concentration. Through the dichroic lens and filter, useless light was filtered and finally the reemit light focused on a photodiode for opto-electric conversion. After that, a stepper motor drived the test strip by stepwise pulling-in and pulling-out to scan the length axis of the strip.

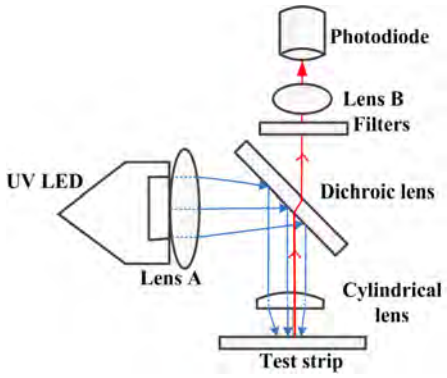


Fig. 2. The structure of optical measurement module

As shown, T and C line on the test strip (Fig. 1) were rectangular shape. In this regard, the excitation light focused on the test strip was designed as rectangular shape beam to make the best use of the optical system. By precise calculation of the optical path parameters, the area of the rectangular shape beam was, S_{beam}=3.0mm×1.5mm.

2) Background Acquisition Module

The background acquisition module was a transresistance amplifier circuit, which was used to convert background current (nA) to voltage signal (mV). As shown in Fig. 3, the main component of this circuit was an operational amplifier (OPA657). Besides, high and low frequency noise existed in circuit was filtered by a band-pass filter [13]. At this circuit, internal 12 bit A/D module of the processor performs the background signal sampling and the reference voltage was 3.3 V. It was recommended that the input analog signal of the A/D module should be higher than half of reference voltage to make a better use of the A/D resources. Therefore, in the post stage of this circuit, a two-phase scaling circuit with OPA657 was implemented to amplify the background voltage from mV level to V level.

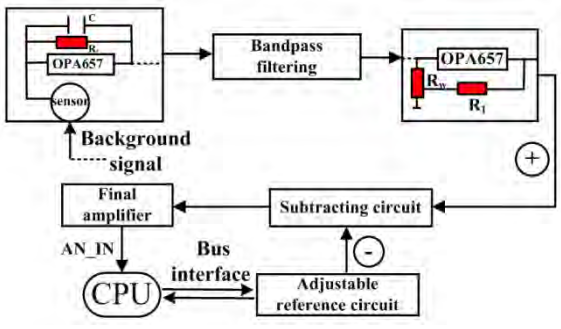


Fig. 3. The block diagram of system background processing

3) Background Adjustment Module

The background adjustment module was mainly composed of two minor modules — background adjusting module and control module. On one hand, the background adjusting module included an adjustable

reference voltage unit, a subtraction unit and an output amplifier unit. There were three main components in adjustable reference voltage unit, which were voltage regulator, digital potentiometers and voltage divider resistors. The voltage regulator generated stable reference voltage which can be coarsely or finely adjusted by two digital potentiometers. The whole procedure was controlled by a processor to achieve self-adaptive resistance adjustment.

On the other hand, a 32-bit ARM processor (STM32) was implemented in the control module to control the output voltage of the background adjusting module. First, A/D conversion unit of the processor sampled the output voltage of the background adjusting module, following the digital signal filter. Then, after processing the test result by a specific algorithm, the baseline estimation value and the threshold value were obtained. If the difference between baseline estimation value and standard baseline value was larger or lower than the threshold value, a feedback system was triggered to adjust the baseline value. Finally, when the absolute difference between final baseline estimation value and standard baseline value converged to the threshold value, a test strip detection starts.

III. Background Discrimination and Processing Algorithm

A. Discrimination of the Background Area

A proper discrimination of the test strip background area was crucial for a good detection. When the stepper motor drived the test strip to scan the length axis of the strip, the fluorescence signal distribution of the test strip collected from the optical module was obtained, as shown in Fig. 4. L1 and L4 region was the upper and lower edge of the scanning window, where the reflection coefficient between the edge of the strip shell and the background area of the test strip was different. As a result, there was a huge background signal drift. In this regard, L1 and L4 should not be selected as background area. However, in L3 region, most fluorescence signals distributed in T line (test line) and C line (control line). Certainly, in this paper, L2 region was chosen as background area.

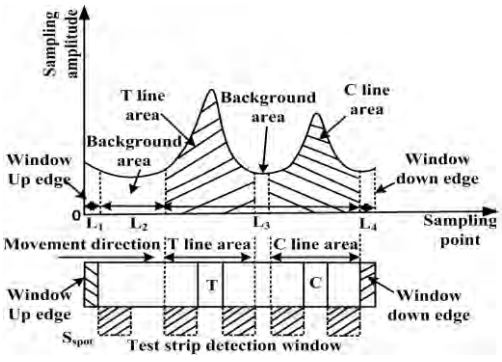


Fig. 4. Fluorescence signal sampling area curve allocation map

B. Data Processing Algorithm

During a test, feature quantity detection algorithm of the system was based on the background baseline. However, with large system noise, the background baseline value was unstable in each detection. Therefore, before operating feature quantity detection algorithm, it was necessary to estimate the baseline value according to the estimation theory. In this feature quantity detection algorithm, the background baseline was detected n times at the beginning of each detection. From least squares estimation theory, arithmetic average of the detected results was the least squares estimate value.

IV. Result and Discussions

To examine the performance of the background baseline adjustment system, ten standard test strips with different concentrations were selected for background baseline detection. Test results were listed in Table I, in which the first column referred to the test strips number followed by second and third column with background baseline value before adjustment and after adjustment respectively, whereas the threshold value was listed in fourth column. As shown, the baseline value of standard strip was stable at the level of 100 which was stored in the control unit, as the standard of adjustment baseline. At a range from 63 to 336, there was a huge background baseline drift, and the difference between background baseline and standard strip baseline intensely exceeded the threshold value. However, after adjustment, strip background baseline fluctuation dramatically decreased. From 1 to 7, baseline drift fluctuation range was less than 10 % of the results before adjustment. Furthermore, the difference between background baseline after adjustment and standard strip baseline was within the threshold value. Therefore, the stability of the background baseline after adjustment was dramatically improved.

TABLE I
Fluorescent Strip Background Detection of Different Schemes

Strip	Before Adjustment	After Adjustment	Threshold
1	1256	102	5.10
2	1124	103	5.10
3	1232	101	5.20
4	1420	96	5.10
5	1187	98	5.10
6	1236	103	5.00
7	1280	99	5.20
8	1460	102	5.00
9	1376	102	5.10
10	1416	97	5.10

To analyze the linearity of the system, test had been conducted with different samples of different concentrations in five standards between 60ng/mL, 180ng/mL, 540ng/mL, 1800ng/mL, and 3600ng/mL. A high performance lateral flow system (ESEQuant, Qiagen, Germany) was prepared as comparison test. The calibration of V_T/V_C – concentration curve was

shown in Fig. 5. Certainly, the linearity of the calibration curve of the proposed system (R^2=0.99826) was better than ESEQuant system (R^2=0.99645). Furthermore, after recognition and adjustment processing of the background baseline, the proposed system had a wider detection range than ESEQuant system.

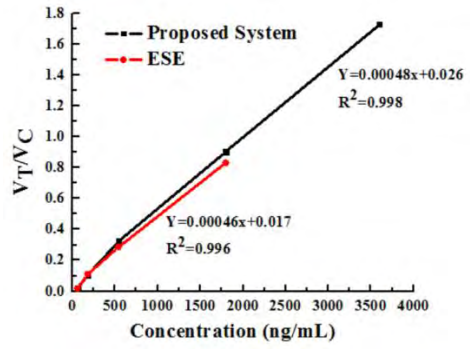


Fig. 5. Discrete strip data distribution Calibration curve of fluorescence concentration and characteristic

Besides, coefficient of variation (CV) had been calculated as for each strip was repeatedly conducted for 10 times. As listed in Table II, the proposed system could confirm stable detections with CV less than 3% in different concentrations. However, CV of ESEQuant system was lower in low concentrations, which indicated that ESEQuant system had better sensitivity. In general, this recognition and adjustment method could dramatically improve the performance of the fluorescence immuno-chromatographic detection system and was very competent against other existing techniques.

TABLE II
COMPARISON OF INSTRUMENT CV OF THE DIFFERENT CONCENTRATIONS STRIP

Concentration (ng/mL)	The proposed instrument	ESE
60	2.54 %	1.17 %
180	2.54 %	1.09 %
540	1.19 %	1.03 %
1800	0.37 %	0.83 %
3600	0.22 %	—

V. CONCLUSIONS

This work proposed a background baseline recognition and adjustment method from the aspect of optical design. It integrated background acquisition, conditioning and control functions which could adjust the photo-electric signal of the background in real time until the sampling value of background baseline was within the threshold. In the detection results of the system, the fluctuation range of baseline was within ±4% of the threshold, whereas the CV of the test strip background was 2.57%. Furthermore, although the occupancy rate of background amplitude for A/D unit was decreased from 31.7% to 2.4%, CV of the system was still lower than 3%, which indicated that the system has much wider detection range than other congeneric detection devices. In general, this recognition and adjustment method could dramatically improve the performance of the fluorescence immuno-chromatographic detection system and was very competent against other existing techniques.

REFERENCES

[1] RAPHAELC, HARLEY Y. Lateral flow immunoassay [M]. Humana Press, 2008.

[2] SUMONPHAN E, AUEPHANWIRIYAKUL S. Interpretation of nevirapine concentration from immuno- chromatographic strip test using support vector regression [J]. Proceedings of 2008 IEEE International Conference on Mechatronics and Automation, 2008: 633-637.

[3] LIN R, SKANDARAJAH A, GERVER R, etal.. A lateral electrophoretic flow diagnostic assay. [J]. Lab on a Chip. 2015.

[4] PYOD, YOOJ. NewTrends in Fluorescence Immuno-chromatography [J]. Journal of Immunoassay and Immuno-chemistry, 2012，33(2): 203-222.

[5] GUI C, DAI X, WANG K, etal.. Design gold immuno-chromatography detection devices based on CCD [J]. Chinese Journal of Biomedical Engineering. 2015(2)：237-242.

[6] LAWRENCE T, CHARLES C H. Use of point-of-care testing for plasma therapy [J]. Transfusion, 2012, 52(5): 56-64.

[7] VASHIST S, LUPPA P, YEO L, etal.. Emerging Technologies for Next-Generation Point-of-Care Testing. [J]. Trends Biotechnol. 2015.

[8] LI X T. Quantitative fluorescence immunoassay test strip detector design and development [D]. Zhejiang University,2014.

[9] XIAO H. Immuno-fluorescence detector hardware design and implementation [D]. Hangzhou University of Electronic Science and Technology, 2014.

[10] LIU X, DU M, LI Y R, etal.. Design of fluorescence Immuno-chromatographic quantitative detection system [J]. Journal of Electronic Measurement and Instrument, 2013, 27(9): 859-866.

[11] GAO Y M, LI TQ, LIN C Y, etal.. Optical signal processing and feature amount immune-fluorescence strip test select [J]. Electronic Measurement and Instrument, 2015, (5): 662-668.

[12] MENDONCA M，CONRAD N L，CONCEICAO F R，et al.. Highly specific fiber optic immune-sensor coupled with immune-magnetic separation for detection of low levels of Listeria monocytogenes and L. ivanovii [J]. BMC microbiology, 2012,12(1): 275-277.

[13] ZHANG N, ZHANG R B, XU P F, etal.. Immuno agglutination photoelectric detection methods latex micro fluidic chips [J]. Chinese Journal of Scientific Instrument. 2013, 06: 196-201.

[14] HAI Q, CHENG W H. Least Square Generalized Pareto Distribution Parameter Estimation [J]. Applied Probability and Statistics. 2013 (2).

[15] DU X H, YIN Y N, ZUO Z H, et al.. Establishment of Dairy Cattle GPR109A Gene SYBR Green I real-time PCR detection method [J]. South Agricultural Sciences, 2015 (3).

Agile Development of a Hospital Information System

S. L. R. Vrhovec*,**

* University of Maribor, Faculty of Criminal Justice and Security, Ljubljana, Slovenia
** University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
simon.vrhovec@fvv.uni-mb.si

Abstract - Agile software development methods have rapidly spread after their formal introduction in the agile manifesto at the dawn of the millennium. They rapidly gained support in the software industry and can be considered a standard today. Agile methods seem to improve software project success rates and offer developers the needed flexibility to adapt to changing user requirements. However, the use of agile methods in large-scale information systems development projects has been rarely studied. There are even fewer insights in the area of agile healthcare information systems development. In this paper, we present insights into an agile hospital information system development in a European hospital. Agile development has been studied from various stakeholder perspectives. Results show that physicians, nurses and administration have diverse opinions on agile practices and different practices may be appropriate for interacting with each stakeholder.

I. INTRODUCTION

Agile software development methods have rapidly spread after their introduction in the late 1990s. They emerged as an answer to the notoriously low share of successful software projects using traditional software development methods [1]. They rapidly gained support in the software industry and can be considered a standard for smaller software projects today. Agile methods seem to improve software project success rates which are slowly but steadily increasing over the recent years and offer developers the needed flexibility to adapt to changing user requirements [2], [3]. Nevertheless, the success rates of larger software projects remain alarmingly high as only 10 percent of larger software projects is successful and the next 52 percent is over budget, over time, or with less functionality than originally planned [2].

Traditional methods still prevail in business environments where software projects are large and complex [4], [5]. Nevertheless, organizations seek how to include elements of agility into larger software projects in order to improve their success [4], [5]. Only few attempts at this can be found in the literature usually based on crossing over agile and traditional methods [5], [6]. The research on the use of agile methods in healthcare where software projects are typically larger has been particularly scarce.

In this paper, we report on the agile development of a hospital information system in order to fill in the existing gap in the literature. We first review the literature on agile methods and present the key characteristics of the Scrum agile method. Then, we report on the use of the agile method as seen by different software users in a hospital – physicians, nurses, and employees in the hospital administration. We conclude the paper with a discussion of the findings and some concluding remarks.

II. AGILE METHODS

Agile methods emerged as an answer to traditional plan-driven methods, such as waterfall development, incremental, prototyping, iterative development, and their variations and hybrids. Traditional methods are characterized with upfront planning, i.e., upfront agreement on requirements, reporting, and following a detailed plan [7], [8]. The key issue with traditional methods is that uncertainty is at its highest in the project planning phase making particularly difficult to make a plan that can be followed throughout the whole project [9]–[11]. Agile methods in contrast promote more flexible planning. Some planning may be done during the course of the project enabling easier adaptation to changing requirements which is achieved with a high degree of interaction and cooperation with the client [7], [12].

Agility is a constant readiness of the software development method to quickly and inherently creates changes, proactively or reactively accepts changes and learns from them while at the same time contributes to the perceived added value for the customer through its collective components and relationships with the environment [12]. To better understand agility, agile manifesto defines the contrast between traditional and agile methods [1], [8], [13]:

- Individuals and interactions over processes and tools.
- Working software over comprehensive documentation.
- Customer collaboration over contract negotiation.
- Responding to change over following a plan.

The agile manifesto includes a side note – even if there is value in the items on the right, agile methods value items on the left more. The manifesto therefore does not advocate the ignorance of development processes, tools,

documentation, contracts and plans as often promoted and seen in practice. Such approach to software development does not pay off in the long term, e.g., due to significantly higher costs of maintenance.

One of the agile methods' key characteristics is frequent delivery of working code typically as a result of frequent short iterations [8], [14]. This is an effective mechanism for getting regular feedback on developed software during the project in contrast to traditional methods where the users interact with the developed software only at the end of the project [8], [14]. Regular user feedback enables the adaptation of upcoming iteration plans. Frequent and short iterations pose a challenge as how to align the goals of individual iterations with the goals of the project as a whole [14]. The project management therefore needs to be aware of the project goals and ensures that the partial goals are aligned with them.

Agile methods also emphasize informal communication which should be direct and personal whenever possible [15]. Informal communication may take place outside formal reporting structures of the software project and without management knowing about it [15], [16]. In this way, it is possible to quickly troubleshoot and fix errors and gather any needed details [15]. Visual models and products, such as user interface prototypes, may prove very useful in enriching the communication between software developers and users as they have technical added value for the developers as well as they can be clearly understood by users [17]. The frequent use of informal communication causes a significant reduction of the role of project management in decision making and increase in the involvement of the project team and customer into it [14]. For example, the customer can set or significantly influence the priorities of functionalities to be developed [1].

At first, agile methods were deemed appropriate for smaller software projects as they advocate lightweight software development [5]. In business environments where software projects are large and complex the traditional methods still prevail [4], [5]. Organizations nevertheless seek ways to include elements of agility into larger software projects in order to improve responsiveness to changes, shorten the time needed for the introduction of new products to the market, and reduce overhead costs [4], [5]. Some attempts at addressing this challenge can be identified based on crossing over agile and traditional methods [5], [6].

Each agile method is composed of specific practices that focus on different aspects of simplification of the software development process [12], [18]. Among the most widespread and researched agile methods, Scrum focuses on project management and Extreme Programming (XP) on development activities on the team level [1], [12], [19]. Other established methods that are considered agile are: Crystal (a family of agile methods that is composed of numerous methods and principles for their adaptation for specific projects), Feature Driven Development (division according to functionality), Dynamic Software Development Method (the first agile method), Adaptive Software Development (based on Rapid Application Development – RAD, establishing order from chaos), Unified Development Process (centered around software architecture, based on use cases), and Lean Software Development (based on principles of the Japan automobile industry) [7], [12], [13], [18], [20].

III. SCRUM

The popularity of Scrum sky rocked in recent years. This agile method is inspired by American football. It is simple, proven to be effective and compatible with other agile methods. More than a comprehensive software development method, it is a lightweight method for project management [6]. Scrum teams are smaller, self-organized and composed of best developers [1], [6]. The work that needs to be done in the project is split into smaller units, typically user stories, that are listed on the priority list, i.e., product backlog [6]. Sprints are iterations that typically last two to four weeks [1], [6]. At the beginning of the sprint, a certain amount of user stories is moved from the product backlog to the sprint backlog [1]. The amount of user stories may vary depending on their size and the capabilities of the project team [1]. At the end of each sprint, working software realizing user stories on the sprint backlog must be available for presentation to the customer or transfer to the production [1], [6]. At the end of the sprint, a sprint review takes place where customer feedback is considered as well as priorities on the product backlog may be updated [6]. The last step in the sprint is sprint retrospective which aims to optimize and improve the Scrum method itself as it is not precisely defined and leaves room for organizations to adapt it to their own environment [6].

IV. METHOD

The case study method was used as research approach due to several reasons. First, case study is the most suitable approach for investigating contemporary events with no control over the environment [21], [22]. Next, the complex interaction of different stakeholders and new software can be best analyzed with a case study [23], [24]. Finally, case study is particularly appropriate for studying software projects in natural organizational settings [25], [26].

The unit of analysis of the case study was an agile hospital information system (HIS) development project in a European hospital. The client was a hospital with approximately 350 employees in 13 departments, including 100 physicians, 200 nurses, and 50 employees in hospital administration. The developer was a European software company with approximately 100 employees. From this point of view, the settings can be considered as typical for healthcare. The HIS project was a key step toward the hospital going paperless. It started with a traditional method during which the first requirements were developed and architectural foundations were designed. However, the traditional method was switched with the agile method Scrum even before the first issue of working code. The software company decided to switch to an agile method due to their prior experience with HIS development and vast knowledge of the field. The project team has been active at international conferences and knows very well the relevant international standards. In

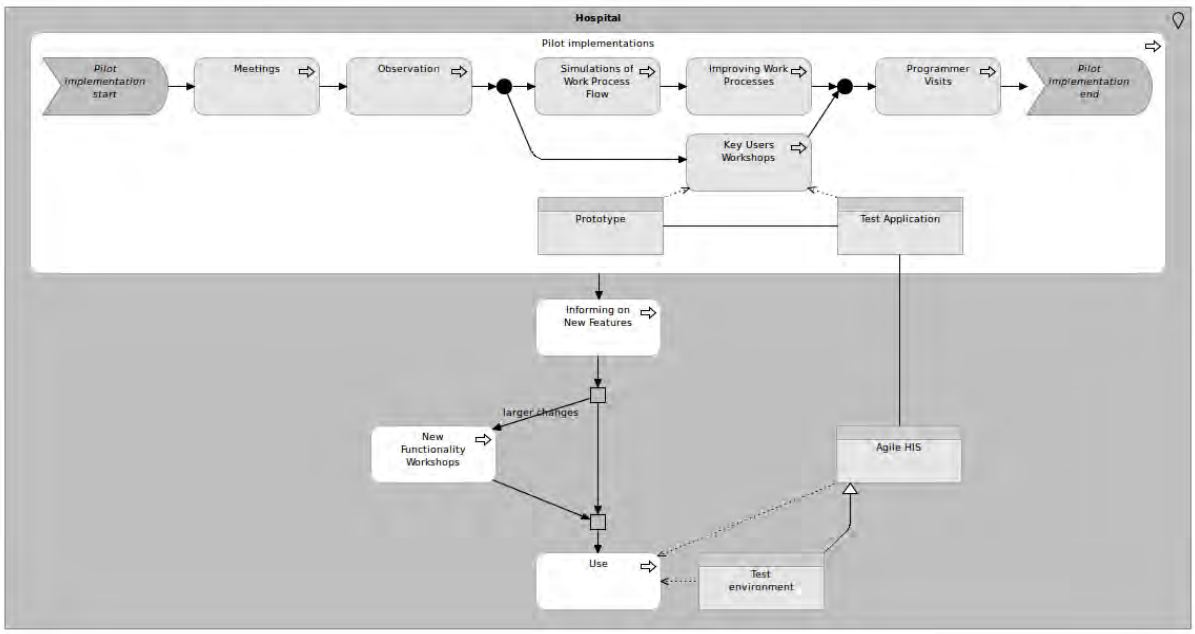


Figure 1. Relations between the interaction elements of the development method

addition, the architecture was well-defined and appropriate for agile development of a complex information system.

We identified three key stakeholders at the hospital: physicians, nurses, and administration. The project was managed by the software company using Scrum with a sprint length of two weeks. The project team had two major roles, i.e., business analysts and programmers. Business analysts were the bridge between the software company and the hospital. This is a deviation from the typical agile practices and is a consequence of healthcare specifics where users are overloaded with everyday work. Instead of the hospital assigning some key users to the project the software company dedicated a few business analysts to the hospital. The programmers had only few direct interactions with hospital employees.

Data triangulation has been used in this study to assure validity and to provide confirmatory data [27]. Various data types and data provided by different stakeholders were compared against each other to ensure consistency of the results [21], [27]. Multiple data collection methods were employed, i.e., open-ended interviews [28], surveys, focus groups [28], and by project documentation. Reliability was assured by following a case study protocol [22]. It included research questions, methods and procedures for data collection and data analysis guidelines.

TABLE I. STAKEHOLDER SATISFACTION WITH INTERACTION ELEMENTS

Interaction Element	Physicians	Nurses	Administration
Prototype	– –	– –	–
Test application	– –	–	–
Test environment	–	–	– –
Simulations of work process flow	– –	–	+ +
Observation	–	+	+
Key users workshops	+	+	+
Informing on new features	–	– –	+ +
New functionality workshops	+ +	+	– –
Improving work processes	+	+ +	+
Pilot implementations	+	–	+ +
Programmer visits	+ +	+ +	–
Meetings	+ +	+ +	– –

V. RESULTS

In the case study, we first identified the interaction elements of the development method. These are those methodology elements (i.e., processes and products) that hospital stakeholders interact with. We documented 12 interaction elements in the used agile method, comprising of 9 processes and 3 products. The relations between interaction elements are presented in Figure 1. The stakeholders' perceived usefulness of interaction elements sorted by increasing perceived usefulness can be seen in Table I. The descriptions of each interaction element along with the findings from the study are presented in the next subsections.

A. Prototype

Prototype is a user interface without the possibility to enter data. It is used by key users at key user workshops. Most users very rarely come into contact with this interaction element if at all. It is used exclusively very early in user interface development and does not provide any functionality, e.g., data entry. The users cannot understand what is the prototype supposed to do and therefore cannot offer any useful feedback.

B. Test Application

Test application is a simple user interface with the capability of entering data and is the next stage of user interface development after the prototype. Similar to the prototype, it is used by key users at key user workshops. The test application is an important tool for acquiring requirements however the users do not perceive it as such. This may be attributed to the fact that the test application eventually evolves to the final user interface which is much more visually appealing comparing to the quickly prepared test application.

C. Test Environment

An environment with the testing version of HIS in which new HIS functionality can be tested independently of the production HIS. Users are free to test new functionality in the test environment prior to the deployment to the production however it is rarely used. Everyday work does not leave users with enough time to test them peacefully therefore they would prefer training on real cases in the presence of someone who has already mastered the new functionalities.

D. Simulations of Work Process Flow

Simulations of work process flow of the physicians, the nurses, and the employees in the hospital administration takes place after observation done by business analysts. It serves as a confirmation of observations and a foundation for subsequent improvement of work processes. Administration sees simulations of work process flow as very useful as it is easy to provide good feedback while the physicians and nurses do not. Physicians do not perceive the simulations as useful for two reasons. First, they have no influence over the work processes as they are entirely in the hands of the heads of departments. Second, they fully trust the software company to be capable of making quality proposals based on best world practices from medical informatics to the heads of departments. Concerning the nurses, the business analysts do not see much merit in conducting these activities with them as it is hard to obtain good feedback from them. Instead, they prefer to rely more on the direct observation of their work.

E. Observation

Observation of physicians, nurses, and employees in the hospital administration at their everyday work is done by business analysts. Nurses and administration perceive observation as useful. Observation is among the most important activity for acquiring requirements however the physicians do not perceive it as useful because they do not see a connection between it and the end product.

F. Key Users Workshops

Workshops for key physicians, nurses, and employees in the hospital administration are conducted by business analysts. The workshops are intended for gathering requirements with the help of prototypes and test applications that are prepared based on meetings and observation. All users perceive key users workshops as useful. The participants of the workshops were frequently changing in the past. Various individuals have been invited to the workshops however relatively few stayed. Those who stayed are sufficiently motivated and have good relations with other participants. Heads of departments and others with broad overview over the work at the whole hospital regularly attended the workshops.

G. Informing on New Features

Informing on new features was usually done through email and included short instructions for use. Only administration perceived informing as useful as they use email much more regularly than physicians and nurses who are preoccupied with every day work. Business analysts perceive the informing as just a formality that cannot be avoided. They are convinced that it is not effective due to low email usage among users. However, this does not hold true for the administration who work predominantly on the computers.

Administration and physicians are informed directly on their personal email. The nurses are however informed only through the head nurse who then forwards the message to others either through email or newsletters. This results in different opinions about the timeliness of messages at different departments. At some departments, the nurses get the messages a few days before which they consider too soon as they forget about the new functionality in the following days. At other departments, the nurses complain that they always get informed about new functionalities too late, e.g., just a few hours or on the eve before its deployment.

H. New Functionality Workshops

Prior to larger changes, the business analysts conduct workshops for all users on new features being deployed. Physicians and nurses perceive the new functionality workshops as useful. The workshops help them to better understand the work of each other and relate it to their own. They complain only about the workshop dates as they often cannot attend them due to their everyday work. Due to a low participation of physicians and nurses, the business analysts do not perceive it as too valuable. The administration does not have issues with participating at the workshops however they perceive the workshops as a useless waste of time as only a small fraction is dedicated to them. The rest is dedicated to features that they have no interest in.

I. Improving Work Processes

Improving the work processes of physicians, nurses, and administration is conducted by business analysts. All users perceive it as useful. Physicians perceive this activity as useful because business analysts often propose how to apply the best world practices to their workplace. Business analysts however do not see this activity as a key one. They try to base improvements on best world practices instead of relying on the ideas and suggestions of users. It is even more obvious with the administration that has already prescribed and standardized work processes. This is sometimes perceived as arrogant by the administration who thus feel like they are not being treated in a way that they deserve to.

J. Pilot Implementations

Pilots are typically implemented at individual departments when functionality is intended for physicians and nurses or key users when it is intended for the administration. New functionality is first implemented as a pilot and then deployed for the whole hospital as a general rule. Physicians and administration see merits in this way of software development.

Nurses and some physicians have however differing opinions about this activity. This can be attributed to two key issues. First, the pilot implementations are conducted more often at some departments than others. The perceived usefulness of pilot implementations is therefore much lower at departments that do not participate often if at all. Second, the departments that have participated in pilot implementations perceive lower usefulness after the pilot implementation is over. During pilot implementations, i.e., when the software company dedicates to them, they are very pleased with it as they can see the results quickly. However, they are very unpleased after pilot implementations when the software company is much less accessible and the time to implementation significantly longer as they focus on other tasks. The administration does not have such problems as they are a part of a single department.

K. Programmer Visits

Occasionally, programmers visit users at the hospital when they encounter problems that the business analysts cannot solve themselves. Programmer visits are seen as useful by physicians and nurses as the elimination of the intermediary business analysts usually enables better understanding of the situation by the programmer. Even though occasional visits by programmers are useful, the role of business analysts is valued by all involved stakeholders. In contrast to the programmers, they have extensive understanding of the domain as well as good communicating skills needed for acquiring requirements. Administration does not see much merit in this activity as most of their requirements are prescribed by law and hospital rules.

L. Meetings

Meetings between business analysts on one side and physicians, nurses, or administration on the other are a starting activity of acquiring requirements. Physicians and nurses perceive meetings as very useful however the administration does not. In latter's opinion, the business analysts have an inappropriate arrogant attitude during the meetings raising some dissatisfaction among them.

Business analysts do not consider the meetings as the most valuable activity for acquiring requirements as other activities provide much more valuable information. They nevertheless see value in them as a key activity for good relations with users and solving organizational issues.

VI. Discussion and Conclusion

The results of the study show that different stakeholders can have very different opinions on the same interaction elements. Analysis considering these differing perceptions may provide in-depth understanding of the situation and may significantly contribute to the preparation of adequate improvement measures. For example, it came to light that new functionality workshops were not appropriate for the administration and separate workshops should be prepared and tailored for them. The study also showed that the general beliefs do not always hold true. One of the biggest surprises for the software company arising from the results of this study was that pilot implementations were not perceived as useful by all users. They were convinced that physicians, nurses, and administration see them as one of the most valuable elements. However, the strategy of conducting pilot implementations and happenings after them arose some discontent among nurses and some physicians. Based on these findings we can claim that confronting different stakeholder perspectives can positively contribute to the development of adequate improvements of interaction elements.

The findings presented in this paper shed light into agile development of complex software in hospital settings. It contributes to better understanding of both the application of agile methods to larger software project and the use of agile methods in healthcare. Software development in healthcare is a special topic of high interest in research as it encompasses a problematic interaction between fundamentally very different stakeholders. This interaction and own specifics of individual stakeholders often results in resistance to change that may undermine the implementation efforts. However, with the present research we shed some light that may help implementers in healthcare better understand and prepare for similar undertakings.

References

[1] R. Hoda, J. Noble, and S. Marshall, "Self-Organizing Roles on Agile Software Development Teams," IEEE Trans. Softw. Eng., vol. 39, no. 3, pp. 422–444, Mar. 2013.

[2] The Standish Group, "CHAOS Manifesto 2013: Think Big, Act Small," 2013.

[3] K. El Emam and A. G. Koru, "A replicated survey of IT software project failures," IEEE Softw., vol. 25, no. 5, pp. 84–90, 2008.

[4] J. Pernstål, R. Feldt, and T. Gorschek, "The lean gap: A review of lean approaches to large-scale software systems development," J. Syst. Softw., vol. 86, no. 11, pp. 2797–2821, Nov. 2013.

[5] G. van Waardenburg and H. van Vliet, "When agile meets the enterprise," Inf. Softw. Technol., vol. 55, no. 12, pp. 2154–2171, Dec. 2013.

[6] D. Stankovic, V. Nikolic, M. Djordjevic, and D.-B. Cao, "A survey study of critical success factors in agile software projects in former Yugoslavia IT companies," J. Syst. Softw., vol. 86, no. 6, pp. 1663–1678, Jun. 2013.

[7] A. Tarhan and S. G. Yilmaz, "Systematic analyses and comparison of development performance and product quality of Incremental Process and Agile Process," Inf. Softw. Technol., vol. 56, no. 5, pp. 477–494, May 2014.

[8] S. Wood, G. Michaelides, and C. Thomson, "Successful extreme programming: Fidelity to the methodology or good teamworking?," Inf. Softw. Technol., vol. 55, no. 4, pp. 660–672, 2013.

[9] R. Atkinson, "Project management: cost, time and quality, two best guesses and a phenomenon, its time to accept other success criteria," Int. J. Proj. Manag., vol. 17, no. 6, pp. 337–342, 1999.

[10] K. de Bakker, A. Boonstra, and H. Wortmann, "Does risk management contribute to IT project success? A meta-analysis of empirical evidence," Int. J. Proj. Manag., vol. 28, no. 5, pp. 493–503, Jul. 2010.

[11] K. Jugdev and R. Müller, "A retrospective look at our evolving understanding of project success," Proj. Manag. J., vol. 34, no. 3, pp. 19–31, 2005.

[12] X. Yu and S. Petter, "Understanding agile software development practices using shared mental models theory," Inf. Softw. Technol., vol. 56, no. 8, pp. 911–921, Aug. 2014.

[13] B. Tessem, "Individual empowerment of agile and non-agile software developers in small teams," Inf. Softw. Technol., vol. 56, no. 8, pp. 873–889, Aug. 2014.

[14] M. L. Drury-Grogan, "Performance on agile teams: Relating iteration objectives and critical decisions to project management success factors," Inf. Softw. Technol., vol. 56, no. 5, pp. 506–515, May 2014.

[15] M. Korkala and F. Maurer, "Waste identification as the means for improving communication in globally distributed agile software development," J. Syst. Softw., vol. 95, no. 9, pp. 122–140, Apr. 2014.

[16] J. D. Herbsleb and A. Mockus, "An empirical study of speed and communication in globally distributed software development," IEEE Trans. Softw. Eng., vol. 29, no. 6, pp. 481–494, Jun. 2003.

[17] J. M. Rivero, J. Grigera, G. Rossi, E. Robles Luna, F. Montero, and M. Gaedke, "Mockup-Driven Development: Providing agile support for Model-Driven Web Engineering," Inf. Softw. Technol., vol. 56, no. 6, pp. 670–687, Jun. 2014.

[18] C. de O. Melo, D. S. Cruzes, F. Kon, and R. Conradi, "Interpretative case studies on agile team productivity and management," Inf. Softw. Technol., vol. 55, no. 2, pp. 412–427, Feb. 2013.

[19] J. Garzás and M. C. Paulk, "A case study of software process improvement with CMMI-DEV and Scrum in Spanish companies," J. Softw. Evol. Process, vol. 25, no. 12, pp. 1325–1333, 2013.

[20] B. Losada, M. Urretavizcaya, and I. Fernández-Castro, "A guide to agile development of interactive software with a „User Objectives'-driven methodology," Sci. Comput. Program., vol. 78, no. 11, pp. 2268–2281, Nov. 2013.

[21] K. de Bakker, A. Boonstra, and H. Wortmann, "Risk management affecting IS/IT project success through communicative action," Proj. Manag. J., vol. 42, no. 3, pp. 75–90, Apr. 2011.

[22] R. K. Yin, Case study research: design and methods, 4th ed. Thousand Oaks, California: Sage Publications, 2009.

[23] M. Trkman and P. Trkman, "Actors' misaligned interests to explain the low impact of an information system – A case study," Int. J. Inf. Manage., vol. 34, no. 2, pp. 296–307, Nov. 2014.

[24] D. Vavpotič and T. Hovelja, "Improving the evaluation of software development methodology adoption and its impact on enterprise performance," Comput. Sci. Inf. Syst., vol. 9, no. 1, pp. 165–187, 2012.

[25] J. P. Bansler and E. Havn, "Exploring the role of network effects in IT implementation: The case of knowledge repositories," Inf. Technol. People, vol. 17, no. 3, pp. 268–285, 2004.

[26] P. Darke, G. Shanks, and M. Broadbent, "Successfully completing case study research: combining rigour, relevance and pragmatism," Inf. Syst. J., vol. 8, no. 4, pp. 273–289, Oct. 1998.

[27] J. Molineux, "Enabling organizational cultural change using systemic strategic human resource management – a longitudinal case study," Int. J. Hum. Resour. Manag., vol. 24, no. 8, pp. 1588–1612, Apr. 2013.

[28] M. J. Ahn, O. Zwikael, and R. Bednarek, "Technological invention to product innovation: A project management approach," Int. J. Proj. Manag., vol. 28, no. 6, pp. 559–568, Aug. 2010.

SOA based interoperability component for healthcare information system

Danijel Kučak, Goran Đambić, Vjeran Kokanović
University College Algebra
Ilica 242, Zagreb, Croatia
danijel.kucak@algebra.hr, goran.djambic@algebra.hr, vjerch@gmail.com

Abstract - In this paper, a new SOA based interoperability component developed for the Healthcare Information System project of Special Ho0spital for Pulmonary Diseases in Croatia is proposed. Component has developed as standalone server which enables communication Healthcare Information System with other systems like Laboratory Information System, Radiology Information System, e-lists, e-ordering etc. Managing complex medical information as well as integrating heterogeneous system in Special Hospital for Pulmonary Diseases are solved by introducing HL7 and Web services. Finally, the performance of a communication between Healthcare Information System and other systems based on interoperability component is analysed.

1. INTRODUCTION

Special Hospital for Pulmonary Diseases (SHPD) owns about 100 beds and serves approximately 400 patients a day. The hospital uses multiple information systems to manage patients' demographic data, medical records, equipment, and facilities, as well as pharmaceutical events occurred daily within the hospital, for examples: outpatient registrations, emergency treatments, radiology, laboratory results, etc. In addition, the systems keep tracking routines, activities regarding auditing, billing and accounting with Croatian Health Care Fund.

These systems together form hospital's information (eco) system. Systems, such as the laboratory information system (LIS) or the radiology information system (RIS) are based around Microsoft stack (MS SQL Server, Microsoft .Net framework, Visual Basic 6 and Delphi). As time and technologies advancing, the digitized, multimedia medical data as well as electronic administrative office environment have been dramatically demanding in the recent years. While introducing new systems and the cutting edge technologies into the legacy healthcare systems, it becomes increasingly difficult.

Therefore, since May 2011, SHPD has begun to rebuild and restructure the system in order to provide efficient and integrated services for not only electronic, educational environment establishment but patients, staff and faculty members' usages as well. A modern middleware framework is designed to make our developed HIS available, scalable, robust, of high performance. In order to integrate different platforms and databases, a multi-tiered Service-Oriented Architecture (SOA) [3] is chosen while Health Level Seven (HL7) standard is wildly adapted to cover all data exchange [1]. Hence, the transition from the legacy system into the new one can be applied smoothly.

Furthermore, the proposed framework treats all services as independent modules that can be hot-swapped without interrupting the whole system. This enables our new systems to be deployed and configured in the highest degree of flexibility. In the following part of this paper, we will describe the requirement, design, deployment and performance of our middleware framework as well as completely new HIS.

2. SYSTEM REQUIREMENTS

The most important task of developing a modern HIS is to define and/or adapts architecture for exchanging various medical information and data across the system. In our HIS architecture design, several requirements have to be met while choosing and defining the protocols to be used in our system. First, our system is designed to be operated by all outpatient service staff including clinical doctors, nurses, pharmacists, administrative personnel and etc. Hence, dissimilar and deviated user interface would be needed. Moreover, information from different medical staffs could be quite varied. The software architecture must be able to cover and handle all information. Second, there are many existing medical information systems co-existed with the legacy HIS. These systems are not going to be phased out in a short term. So that, the new HIS must be able to communicate with these old systems, for example, the LIS and RIS.

Furthermore, there are also external systems under Croatian Health Care Fund VPN. The ability of exchanging data between our HIS and other systems is also a must. And last, although we have started developing our out-patient service system, the ultimate goal of the entire project is to design a new HIS that covers all healthcare services in including out-patient, in-patient and emergence services. Thus, the skeleton of our HIS has to be able easily extended or re-used for further development.

HL7 standard is chosen to carry all information between Web Servers and data exchange servers. HL7 is an ANSI standard widely used and adapted in many health institutions and medical instruments. Originally, HL7 was supposed to act as the seventh layer in Open System Interconnection (OSI) model. It means that HL7 was

designed to be directly transmitted over transporting layer protocol, for example: LLTCP/IP. Transmitting HL7 messages over TCP has it benefits, such as simple and easy to implement. But, for a large system like our HIS, using direct TCP connection means that we have to establish dedicated connection channels across the hospital network system. By doing this, we would have to open extra TCP ports for handling multiple and concurrent connections manually without using service bus. These may not only increase the complexity of our system design and the risk of being attacked but also lose the robustness and flexibility of the HIS system.

To address these concerns, we decide not to transmit HL7 message directly over TCP socket. Instead, SOAP (Simple Object Access Protocol) / WSDL (Web Service Description Language) Web services is introduced into our system as a proxy layer to our TCP/IP layer. SOAP/WSDL Web services, or in general, Web services, is mechanism enabling interoperation between software systems. By combining multiple W3C XML standards in different parts of system communication, Web service allows message or data to be exchanged across network and program process.

The most advantage of Web services is that there is almost no restriction of programming language for a system using Web services. Since Web services can be discovered, described and communicated by UDDI (Universal Description Discovery and Integration), WSDL and SOAP respectively [2], any platform which can generate these XMLs may have the ability to consume Web services. The other benefit of using Web services is that it consumes and provides exchange data by SOAP or HTTP GET/POST mechanism. Because SOAP itself is also a kind of content carried by HTTP, Web services can exchange data by normal HTTP channel. This means that by using Web services, data exchange can be achieved without additional network configurations. Because of these, Web services technology is adapted in many major middleware solution architectures such as Microsoft .NET Framework. Therefore, developing a Web services based system could be facilitated by well matured developing tools. With these advantages, Web service is chosen in our HIS as the key role of middleware. We decide not to transmit HL7 directly over TCP, instead, we exchange HL7 message through Web services. The solution architecture lets our system be benefited by the comprehensive and flexible definition of HL7 but leaves the sophisticated connection problems to be handled by Web services.

3. SYSTEM DESIGN

To realize our system, we mapped our architecture to 3 main segments (Figure. 1):

- External Network (CEZIH VPN)
- Proxy (HIS Web Service)
- Integration Server with HL7 Message Library

A. External network (CEZIH VPN)

CEZIH stands for Central Croatian Health Care Information System. Every system in organization that needs to communicate with CEZIH must have granted access from Croatian Health Care Fund. Communication is achieved through VPN client applications or by using IPSec routes on organization's main router. Authentication is achieved through health care staff's personal Smart Card certificates with concrete permissions for actions available on CEZIH. Protocol for data exchange between CEZIH and other systems are HL7 v3.0 messages.

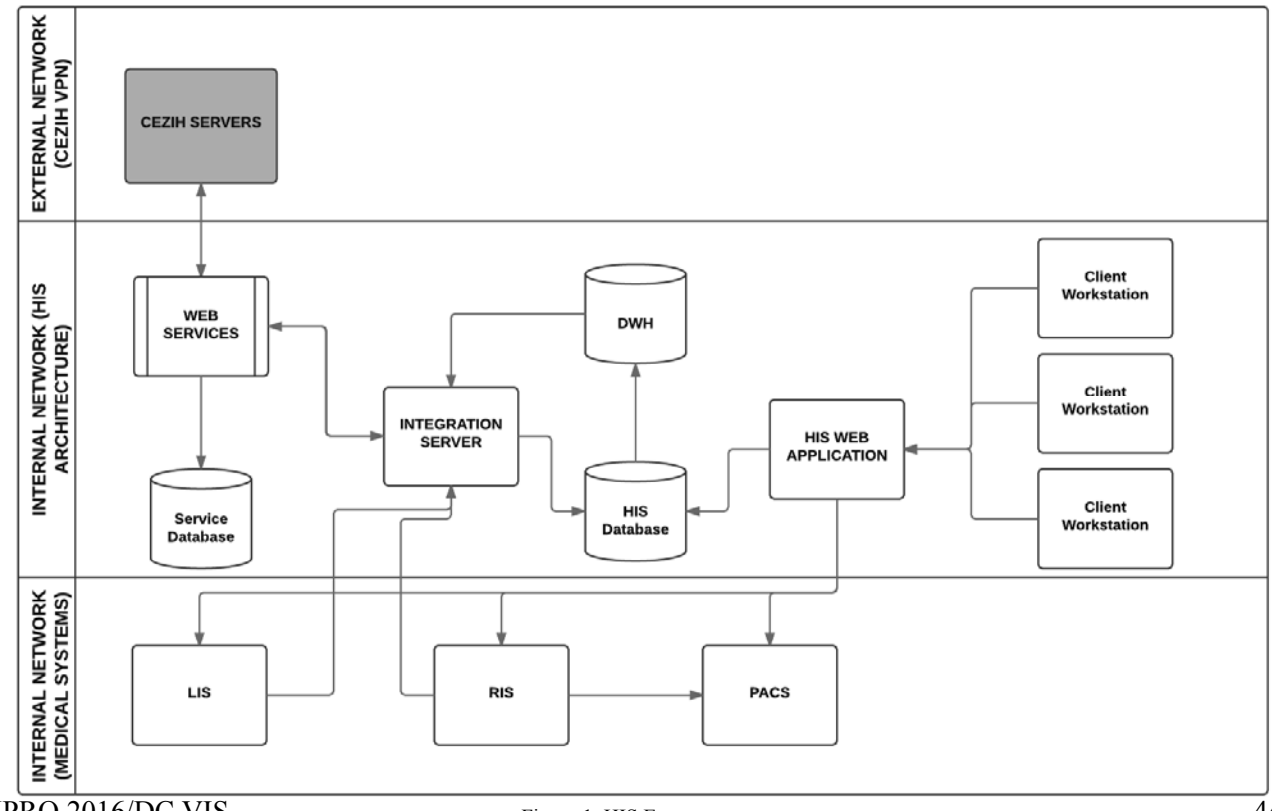


Figure 1. HIS Ecosystem.

Exchange data between HIS and CEZIH is defined by CEZIH by their protocols. CEZIH servers collect data for daily list of ordered patients (e-lists) during night-time in Croatia (around 3AM) and also creates new patient orders to HIS, or declines previously ordered patients (e-ordering) [4].

B. *Proxy (HIS Web Service)*

Proxy has been created for security reasons that we can easily sustain. Every request that comes to our proxy service is validated by SHPD secure server's contract with organization. If contract between CEZIH, SHPD and HIS exist, proxy service will send LLHL7 message to dedicated port asynchronous to organization's Integrations Server. Message that is send is original message that is enriched with data only known between organizations proxy server and Integration Server installed. When Integration Server receives the message, HL7 module interprets the message and sends response back to external proxy caller.

C. *Integration Server*

Every internal and external service (Proxy) must pass through this check point. Integration Server is asynchronous TCP/IP listener server with dedicated threads. To receive and interpret messages, we decided that we need one point, called Integration Server (Figure 2).

By our practice and definition, Integration Server can make decisions to several of implemented HL7 protocols, queue them in FIFO order and make proper response to them. As it is multithreaded service, we created our own extension of HL7 library that supports all HL7 versions between 2.3 and 4.0 (JSON). [5]

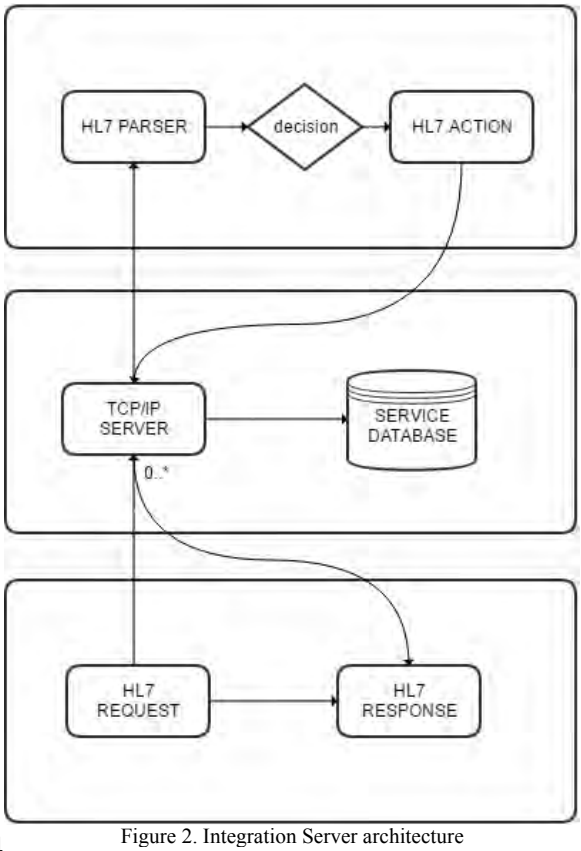


Figure 2. Integration Server architecture

D. *HL7 Message Library*

In our design, HL7 messages are utilized to format all information exchanged across systems. Since complex business logic process may be applied onto HL7 messages while inputting, presenting and data exchanging, we used an HL7 software library NHapi. We decided to create a DOM HL7 library because HL7 message is well structured. A DOM-like implementation could make it easier for developer to access elements in each message.

First of all, a C#-based class library is generated by parsing the XML schema of HL7 versions 2.3, 2.5 and 3.0. Every element in HL7 definition is mapped as an object including fields, segments, and messages. Then, these objects are instantiated at service run-time. When an HL7 message is needed, instance of the message is created and manipulated by program logic. Later, the message could be written out as an XML document and prepared to be carried by SOAP. On the other hand, when an HL7 message is received from SOAP content, an instance of HL7 object could also be created by reading back the XML document via the same library, therefore, both sender and receiver could process the message in a symmetric manner.

4. PERFORMANCE ANALYSIS

"Integration Server" component entered fully service on 2015/3/1. Before the system was deployed, the most concerning issue was too long average patient processing time. After running component in SHPD, average processing time have fell by approximately 35 percent according to statistics (Figure 3).

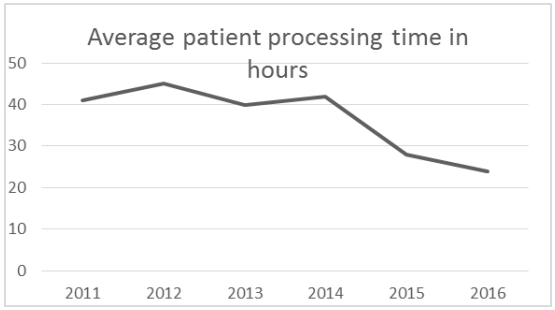


Figure 3 Average patient processing time

5. CONCLUSION AND FUTURE WORKS

In this paper, "Integration Server", component implemented in Special Hospital for Pulmonary Diseases, based on HL7 standard and SOAP Web service is presented. The HL7/Web service framework has successfully provided an effective solution for the challenges of integrating various and heterogeneous information systems while developing a large scale HIS.

The greatest value of Integration Server is its simple architectural design that it deliberately limited to strictly typed messages whose content is validated by both sides. Because of this design, serialization and deserialization of messages from text to C# objects is fast and efficient because it does not use unnecessary generalization of all messages like eg. Mirth Connect server [6] does, which is one of the most popular commercial solution for integration purposes in healthcare sector.

On the other hand, Integration Server configuration is made by manufacturer through simple interface with the minimum option, while the buyer is only concerned about whether a service is working properly.

A new HIS based on the proposed framework is developed and currently serving about 400 patients a day in SHPD.

Further development of the framework is also under study including using HL7 4.0 and integrating other external subsystems.

LITERATURE

[1] I. Henderson, HL7 Messaging, 2nd edition, Tech, 2007

[2] J. Snell, D. Tidwell, P.Kulcenko, Programming Web Services with SOAP, O'Reilly, 2001

[3] N. M. Josuttis, SOA in Practise, O'Reilly, 2007

[4] CEZIH e-lists HL/ messages specifications: http:// http://www.cezih.hr/dokumenti/eListe_Specifikacija_HL7_poruka_20152605_07_03.pdf

[5] B. Smith, Beginning JSON, Apress, 2015

[6] https://www.nextgen.com/Interoperability/Mirth-Solutions/Connect-Pricing?tab=true

Wireless Intrabody Communication Sensor Node Realized Using PSoC Microcontroller

F. Grilec*, Ž. Lučev Vasić*, W. Ni**, Y. Gao**, M. Du**, M. Cifrek*
* University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** Fuzhou University, Laboratory of Medical Instrumentation and Pharmaceutical Technology, Fuzhou, China
grimcro@gmail.com, zeljka.lucev.vasic@fer.hr

Abstract - The aim of this paper is the development of a frequency modulated signal generator and receiver for intrabody communication (IBC) using Programmable System-on-Chip (PSoC) microcontrollers by Cypress. The program is developed in Cypress PSoC Creator and PSoC Designer programs for the CY8C27643 – 24PVXI and CY8C5888LTI – LP097 microcontrollers. Based on the input received from the serial bus, the signal generator synthesizes a FSK signal using digital-to-analog conversion. The signal is transmitted between transmitter and receiver IBC electrodes placed on the human body. The received signal is then demodulated using low pass and band pass filters and a correlator and forwarded back to the computer using the serial bus.

I. INTRODUCTION

Biotelemetry is remote monitoring, measuring and recording of a living organism's function, activity or condition. A network of sensor nodes placed on or implanted inside the body of a subject is called a Body Area Network (BAN). In a traditional wireless system, the system would consist of multiple transmitters and one receiver, all communicating via a standard wireless link [1], which creates electromagnetic radiation that can lead to tissue heating and irritation. High transfer speeds require a high amount of energy consumption, which leads to often recharges or battery replacements in a battery-powered wireless system.

In order to reduce the overall energy consumption, intrabody communication (IBC) systems were developed, [2]. This way of communication uses a lower frequency and less energy for communication over short distances employing the conductivity of the human body. In a typical IBC system the transmitters communicate with a receiver on the body using intrabody communication principle, while the receiver communicates with the remote system using a standard wireless link. The transmitter, i.e. signal generator, encodes and modulates the measured biological analog signal, while the receiver does the opposite and sends digital data to the central unit. The characteristics of data transfer for an IBC system depend on the tissue type, signal path, amplitude, modulation and signal frequency [1]. At higher carrier frequencies tissue conductivity on all tissue types is higher, which leads to lower power consumption. IBC transmitters and receivers can be designed using a wide variety of available off-the-shelf and proprietary chips and microcontrollers for communication [3-5].

Programmable System-on-Chip (PSoC) is a family of microcontroller integrated circuits by Cypress Semiconductor. These chips include a CPU core and mixed-signal arrays of configurable integrated analog and digital peripherals. The integrated circuit is composed of a core, configurable analog and digital blocks with programmable routing and interconnects, [6]. These configurable blocks distinguish the PSoC from other microcontrollers. Each block can be assigned a wide range of functions and interconnected on-chip. Unlike for an application-specific integrated circuits (ASIC), there is no special manufacturing process required to create the custom configuration – only startup code created by a software. The Cypress PSoC systems enable easy programming with the use of programmable digital blocks. The programming consists of a drag and drop canvas interface where the user can place all of the desired analog and digital blocks and route them to the appropriate input and output pins. Each block can be assigned to a predefined functionality with specific parameters, and it's connected with some overhead in the C programming language. The code in C can contain functions and procedures, and interrupt functions. The Cypress PSoC microcontrollers also offer low power consumption and high analog flexibility.

This paper is focused on creating single chip solutions for making a wireless signal generator (transmitter) and a receiver which use the human body as a medium for communication.

II. PSOC-BASED IBC SYSTEM

Using the aforementioned PSoC microcontroller systems implied boundaries due to their processing powers and available settings, which reduced the best possible preformance of the resulting system. During the design process multiple PSoC microcontrollers were tested due to their different limitations, and the final PSoC-based IBC system was realized using a PsoC 5 microcontroller as a signal generator and a PsoC 1 microcontroller as a signal receiver, achieving the system with both as single chip solutions. Startup code was created by Cypress' PSoC Designer for PSoC 1 or PSoC Creator for PSoC 3–5 IDE. The PSoC 1 uses an M8C core, while the newer generations of PSoC 4 and 5 use the ARM cores. With the upgrade of cores, the systems are more oriented for a digital application, while still offering good analog programming options.

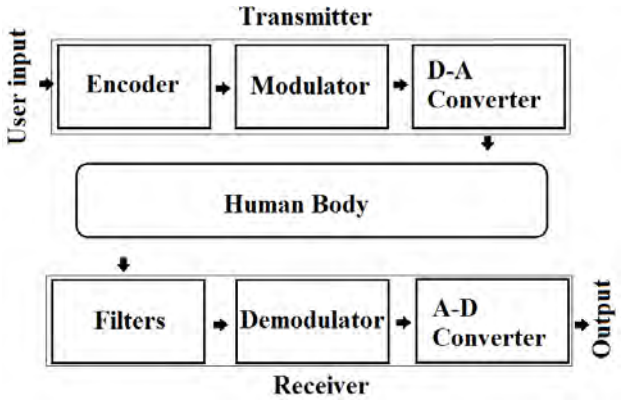


Figure 1. Block diagram of the PSoC-based IBC system

A developed PSoC-based IBC system is presented in Fig. 1. It functions as a COM-2-COM communication system using the human body as a communication medium. It consists of a signal generator as an IBC transmitter, a receiver and two battery-powered laptops.

On the laptop connected to the signal generator the user enters characters into the terminal, which are then encoded, modulated and sent to the receiver. The receiver performs demodulation and decoding of the signal, printing it out on the terminal on another laptop. The laptops are not connected together, nor grounded. The functionality of the system's parts is explained in the following sections.

A. Signal Generator

The signal generator is developed using the CY8C5888LTI–LP097 microcontroller, [7]. Due to the limitations of the receiver and the generator, FSK (frequency shift keying) modulation was selected for this paper, [8]. Digital data representing bits of a different values is encoded as a different frequency, sent using a continuous sine wave, as shown in Fig. 2.

The system uses data from a sine wave lookup table in the memory and transmits it out using voltage digital-to-analog converters with the carrier frequency corresponding to the current input bit. The programmed generator inner structure is depicted in Fig. 3. As can be seen, it consists of several blocks: UART (universal asynchronous receiver/ transmitter), multiplexer, DMA (direct memory access), VDAC8 (8-bit voltage digital to analog converter), and an OpAmp (operational amplifier).

Since the PSoC 1 chip used in this design can reach a maximum carrier frequency of 50 kHz, the PSoC 5 was selected to reach frequencies which would be better for

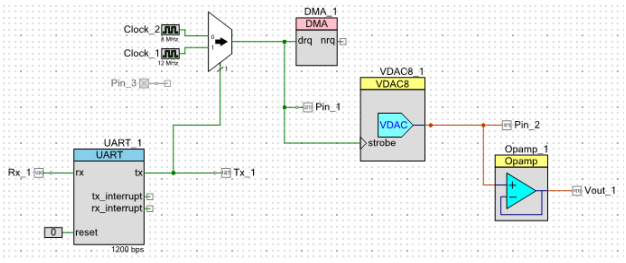


Figure 3. FSK Modulated Signal generated by the PSoC1

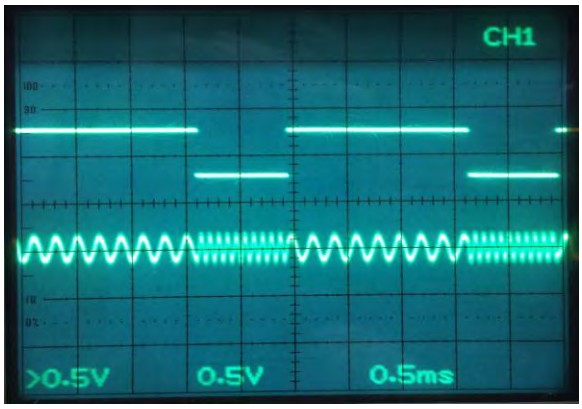


Figure 2. FSK Modulated Signal generated by the PSoC1

operating on human tissue. Due to the receiver's limitations, the resulting carrier frequencies f_c were 167 kHz and 187 kHz with an amplitude of 5 V, generated by system clocks at 2.67 MHz and 3 MHz, respectively.

The signal generator receives an ASCII encoded character from the laptop serial connection. The character bit values are used as a selector signal for the carrier frequency. Depending on the value, one of the two system clocks is used to generate the sine values of voltage through the digital-to-analog converter at the corresponding carrier frequency. Depending on the size of the lookup table, different maximum carrier frequencies can be achieved, as shown in Table 1.

The DMA unit enables the program to write the lookup table's values directly into the VDAC as an alternative to calling the API function, which would significantly slow down the system. Due to the limitations of the receiver, the maximum transmitter frequency was not selected so both chips could operate properly.

B. Receiver

The receiver was built using the CY8C27643 – 24PVXI microcontroller, [9]. Since this is a PSoC 1 microcontroller, it has less processing power then the newer chips, but a wider array of analog processing options. The receiver follows the block diagram shown in Fig. 4.

The input signal is processed through a band pass Butterworth filter with the center frequency of 177 kHz

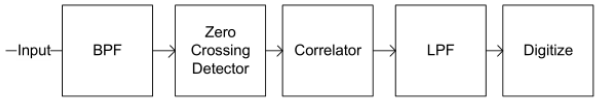


Figure 4. Receiver block diagram

TABLE I. MAXIMUM CARRIER FREQUENCIES FOR VDAC8 WITH A 12 MHZ CLOCK

Sine table elements	f_c, kHz
64	37
16	147
10	243
8	301

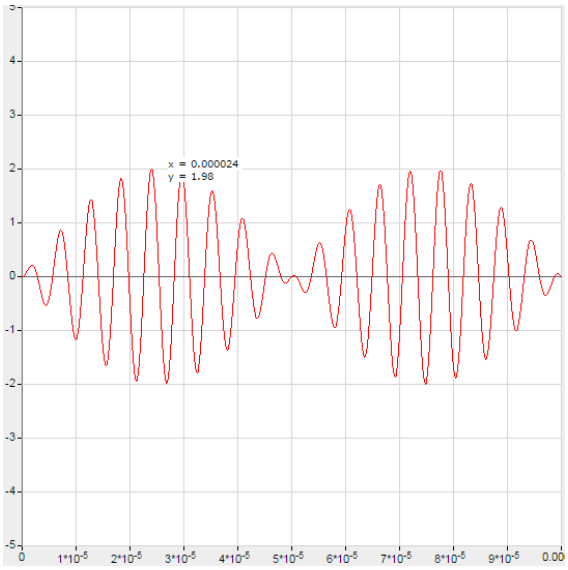


Figure 5. Local maximum of the *Diff* expression with carrier frequencies of 167 kHz and 187 kHz

and a bandwith of 20 kHz. This was the limiting factor on the maximum carrier frequency of the signal generator, since the filter could not work with the rest of the system if the carrier frequency was over 200 kHz.

There are multiple options for detecting a FSK encoded signal and due to the unique structure of the microcontroller, period measurement was the most applicable solution. The processing of the correlator block comes down to a XOR operation between the input signal and it's delayed copy. The delay is calculated from maximizing the following equation:

$$Diff = \cos(2\pi f_{low} d) - \cos(2\pi f_{high} d)\,. \qquad (1)$$

where d is the delay in seconds, f_{low} is 167 kHz and f_{high} is 187 kHz. This results with the maximal difference at a delay of 24 μs (shown in Fig. 5), and the correlator working at a frequency of 1 MHz due to the shift register having a capacity of 24 bits.

The resulting waveform is a square wave with a high duty cycle for f_{high}, and a low duty cycle for f_{low}. This kind of signal processing is shown in Fig. 6. The resulting waveform from the correlator is driven through a low pass filter and a comparator to create a square wave for the UART interface to output the data to the computer terminal through the serial connection.

III. Measurement Results

The system was firstly tested using a direct wired link between the microcontrollers. Even with this connection,



Figure 7. Input and output terminal characters

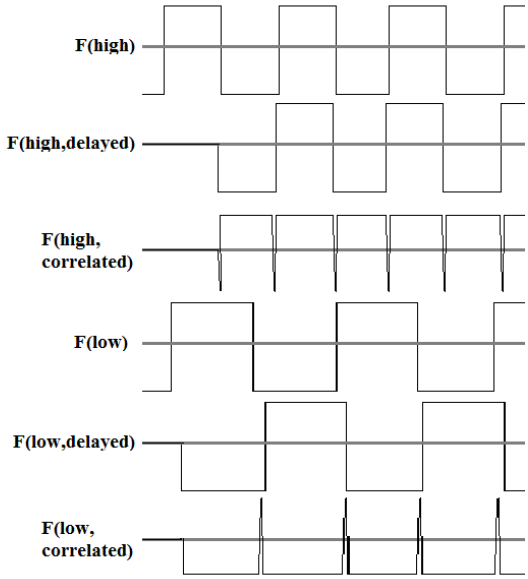


Figure 6. Input square signal and correlator output wave form

an occasional error in transfer would occur, as shown in Fig. 7.

Then, the system was connected to the human body using Ag/AgCl adhesive electromiography electrodes, as shown in Fig. 8. Different connection types were tested, as well as different distances. Several different ways of connecting the electrodes to the body were used for testing: i) both signal and both grounding electrodes connected to the body, ii) only transmitter and receiver signal electrodes connected to the body, while the ground electrodes remain unconnected, and iii) both receiver electrodes and transmitter signal electrode connected to the body, transmitter ground electrode unconnected. The worst transmission was achieved in case i), while the best results were achieved in case iii), as in Fig. 8. The similar results were obtained when changing the placement of the electrodes to different transmitter-receiver distances and when placing the PSoC-based IBC system on different arms.

The actual results were sporadic, and would benefit from a different modulation method and higher processing power for a higher carrier frequency, but prove how much is achievable with a single PsoC chip.

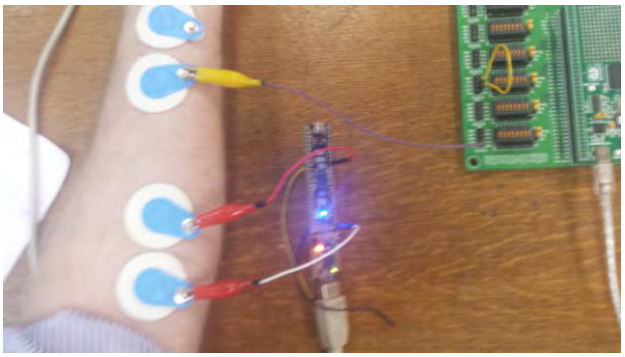


Figure 8. Measurement setup with three electrodes connected to the subject's forearm

IV. CONLUSION

The goal of this paper was a development of a intrabody communication system using single chip solutions without any additional external components. This was achieved using Programmable System-on-Chip (PSoC) microcontrollers by Cypress. The system consists of a signal generator that synthesizes a FSK signal using digital-to-analog conversion, and IBC receiver on which demodulation is performed. Communicaton between transmitter and received electrodes is realized through the human body.

The developed system has achieved the desired functionality. Even with the carrier frequency limitations, functioning at frequencies that are not optimal for IBC use the system works with a low error rate and is a great demonstration of what the PSoC systems are capable of.

ACKNOWLEDGMENT

This research has been supported by the bilateral Croatian-Chinese project "Intrabody Communication as a Key Technology for Internet of Things in Health Applications".

REFERENCES

[1] Ž. Lučev, I. Krois, and M. Cifrek, *Intrabody communication in biotelemetry*, vol. 75 of *Lecture notes in electrical engineering*, pp. 351–368. Springer Berlin Heidelberg, 2010.

[2] T. G. Zimmerman, "Personal area networks: Near-field intrabody communication," *IBM Systems Journal*, vol. 35, no. 3–4, pp. 609–617, 1996.

[3] H. Lee, H. Cho, and H.-J. Yoo, "A 33 μW/node duty cycle controlled HBC transceiver system for medical BAN with 64 sensor nodes," in *Custom Integrated Circuits Conference (CICC), 2014 IEEE Proceedings of the*, pp. 1–8, Sept. 2014.

[4] Ž. Lučev, I. Krois, and M. Cifrek, "A multichannel wireless EMG measurement system based on intrabody communication," in *Proc. XIX IMEKO World Congress, Fundamental and Applied Metrology* (P. S. Girao, ed.), pp. 1711–1715, Sept. 2009.

[5] J. Bae, K. Song, H. Lee, H. Cho, and H.-J. Yoo, "A 0.24-nJ/b wireless body-area-network transceiver with scalable double-FSK modulation," vol. 47, pp. 310–322, Jan. 2012.

[6] Cypress Programmable System-on-Chip (PSoC). http://-www.cypress.com/products/32-bit-arm-cortex-m-psoc, 2015.

[7] Cypress PSoC® 5LP: CY8C58LP family datasheet. http://-www.cypress.com/file/45906/download, 2015.

[8] D. Seguine, "Simplified FSK signal detection, application note #2336," 2014.

[9] Cypress PSoC®: CY8C27X43 datasheet. http://-www.cypress.com/file/134906/download, 2014.

Detection of Heart Rate Variability From a Wearable Differential ECG Device

Jure Slak*, Gregor Kosec**
*student at Jožef Stefan Insitute, Department of Communication Systems, Ljubljana, Slovenia
**Jožef Stefan Insitute, Department of Communication Systems, Ljubljana, Slovenia
jure.slak@student.fmf.uni-lj.si, gregor.kosec@ijs.si

Abstract – The precise heart rate variability is extracted from an ECG signal measured by a wearable sensor that constantly records the heart activity of an active subject for several days. Due to the limited resources of the wearable ECG device the signal can only be sampled at relatively low, approximately 120 Hz, frequency. Besides low sampling rate the signal from a wearable sensor is also burdened with much more noise than the standard 12-channel ambulatory ECG, mostly due to the design of the device, i.e. the electrodes are positioned relatively close to each other, and the fact that the subject is active during the measurements. To extract heart rate variability with 1 ms precision, i.e. 10 times more accurate than the sample rate of the measured signal, a two-step algorithm is proposed. In first step an approximate global search is performed, roughly determining the point of interest, followed by a local search based on the Moving Least Squares approximation to refine the result. The methodology is evaluated in terms of accuracy, noise sensitivity, and computational complexity. All tests are performed on simulated as well as measured data. It is demonstrated that the proposed algorithm provides accurate results at a low computational cost and it is robust enough for practical application.

I. Introduction

It is well known that the morphology of ECG signals changes from beat to beat as a consequence of physical activity, sensations, emotions, breathing, etc. of the subject [1]. The most straightforward measure of these changes is the heart rate variability (HRV), i.e. small variations of beat duration. HRV characterizes the timings of hearth cells repolarization and depolarization processes. The HRV is typically determined by measuring the intervals between two consecutive R-waves (RRI) or intervals between R and T waves (RTI). Several vital signals can be identified from the HRV and therefore it is often used as a health status indicator in different fields of medicine, e.g. neurology [11], cardiac surgery [5], heart transplantation [3] and many more. Typical HRV values of healthy subjects are approximately 40 ms for RRI and 2 ms for RTI [2] (see Figure 1). Therefore it is important to detect considered waves with at least 1 ms accuracy for practical use [6]. This paper deals with the detection of HRV in ECG signal provided by a Wearable ECG Device (WECGD) that is paired with a personal digital assistant (PDA) via Bluetooth Smart protocol [4]. The WECGD, due to the hardware limitations, only measures the signal, while the PDA takes care of data visualization, basic analysis and transmission of the data to a more powerful server for further analyses. In contrast to a standard ambulatory 12-channel ECG measurement, where trained personnel prepare and supervise the measurement of subject at rest, the WECGD works on a single channel, the subject is active and since the WECGD is often placed by an untrained user its orientation might be random, resulting in additional decrease of signal quality. In order to maintain several days of battery autonomy The WECGD also records the heart rate with significantly lower frequencies and resolution in comparison to ambulatory measurements. In this paper we analyse a possible local, i.e. only short history of measurement data is required, algorithm for detection of heart rate variability with 1 ms precision of a signal recorded with 120 Hz.

The rest of the paper is organized as follows. In section II a detection algorithm is presented, in section III results are discussed, and in section IV conclusions are presented.

II. Detection method

In order to evaluate the HRV, the *characteristic point* of each heart beat has to be detected in the signal, which is provided as values of electric potential sampled with a frequency 120 Hz. Since the HRV is computed from differences of consequent characteristic points the choice of the characteristic point does not play any role, as long as it is the same in every beat. In this work we choose to characterise the beat with a minimal first derivative, in another words, we seek the points in the signal with the most violent drop in electric potential (Figure 1) that occurs between R and S peaks.

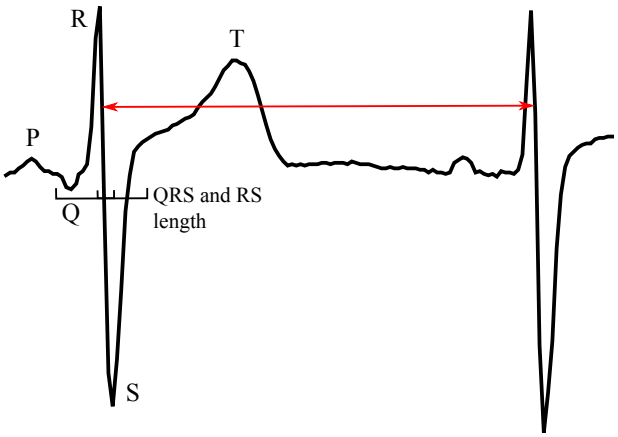


Figure 1. Beat to beat time between two characteristic points.

The detection method is separated in two stages, namely global and local. The goal of the global method is

to approximately detect the characteristic point, while the local method serves as a fine precision detection, enabling us to detect HRV with much higher accuracy.

A. *Coarse global search*

In the first step the algorithm finds a minimal first derivative of a given signal on a sample rate accuracy, i.e. $1/\nu$. The global search method is next to trivial. The algorithm simply travels along the signal, calculating the discrete derivative and storing the position of minimal values found so far. Since the points are sampled equidistantly, minimizing $\frac{\Delta y}{\Delta t}$ is equal to minimizing Δy. The middle of the interval where the largest drop was detected is taken as the global guess t_G. The results of the global search are presented in Figure 2.

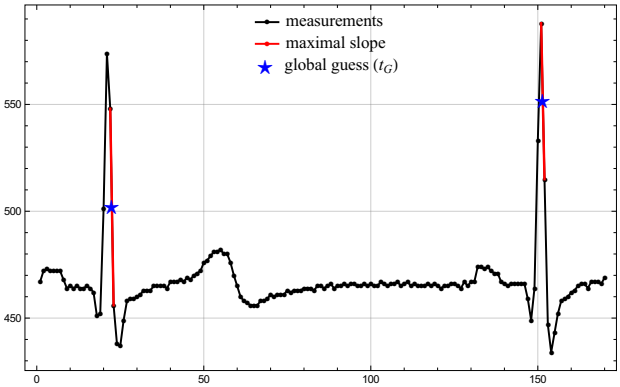


Figure 2. Global search detection of two beats.

B. *Fine local search*

The global search provides only coarse, limited to sample points, positions of the characteristic points. To push accuracy beyond $1/\nu$, the signal has to be represented also in between the sample points. A monomial approximation function based on a Moving Least Squares approach [10] is introduced for that purpose.

The value of the electrical potential at arbitrary time t_0 is approximated. Denote the vector of n known values near t_0 by $\boldsymbol{f}$ (called *support*), and the times at which they were measured by $\boldsymbol{t}$. The approximation $\hat{f}$ of $\boldsymbol{f}$ is introduced as a linear combination of m in general arbitrary basis functions $(b_j)_{j=1}^m$, however in this work only monomials are considered.

$$\hat{f} = \sum_{j=1}^{m} \alpha_j b_j \tag{1}$$

The most widely used approach to solve above problem and find the appropriate $\hat{f}$ is to minimize the weighted 2-norm of the error, also known as the Weighted Least Squares (WLS) method.

$$\|\boldsymbol{f} - \hat{f}(\boldsymbol{t})\|_w^2 = \sum_{i=1}^{n} (f_i - \hat{f}(t_i))^2 w(t_i), \tag{2}$$

In the above equation w is a nonnegative weight function.

The only unknown quantities are the m coefficients $\boldsymbol{\alpha}$ of the linear combination, which can be expressed as a solution of an overdetermined linear system $WB\boldsymbol{\alpha} = W\boldsymbol{f}$, where W is the $n \times n$ diagonal weight matrix, $W_{ii} = \sqrt{w(t_i)}$ and B is the $n \times m$ collocation matrix, $B_{ij} = b_j(t_i)$. There are different approaches towards finding the solution. The fastest and also the least stable and accurate is to solve the Normal System $B^\mathsf{T}W^\mathsf{T}WB\boldsymbol{\alpha} = B^\mathsf{T}W^\mathsf{T}W\boldsymbol{f}$, a more expensive but also more stable is via QR decomposition, and finally the most expensive and also the most stable is via SVD decomposition [9]. The resulting vector $\boldsymbol{\alpha}$ is then used to calculate $\hat{f}(t)$ for any given t. The derivatives are approximated simply by differentiating the approximating function, $\hat{f}' = \sum_{j=1}^{m} \alpha_j b_j'$.

The WLS approximation weights the influence of support points using the weight function w. Usually, such a weight is chosen that points closest to t_0 are more important in the norm in comparison to the nodes far away. Naturally such approximation is valid only as long as the evaluation point is close to the t_0. A more general approach is a Moving Least Square (MLS) approximation, where coefficients α are not spatially independent any more, but are recomputed for each evaluation point. Naturally, such an approach is way more expensive, but also more precise. A comparison of both methods, i.e. WLS and MLS, is shown in Figure 3.

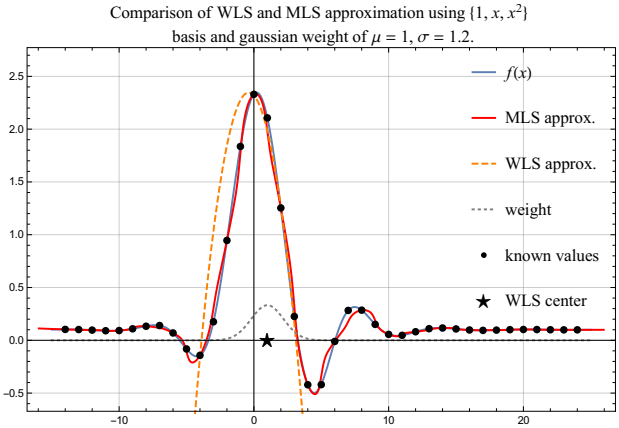


Figure 3. MLS and WLS approximation of a heartbeat-like function $f(x) = \frac{\sin x}{x}\frac{|x+8|-|x-5|+26}{13((\frac{x-1}{7})^4+1)} + \frac{1}{10}$, with measurements taken at points $\{-14, \ldots, 24\}$.

The task of finding the minimal value of the first derivative is equivalent to the task of finding the zero of the second derivative. This zero will be our local approximation t_L of the beat time, $\hat{f}''(t_L) = 0$. Therefore an approximation function with a non constant second derivative, i.e. approximation function with a minimal 3rd order monomial basis, is constructed. The most straightforward approach to find its root is the simple bisection. Bisection requires initial low and high bounds that can be estimated from characteristic point t_G provided by the global method. Using the fact that the QRS intervals last approximately $\Delta t_{\text{QRS}} = 0.1\,\text{s}$ [8, 7], we can seek for the root of the second derivative on an interval $[t_G - \Delta t_{\text{QRS}}/2, t_G + \Delta t_{\text{QRS}}/2]$, and at given sample rate this translates to the search interval of two sample points away from t_G in each direction.

C. HRV calculation and error estimation

Given sampled heartbeat, the fine local search described in section B produces the vector of $\ell + 1$ detected times $\boldsymbol{t_L} := (t_{L,i})_{i=1}^{\ell+1}$ of the RS slopes. Their successive differences represent a vector $\hat{\boldsymbol{r}}$ of detected beat to beat times, the durations of RR intervals.

$$\hat{\boldsymbol{r}} = (\hat{r}_i)_{i=1}^{\ell}, \quad \hat{r}_i = t_{L,i+1} - t_{L,i}$$

Let $\boldsymbol{r}$ be the vector of (usually unknown) actual beat to beat times. Then the heart rate variability (HRV) h is defined as

$$h := \mathrm{std}(\boldsymbol{r}) = \sqrt{\frac{1}{\ell}\sum_{i=1}^{\ell}(r_i - \bar{r})^2},$$

where $\bar{r}$ stands for the average beat to beat time, $\bar{r} = \sum_{i=1}^{\ell} r_i/\ell$. The HRV estimation $\hat{h}$ is calculated as the standard deviation of the detected times $\hat{\boldsymbol{r}}$.

In the following analyses the actual vector $\boldsymbol{r}$ will be known, since the synthesized heartbeat will be analysed. The most obvious error measures are the absolute error of HRV, $e_h = |\hat{h} - h|$ and the absolute error of the average heart beat $e_{\bar{r}} = |\hat{\bar{r}} - \bar{r}|$. Using the vector of errors $\boldsymbol{e} = |\boldsymbol{r} - \hat{\boldsymbol{r}}|$ the average error $e_a = \sum e_i/\ell$ and the maximal error $e_M = \max(\boldsymbol{e})$ can be assessed.

III. Results and Discussion

A. Approximation set-up

In first step the approximation free parameters, i.e. weight function, support size, and number of basis functions, have to be assessed. A single heartbeat is extracted and approximated with all possible combinations of basis functions with orders from 2 to 10 and symmetric supports of sizes from 3 to 15 using both WLS and MLS. An global algorithm described in section A was used to produce the initial guesses. For demonstration four sample cases are presented. The weight function was the same in all four cases, a Gaussian distribution with $\mu = t_G$ and $\sigma = m/4$, which makes sure that all support points are taken into account, but the central ones are more important.

The simplest case is when the support size is the same as number of basis functions resulting in an interpolation. In this case, the weight function is not important, making WLS and MLS entirely equivalent as seen in Figure 4.

In case of small support and small order of monomial basis WLS performs worse than MLS and the approaches differ significantly. However, as we increase the order of the polynomial basis the difference within the bisection interval becomes negligible. This transition can be observed in Figures 5 and 6.

As predicted, the support size is important. Both methods perform badly when too many surrounding measurements are taken into account while still using a low order polynomial approximation. Note that in Figure 7 the initial guess is barely improved and the beat shape is skewed away from the RS drop.

The conclusion is, that for our purposes MLS approximation is unnecessary, as WLS provides good enough results, when used appropriately. Further analysis to determine the best choice of parameters m and n is presented in section C.

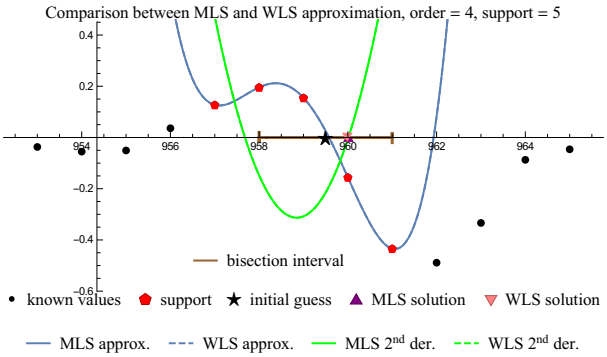


Figure 4. Sufficiently small support implies interpolation, making weight function useless and MLS equal to WLS.

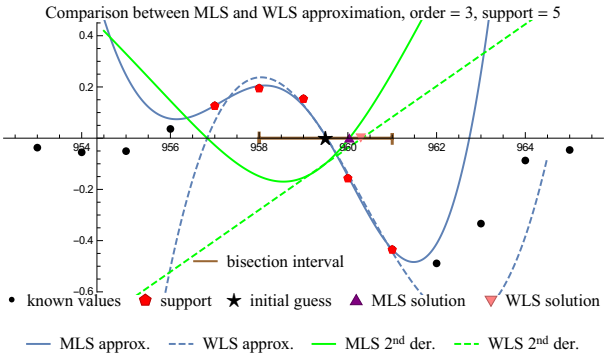


Figure 5. MLS and WLS differ when approximating with a low order polynomial.

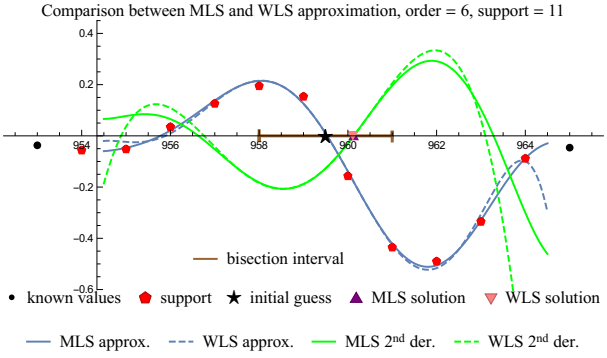


Figure 6. MLS and WLS match when approximating with a high order polynomial.

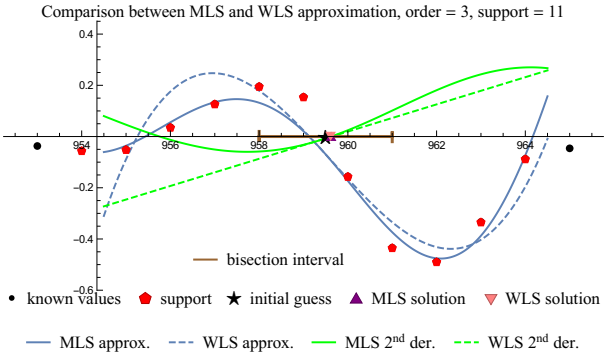


Figure 7. Expected bad behaviour with too many support points and a low order approximation.

B. Computational complexity

The presented algorithm is a streaming algorithm, requiring a buffer to store the current beat in the signal. Let b be the number of measurements per beat, stored in a buffer of length b. The global part of the algorithm makes $O(b)$ operations, being a simple local search. The local part is more expensive. First a $n \times m$ matrix is constructed in $O(mn)$ and the right hand side vector is copied from the buffer. The system is then solved using SVD decomposition in $O(mn^2 + n^3)$. Note that as $m = O(n)$, this step takes $O(n^3)$. The minimal first derivative is found using bisection. To achieve tolerance ε, $\lceil \log_2(1/\varepsilon) \rceil$ function evaluations are needed, each costing $O(m)$ operations. Total time complexity is therefore equal to $O(b + n^3 + m \log(1/\varepsilon))$. Note, that using MLS would require $O(n^3)$ for each function evaluation, resulting in a significantly worse time complexity of $O(b + n^3 \log_2(1/\varepsilon))$. The calculation of average and variance is done later, after the wanted amount of signal has already been analysed.

In practice the algorithm executes very fast, using typical values $b = 150$, $m = 6$, $n = 11$ and $\varepsilon = 10^{-10}$ it runs approximately 0.27 s to analyse 1000 heartbeats ($\approx 10^5$ data points). The algorithm was compiled from C++ source code using `g++5.3.0` with `-O2` flag and run on an `Intel(R) Core(TM) i7-4700MQ` processor.

C. Simulated heartbeat with known variability

The first set of tests for the presented method was performed using a simulated heartbeat. A single real heartbeat was taken and then replicated thousand times, each time shifted by a random offset T, distributed normally around zero, $T \sim \mathcal{N}(0, \sigma^2)$, with $\sigma = \frac{1}{2\nu} = \frac{1}{2}\Delta t$. This means that a decent amount of measurements will be more than Δt apart, a difference that must be detected by global search for method to work. However, around half of the measurements are less than Δt apart, forming suitable ground for testing the precision of the local search. At given sample frequency, σ equals 4.167 ms.

The generated beat to beat times, coarsely detected and finely detected times are presented in Figure 8.

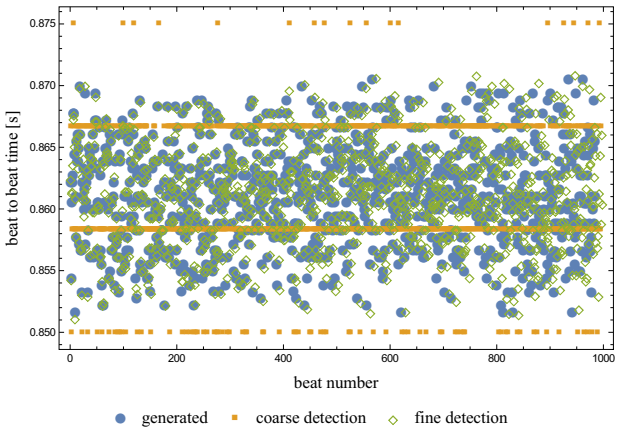


Figure 8. Generated beat to beat times and their global detection.

Beat to beat time precision is significantly improved by the local search. As seen in Figure 9, the distributions of generated and detected heart beats match very well.

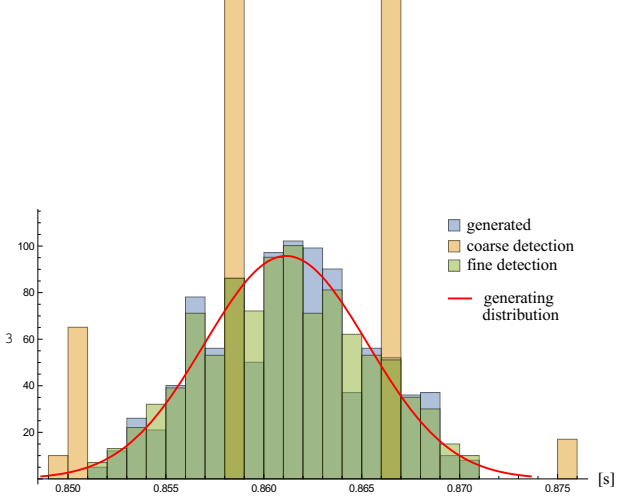


Figure 9. Generated RRI times and their global detection. The middle two coarse detection columns continue off the chart, but are not shown completely for clarity.

Results of RRI and HRV detection by global and local search are presented in Table 1. The generated times were taken as precise and the algorithm was run to produce global and local approximations. Then the average RRI time and HRV were calculated for each data set separately. The average RRI time is estimated very well with both methods, but the precision of the global method is not satisfactory when measuring heart rate variability. The precision is significantly improved using the local search.

A chart showing the average error of the detected times is shown in Figure 10. The red values indicate invalid region with more basis functions than support points. Although the under-determined system could be solved with SVD, in this paper we do not consider that option. Increasing the order of approximations improves results in the beginning, however ultimately a stability limit is reached where approximation system becomes unstable, i.e. the condition number of the matrix becomes critical, resulting in decrease of accuracy to the level of a coarse detection approach. The best results are obtained with order of basis 10–12 and 13–15 support values. From Figure 10 we can also conclude that using MLS is not beneficial enough to justify its computational cost.

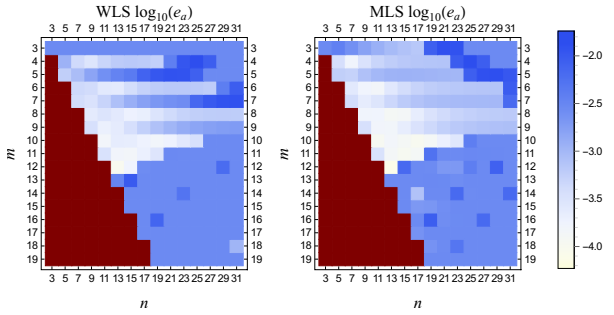


Figure 10. Comparison of WLS and MLS errors using different orders and support sizes.

TABLE 1. RESULTS AND ERRORS OF THE RRI AND HRV DETECTION.

quantity [s]	generated	coarse	fine
$\bar{r}$	0.861136	0.861139	0.861136
$e_{\bar{r}}$	0	$3.34 \cdot 10^{-6}$	$2.83 \cdot 10^{-8}$
h	0.004102	0.005324	0.004137
e_h	0	0.001222	$3.52 \cdot 10^{-5}$
e_a	0	0.002969	0.000263
e_M	0	0.007778	0.000829

D. Noise analysis

The method presented in this paper relies heavily on the values of the derivative. Normally, derivatives are very sensitive to noise. To analyse noise sensitivity of the algorithm a real heart beat signal was taken and normalized by subtracting the average and then divided by maximal absolute value, transforming measurements onto the interval $[-1, 1]$. A uniform noise of level p was then applied to every measurement. Specifically, a uniformly distributed random number on the interval $[-p, p]$, for fixed $p \in [0, 1]$ was added to every measurement point.

Example of the noised and normalized beat and the behaviour of the algorithm is shown in Figure 11.

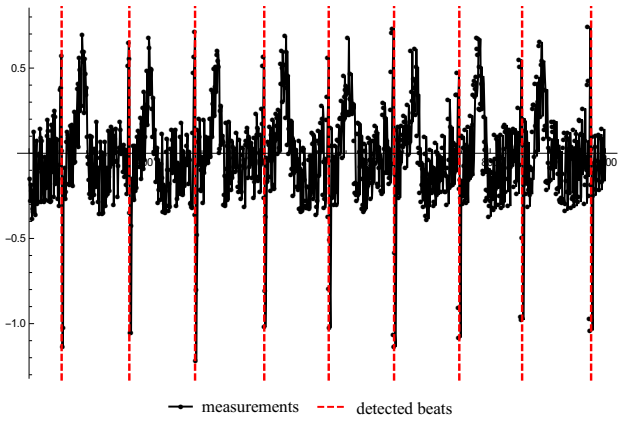


Figure 11. Detecting heart beat in a 25% uniformly noised signal.

An approximate calculation of the critical noise level can be made. Let ν be the sample frequency and p the noise level. The critical noise level is such, that the maximal noise derivative is comparable to the maximal derivative of the signal. The maximal noise derivative is achieved when to subsequent measurements take extreme values of p and $-p$, resulting in drop of magnitude $2p$. In the actual heartbeat, the maximal drop appears during R and S peaks. Taking maximal number of nodes in QRS complex to be $\lceil 0.1\nu \rceil$ as before, and with respect to Figure 1 the RS drop contains at most one third of them, we approximate the maximal drop between two subsequent measurements as the total RS drop divided by number of nodes included. Taking into account that the RS drop equals 2 after normalization, we obtain

$$\Delta y_{\max} = \frac{2}{\frac{1}{3}\lceil 0.1\nu \rceil} \approx 60\frac{1}{\nu}.$$

The critical noise level is such, that both noise levels are approximately equal

$$2p_{\text{crit}} \approx 60\frac{1}{\nu}.$$

At sample frequency of 120 Hz this yields

$$p_{\text{crit}} \approx 0.25.$$

The algorithm was tested as described above for noise levels $p \in [0, 1]$ with a step of 1%. Results are presented in Figure 12. Note that the rise in error corresponds nicely with predicted critical noise.

A little bit more can be achieved using simple statistical analysis. When noise levels are around critical, a few false beats may be recognized among the actual ones. The RRI time before and after the false beat will deviate greatly from the average RRI time. Therefore, we can choose to ignore extreme values, as they are very likely to be wrong. Average and standard deviation are both sensitive to extreme data points, however the median is not, and while extreme values occur rarely, the median should give a good sense of what the "correct" beat to beat time is. To detect the extreme values a Median Absolute Deviation (MAD) was chosen as the simplest and robust enough approach. In this method, the median of differences from the median is taken as the estimate of data variability. All values that differ more that five times MAD from the median are ignored. The results of applying this technique are also shown in Figure 12.

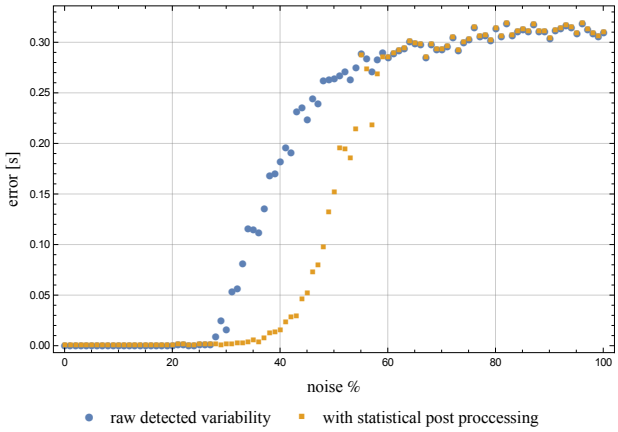


Figure 12. Variability detection at different noise levels.

E. Actual heartbeat analysis

The presented method was tested on sample inputs from the wearable sensor. Detected beats are presented in Figure 13. Average beat to beat time of the subject was $\hat{r} = 0.888264\,\text{s}$ and heart rate variability equals $\hat{h} = 0.282443\,\text{s}$.

IV. Conclusion

In this paper a simple MLS based algorithm for accurate detection of HRV from a low sample rate ECG signal, typically provided by a wearable sensor, is demonstrated. The algorithm is formulated in a fairly general way. Most of the approximation parameters can be easily changed. However, to keep the analyses within the reasonable limits, only the order of monomial basis and the support size are varied to find the optimal set-up. It is demonstrated that increasing order of basis as well as support size improves results up to roughly $m = 12$ and/or $n = 20$, at which point the system matrix becomes ill-conditioned. Based on the presented analyses a basis of 10th order supported with 15 nodes is claimed to be the optimal set-up. It is also demonstrated that using a way more computationally expensive MLS, in comparison to WLS, do not improve accuracy enough to justify it.

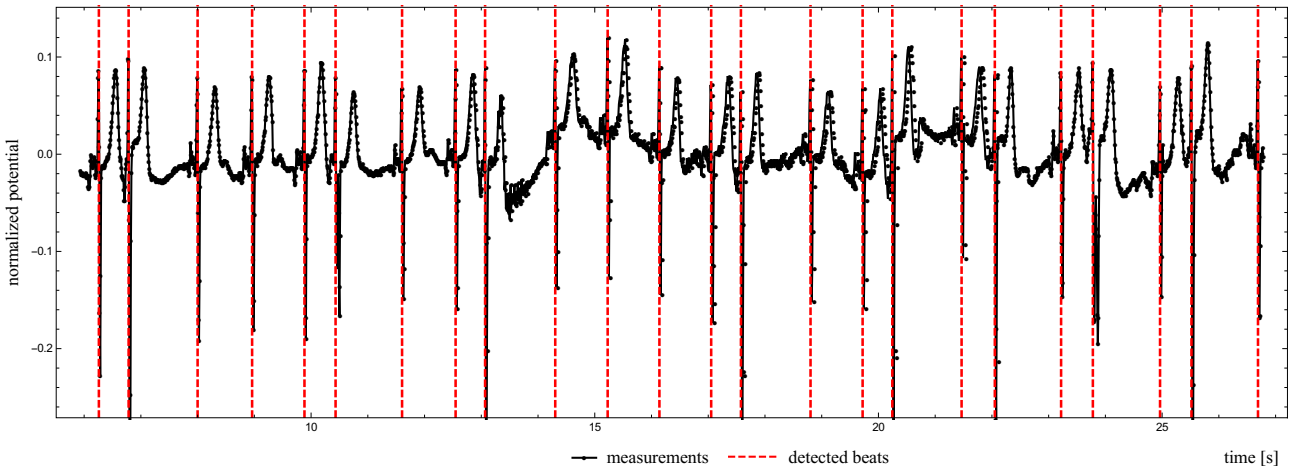


Figure 13. Detection of heart beat from the input measured by wearable sensor.

Based on results from approximation analyses a fast two-stage streaming algorithm for HRV detection is developed. The algorithm is tested on synthetic as well as actual data, achieving good performance. It is demonstrated that the detected beat times of a simulated heartbeat differ from the actual ones with an average absolute error of 0.263 ms at sample frequency of 120 Hz. In other words the detection accuracy is roughly ten times more accurate in comparison with a coarse one. The method is also stable up to approximately 25% noise and computationally extremely effective, successfully processing 1000 heartbeats in approximately 0.27 s on a standard laptop computer.

In future work we will focus on full analysis, including different basis and weight functions, as well as non-linear MLS to fit also the basis shape parameters.

Acknowledgment

The authors acknowledge the financial support from the state budget by the Slovenian Research Agency under Grant P2-0095.

References

[1] M. L. Appel, R. D. Berger, J. P. Saul, J. M. Smith, and R. J. Choen. Beat to beat variability in cardiovascular variables: noise or music? *J. Am. Coll. Cardiol.*, 14:1139–1148, 1989.

[2] V. Avbelj, R. Trobec, and B. Geršak. Beat-to-beat repolarisation variability in body surface electrocardiograms. *Med Biol Eng Comput*, 41:556–560, 2003.

[3] S. Frljak, V. Avbelj, R. Trobec, B. Meglič, T. Ujiie, and B. Geršak. Beat-to-beat QT interval variability before and after cardiac surgery. *Computers in Biology and Medicine*, 33:267–276, 2003.

[4] H. Gjoreski, A. Rashkovska, S. Kozina, M. Luštrek, and M. Gams. Telehealth using ecg sensor and accelerometer, 2014.

[5] J. Kšela, V. Avbelj, and J. M. Kališnik. Multifractality in heartbeat dynamics in patients undergoing beating-heart myocardial revascularization. *Computers in Biology and Medicine*, 60:66–73, 2015.

[6] A. Rashkovska and V. Avbelj. Signal processing methods for ST variability assessment in ECG, 2013.

[7] P Rijnbeek. New normal limits for the paediatric electrocardiogram. *European Heart Journal*, 22(8):702–711, April 2001.

[8] Peter R. Rijnbeek, Gerard van Herpen, Michiel L. Bots, Sumche Man, Niek Verweij, Albert Hofman, Hans Hillege, Matthijs E. Numans, Cees A. Swenne, Jacqueline C.M. Witteman, and Jan A. Kors. Normal values of the electrocardiogram for ages 16–90 years. *Journal of Electrocardiology*, 47(6):914–921, November 2014.

[9] L. N. Trefethen. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, 1997.

[10] R. Trobec and G. Kosec. *Parallel scientific computing: theory, algorithms, and applications of mesh based and meshless methods*. Springer, 2015.

[11] V. Urbančič-Rovan, B. Meglič, A. Stefanovska, A. Bernjak, K. Ažman-Juvan, and A. Kocijančič. Incipient cardiovascular autonomic imbalance revealed by wavelet analysis of heart rate variability in Type 2 diabetic patients. *Diabetic Medicine*, 24:18–26, 2007.

Penetration of the ICT Technology to the Health Care Primary Sector - Ljubljana PILOT

Tonka Poplas Susič* and Uroš Stanič**
* Health Centre Ljubljana, Ljubljana, Slovenia, antonija.poplas-susic@zd-lj.si
** Kosezi d.o.o., and Biomedical Research Institute –BRIS, Ljubljana, Slovenia, uros.j.stanic@gmail.com

Abstract – ICT technology on primary care sector can significantly improve the level of health care and consequently reduce its costs. Some initial results are shown in this paper obtained from the pilot study in the region of Ljubljana, Community Health Centre (CHC). ECG body sensors have been made available to several volunteers with potential disturbances in the heart rhythm in order to reach evidence based diagnosis and further treatment. Initial results from the screened volunteers show that arrhythmias can be excluded or confirmed and specified by this sensor without need to refer patients immediately to the secondary care level. This could have an important impact on the length of waiting queues and on the time needed for the competition of diagnostic process, which implies time and money savings. The ECG body sensor did not disturb patients and can be applied also to the disabled people at home. It enables patients centered medical care with early diagnoses, which will substantially influence the future health treatment.

I. INTRODUCTION

According to the OECD data, Slovenia has among EU countries the lowest number of subjects that could not afford health care services because of financial or other reasons [1]. Slovenia is in the group of 17 EU countries that assure the health care services to all its citizens and even to immigrants. The comprehensive analysis of WHO also revealed that Slovenia has a good coverage and available primary health care sector [2]. These facts motivated a large national consortium with over hundred stockholders from industry and academia to submit the e-Health and m-Health (EMZ) national project proposal on the cohesion funds call as a part of the SPS national system [3]. The ambitious project goals until 2020 are to reach 30% of the national population by ICT supported services with ultimate effects reflected in 10% of lower costs for health care [4] that could result in 2% of the GDP growth and could generate 5000 new work places. The EMZ realization has already started with several project pilots based on the multifunctional ECG body sensor of vital functions [5].

The Slovene approach within the EMZ project is somewhat similar to the Danish one, but compared to Denmark lacks system approach [6]. Namely, the JRC study is based on data collection depicting Danish conditions and serves as a benchmark for similar studies in the rest of the EU-27 countries. For example, one of positive characteristics of Danish health care system is that the percentage of people having a long-standing illness or health problem (in 65-74% of population) is lower in Denmark (37.9%) compared to the EU-27 average (54.4%) (source Eurostat 2009). The difference of 15% reveals the higher quality of life of elderly population but also represents considerable saving of money. One third of Danish population suffers from one or more chronic diseases. As 80% of the Danish healthcare system resources incurred by chronically ill patients (19.5 bn. EUR) the saving of 15% is around 3 bn. EUR. They concentrated their attention to the following eight diseases: diabetes, cancer, cardiovascular diseases, osteoporosis, muscle and skeleton diseases, asthma and allergies, COPD, and mental illnesses. For each of these populations the calculation was made, number of patients and the costs of treatment. Based on this data, they established Public Welfare Technology Foundation that (PWT) that funds two types of projects; Local demonstration (pilot) and national implementation projects. Similar to Danish approach, it is expected that the beneficial results of the Ljubljana PILOT will be implemented nationally as well.

It was shown that by using three differential leads (DLs) it is possible to synthesize high quality 12-lead ECG [7], [8]. Further, based on the ECG sensor measurements it is possible to analyze the timing of ECG signals and, based on this, to diagnose some arrhythmias [9]. The practical validation of these hypotheses is among the main goals of the presented pilot.

The design of pilot system is tailored to the specific needs of the Community Health Centre Ljubljana, which consists of seven units, offering continuous care to the citizens located in their region and to all daily migrants and visitors of Ljubljana, on the area of preventive [10] and curative [11] medical care. CHC is a development oriented institution in primary health care with more than 1400 employees (129 general practitioners) and with more than 440.000 registered patients. The medical personnel are committed to deliver immediate care, when the patients visit them. They wish to ensure a high-quality and time-wise optimal access to health care services for all of their users in all segments of acting. The patients come from Ljubljana and its periphery with rural areas and are treated within the medical doctrine and defined ethics aspects.

It is known, that primary care level solves more than 80% of patients health problems and just some patients are referred for treatment to the secondary/tertiary care level

[12]. Fast evolving technology can improve procedures and measures taken by family physicians.

The proposed pilot system which is for the first time implemented in primary care e.g. CHC could trigger a broad penetration of the ICT in medicine in the primary care level on the national level that could improve and complete the integrated health care with evidence based decisions. Also disable patients that cannot visit their physician could be supervised and treated by ICT. The pilot introduced in Ljubljana is providing responses of medical personnel to the system user interfaces and also the system acceptability level assessed with regard to patients, families and caregivers. The participating patients, occasional experiencing difficulties with heart rhythm, will be provided with a comprehensive care in terms of screening with preventive and curative treatments, and diagnostic. Initial results of the study including the acquired measurements and system efficiency are presented and discussed in the paper.

The rest of the paper is organized as follows. In Section II. the methodology used in the pilot measurements, the participants data and the ICT equipment is described. Section III. presents some illustrative measured results and some of the first feedbacks from patients and medical professionals. In Section IV., the first results from the pilot measurements are shown and some potential benefits and limitations are discussed. The paper concludes with the description of future broadening of the pilot to the rest of Slovenia.

II. METHODS

A. Methodology

Differential leads (DL) [7] are bipolar leads that measure the potential between two closely placed body-surface electrodes. The DL measurements can be obtained by a body sensor that is fixed on the chest in the vicinity of heart. In Fig. 1., a set of possible positions for ECG sensor are shown. Sensor leads are marked with two connected circles. For example, the potential difference between standard leads (V2-V1) can be measured by a body sensor, with its electrodes (marked with black circles in Fig. 1.) placed on standard positions of V1 and V2.

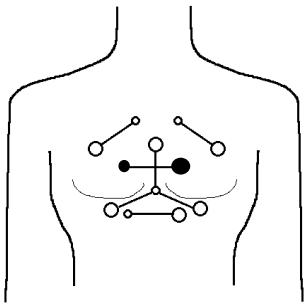


Fig. 1. Several possible positions of ECG body sensors.

If the body sensor is equipped with a small rechargeable battery and radio it can be implemented as a minimally obtrusive wireless body sensor [13]. We use a prototype sensor shown in Fig. 2., termed as PCARD sensor in the rest of the paper, that is able to measure the ECG for three days without charging. The acquisition of ECG signal is implemented with the sampling rate of 125 Hz and the resolution of 10 b/sample. For details about the recording device and procedure please refer to [14].

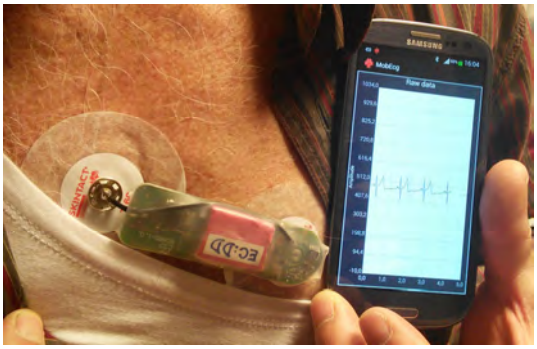


Fig. 2. An example of PCARD sensor.

B. Patients selection

Thirteen volunteers (8 male, 5 female, mean age ± SD = 50.6 ± 9) with no previous medical record related to arrhythmia problems were included in the initial study for validation of the measurement protocol, and for testing the usability and acceptance level of the ICT equipment.

In January 2016, general practitioners (GPs) in Health Centre Ljubljana started to identify patients with potential periodic heart rhythm problems which are difficult to be confirmed by the outpatient electrocardiography (ECG). Indicated problems were: palpitation, attacks of atrial fibrillation, and occasional disturbance of hearth frequency. After detection, patients were asked to permit reporting their name/surname and mobile phone number to the contact person in the Simulation Centre (SIM) in the CHC. All candidates were informed by their GP that the study is a pilot and that they can refuse contribution in any moment during the study. All pilot participants have signed a written consent.

The contact-person in SIM systematically contacted every patient on the list and discussed with them all aspects of usefulness, technical data, reporting, health related side-effect, writing diary and handling the sensor. Additionally, every patient got written information about the use and handling of the ECG sensor.

Twenty volunteers out of 37 wanted to participate in screening and no one had a heart related health problems earlier. Protocol of the study was sent to the Ethic Committee for consideration.

C. Measurement procedure

A trained medical technician explains and shows patients how to fix the body sensor on a most appropriate position from Fig. 1., how to apply skin electrodes and how to connect the ECG sensor on them and how to start the measurement with the mobile application on the phone. Additionally, they are instructed how interrupt and start a measurement.

Pilot participants received a diary to note observed difficulties and troubles. They were also asked to record

their wellbeing and mark unusual events. They received contact data of the assigned professional, whom they can contact if necessary. In case of urgent/life threatening situations they were advised to call 112.

III. RESULTS

Up to date, thirteen healthy volunteers ECGs have been recorded. Additionally, five patients return their sensors with the recorded long term ECG measurements. Participants wore the sensor from few hours to more days depending on the rhythm disturbances. The summary of pilot participants is given in Table I.

	Actul No.	Planned No.	Avg. age ± SD (min, max)	Avg. duration of measure. ± SD (min, max) in h
Volunteers	20	100	52.4 ± 14.3 (25, 82)	8.8. ± 23.7 (0.03, 90)
Patients	5	50	48.2 ± 11.4 (34, 60)	1.2 ± 0.4 (1, 2)

Table I. Summary pilot participants data.

A typical history of a ten-day measurement, as visible for the GP in the custom ECG analysis program, is shown in Fig. 3. We can see that the measurement is composed of more files, represented as colored rectangles, which is a consequence of measurements interruption, because the participants have been instructed to stop the measurement, for example, before taking a shower, or if they noticed that the sensor has lowered the battery level, etc.

For easier and faster evaluation, the rectangles are colored in order to indicate the standard deviation of RR intervals (RRI), which is in some rough correlation with potential arrhythmias or extreme dynamic changes in the hearth beat as a consequence of physical activity.

The GP can simply click on a selected rectangle to visualize a particular measurement. In Fig. 4., an one hour segment of the encircled measurement from Fig. 3. is shown. The ECG signal is in red, while its RRIs are marked with blue crosses. The y coordinate of each cross represents the difference in time between current and previous R-time. Several non expectable changes in RRIs are easily noted. Some of them are a consequence of false detected R waves, but some of them indicate an actual abrupt change of heart rhythm. Now, the user can zoom the interesting parts, e.g., significant changes in RRI marked by left arrow in Fig. 4., or a discrete drop in RRI (from 1.2 sec → 1 sec) marked by the right arrow.

Zoomed ECG around the left arrow discovers a series of premature atrial beats shown in Fig. 5, again with ECG signal by red curve and RRI by blue crosses. In Fig. 6 further zoom around the start of this sequence is applied to visualize details of premature beats morphology. It is clear that based on this record we can exclude atrial fibrillation as one of the most important reason for arrhythmia and/or other danger arrhythmia at that very moment and diagnose palpitation further, if needed according to patient's clinical condition.

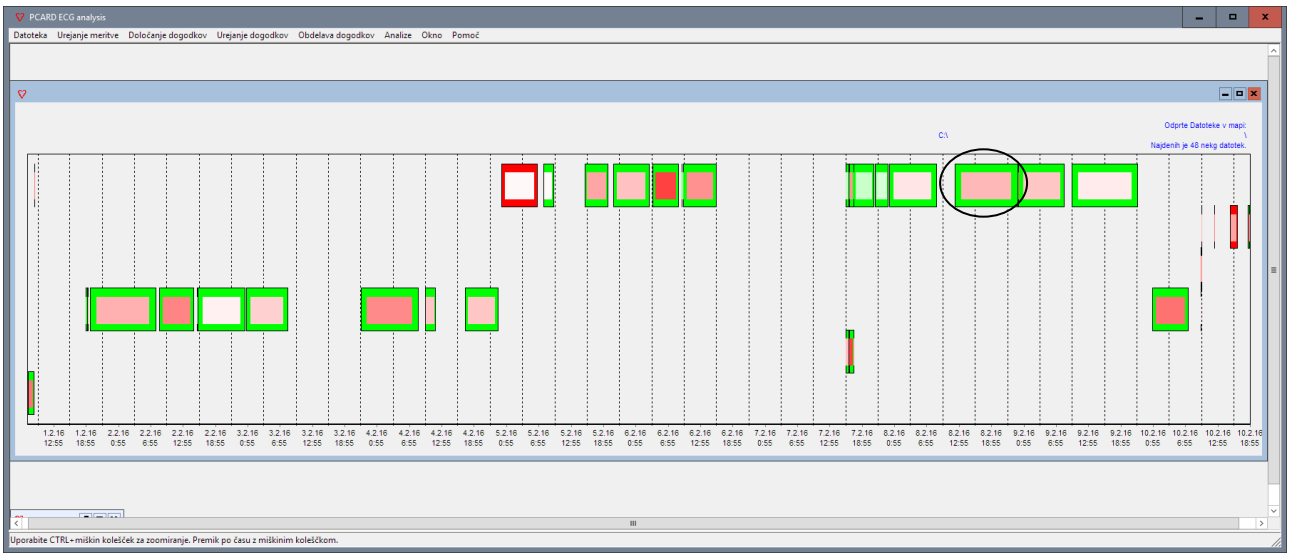


Fig. 3. A typical longterm measurement overview recorded by PCARD sensor.

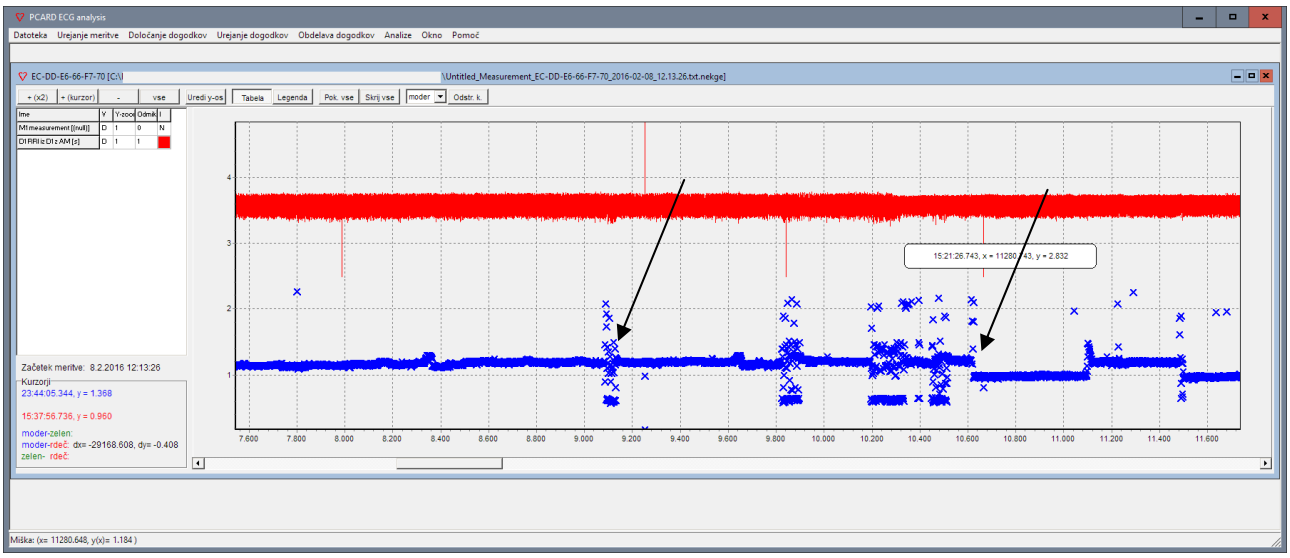


Fig. 4. One hour segment of the encircled measurement from Fig. 3.

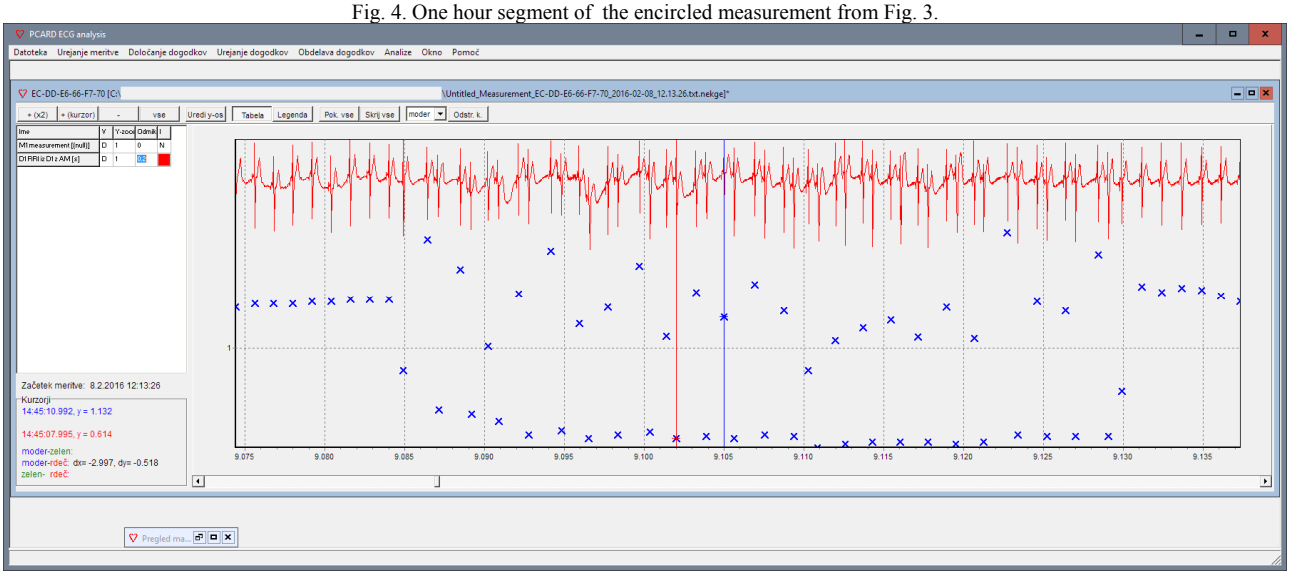


Fig. 5. Zoomed section around the left arrow from Fig.4.

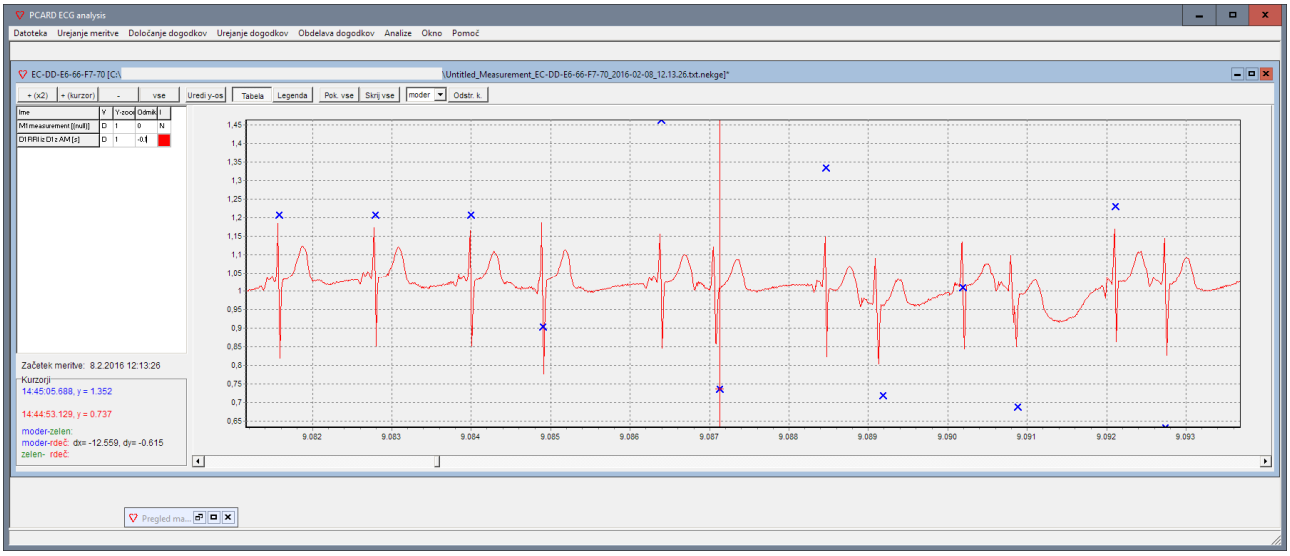


Fig. 6. Further zoom of Fig.5. for detailed view of premature beats morphology.

Patients did not fulfill their diaries with time of emergence of palpitation although they were strongly asked to do that. They reported to feel safe with ECG sensor and nobody complained about side effects. By removing the patch no redness or/and itching of the skin was observed. With no significant reason, one of the patients wanted to prolong the period of wearing sensor and it was allowed. After finishing recording ECG, patients were asked for their feed-back and there wasn't any suggestion to improve the methodology of the planned pilot.

IV. DISCUSSION

Clear benefits of the proposed pilot on the primary level is in usefulness of the sensor for the family physicians since it guarantees better evidence-based diagnostic process, in optimizing health care resources and in fast diagnosis. Most noticeably, waiting list could be shortened and less time could be used to conclude diagnostic process (time and money saving treatment). First feedback from the pilot participants indicates that ECG body sensor does not disturb patients nor their family members. It enables patients centered medical care and early diagnosis and will in this way substantially influence the patient treatment in the near future.

Potential limitations are in the personal response of patients to the ICT equipment. Some patients are interested and eager to manage their measurements alone and to follow their ECG on their phone. Other patients are either not interested or not able to actively follow the measurement therefore they can just make a measurement. According to some reports [4], following the ECG parameters on the screen can disturb patients and make them nervous. These kinds of volunteers decided for "recording without seeing" and totally believed in the purpose of the study. After the measurement they bring the sensor to the medical centre or sent the recorded files from the phone to their doctor, who analyzes them.

Analyzing the first measurements, we can hypothesize that arrhythmias and atrial fibrillations, as one of the important cause for the stroke or sudden dead, can be excluded or confirmed by this sensor at the point of care without the need to refer patients for diagnostic process to the secondary care level.

The sensor can be worn longer (typically 3 days) compared to the Holter monitor (24 hours). The time is limited by the lasting of battery, which can be easily charged at home. The ECG sensors are not expensive and therefore available for all. The sensor can be removed if no problems are observed. Even that some data can be lost, because of interruption of measurements or occasional lost radio connection; the recordings still provide significant evidences about long term health state from patients' everyday environments.

The light and unobtrusive ECG sensor does not disturb patients, and can be applied also to disabled people at home. The sensor enables patients' centered and personalized medical care provided by the team of different experts, e.g., physician, practice nurse, home care nurse or emergency nurse, technicians, engineers and researchers.

The socio-economic impact of the use of the wireless ECG sensor in primary care sector in the CHC cannot be estimated yet because of low number of pilot participants and even lower number of pathological records. However, more extensive and profound study will be implemented within the developing pilot (100 patients) in order to asses also other factors, e.g., processes' quality indicators, usability, clinical significance, and economic parameters.

Like majority of EU countries Slovenia is also challenged by its growth of aging population. This demographic shift is a serious problem for the government as the number of citizens needing medical care will increase very fast in the future and consequently the associated costs will rise. Some recent studies and projects [15], [4] demonstrated the cost-effectiveness of mHealth services could generate 50-60% reduction of hospital nights and re-hospitalizations and reduce overall elderly care expenditure by 25%.

The planned pilot study intents to gather experiences during the application of the ECG wireless sensors for prolonged remote cardiac monitoring of citizens/patients that are under medical care of GPs from CHC. It is expected that the benefit of the new approach will result in a reduction of visits in CHC and stimulate optimal prescription of medication and present new suggestions for healthier life style. The results from the pilot case will be used to demonstrate to the government and other decision makers how effective the mHealth technology can be. It will be demonstrated that by using the already available ICT technology the coverage of all Slovene regions can be achieved at affordable costs.

V. CONLUSION

The ECG body sensor enables faster and economic diagnostic process for different kind of arrhythmias on the primary level. Further implementation will engage more patients and will be expanded through CHC in 2016. The size and influence of the CHC can contribute to the coordinated widening of the pilot on the national level and possibly expanded to other EU regions.

ACKNOWLEDGMENT

This work was partly financed by the Slovenian Research Agency under Grant P2-0095.

REFERENCES

[1] "OECD Health Statistics 2015," 18. 2. 2016, 2016; http://www.oecd.org/health/health-data.htm.

[2] "WHO - Global Health Observatory (GHO) data," 18. 2., 2016; http://www.who.int/gho/publications/world_health_statistics/2015/en/.

[3] "Workshop EMZ, Proc. 18th Int. multiconf. Information Society," 2015.

[4] C. Klersy *et al.*, "Economic impact of remote patient monitoring: an integrated economic model derived from a meta-analysis of

randomized controlled trials in heart failure," *Eur J Heart Fail,* vol. 13, pp. 450-459, 2011.

[5] M. Depolli, J.M. Kališnik, T. Korošec, A. Poplas-Susič, U. Stanič, and A. Semeja, "PCARD platform fo mHEALTH monitoring," *Informatica,* vol. Accepted for publication, 2016.

[6] "Strategic Intelligence Monitor on Personal Health System, Phase2," *JRC Scientific and Policy Reports*, 2013.

[7] R. Trobec, and I. Tomasic, "Synthesis of the 12-Lead Electrocardiogram From Differential Leads," *IEEE Trans. Inform. Technol. Biomed.,* vol. 15, no. 4, pp. 615-621, 2011.

[8] I. Tomasic, S. Frljak, and R. Trobec, "Estimating the Universal Positions of Wireless Body Electrodes for Measuring Cardiac Electrical Activity," *IEEE Trans. Biomed. Eng.,* vol. 60, no. 12, pp. 3368-3374, 2013.

[9] S. S. Lobodzinski, "ECG patch monitors for assessment of cardiac rhythm abnormalities," *Prog. Cardiovasc. Dis.,* vol. 56, pp. 224-229, 2013.

[10] T. Poplas-Susič, I. Švab, and M. Kolšek, "Community actions against alcohol drinking in Slovenia - a Delphi study," *Drug and alcohol dependence,* vol. 83, pp. 255-261, 2006.

[11] V. Ivetić, J. Kersnik, Z. Klemenc-Ketiš, I. Švab, M. Kolšek, and T. Poplas-Susič, "Opinions of Slovenian family physicians on medically unexplained symptoms : a qualitative study," *Journal of international medical research,* vol. 41, pp. 705-715, 2013.

[12] L. Green, G. Fryer, B. Yawn, D. Lanier, and S. Dovey, "The ecology of medical care revisited," *N Engl J Med,* vol. 344, pp. 2021-2025, 2001.

[13] K. Bregar, and V. Avbelj, "Multi-functional wireless body sensor - Analysis of autonomy," in *36th International Convention on Information & Communication Technology Electronics & Microelectronics (MIPRO)*, 2013, pp. 322-325.

[14] V. Avbelj, R. Trobec, B. Gersak, and D. Vokac, "Multichannel ECG measurement system," in *Tenth IEEE Symposium on Computer-Based Medical Systems*, 1997, pp. 81-84.

[15] "mHealth in Europe: Preparing the ground – consultation results published," *Green paper*, 2014.

Image-based metal artifact reduction in CT images

A.Šerifović-Trbalić*, A. Trbalić **
* University of Tuzla/Faculty of Electrical Engineering, Tuzla, Bosnia and Herzegovina
** Drvodom doo, Tuzla, Bosnia and Herzegovina
amira.serifovic@untz.ba

Abstract - The presence of metal in the scanning field of a CT scanner can create so-called metal artifacts in the reconstructed images. These streak artifacts obscure information about anatomical structures, making it difficult for radiologists to correctly interpret the images or for computer programs to analyze them. Several methods have been proposed in the literature based on the reconstruction of the missing/corrupted projection data using the raw data directly from the tomographs. This paper proposes an image-based strategy consisting of the image registration and the morphological reconstruction. Preliminary results on the CT images from different CT scanners and patients are presented and discussed.

Keywords: CT, metal artifact, image registration, mathematical morphology

I. INTRODUCTION

The presence of metal in the scanning field of a CT scanner can create severe artifacts in the reconstructed images. Metal causes beam-hardening, resulting in dark and bright shading artifacts, in particular dark shading between metal objects. Also, metal has a large linear attenuation coefficient leading to photon starvation in the shadow of the metal object, which can result in severe noise streaks. These streak artifacts obscure information about anatomical structures, making it difficult for radiologists to correctly interpret the images or for computer programs to analyze them.

In the last decade, various metal artifact reduction (MAR) approaches have been proposed. However, there is still no robust solution to this issue and it remains a challenging problem.

Most of the currently available MAR techniques [1, 2] are applied to the raw sinogram data obtained directly from the tomographs which are huge encrypted files that are not easy to decode and manipulate. In general, some assumption on the missing information parts is necessary, and typically the metal-affected parts of the sinogram is replaced by interpolation. Usually, metal is segmented in the image space of each uncorrected slice and the metal artifacted region is forward projected to sinogram space. Then, the affected areas are replaced by linear interpolation [1]. The interpolation can also be bi-directional within a slice and weighted [3, 4], or even comprise interpolation across slices over the longitudinal scan range [5]. The method can be extended by normalizing the sinograms [6] and by combining the high frequencies of the original reconstruction with the low frequencies of a MAR-processed image [7]. Alternatively, the interpolation can be based on a tissue-class model and an automatic segmentation of the initial reconstruction [8]. Adaptive filtering techniques have also been proposed [9], as well as the replacement of the filtered back-projection by algebraic solvers [10-12]. Hybrid approaches combining or iterating the above elements have also been suggested [13, 14].

However, in the DICOM-based applications, information about raw sinograms is not available, so different approaches are needed. Alternatively, MAR algorithms may use forward-projected 'virtual sinograms' produced by forward projection of CT images in DICOM format instead of raw sinogram data [15]. The projection data affected by metallic objects are detected in the sinogram space through segmentation of metallic objects in the CT image followed by forward projection of the metal-only image. Thereafter, the affected sinogram bins are replaced by interpolated values from adjacent projections using the spline interpolation technique. Image-based method was proposed in [16]. This method uses the information of the adjoining slice of an affected one to detect its artifacted areas and remove them.

This paper proposes an image-based strategy based on the image registration and the morphological reconstruction. It makes use of information from the neighboring slice of 3D CT images to compensate for data corruption and reveal diagnostic information from the data contaminated by multiple metal objects.

II. METHOD

The proposed method is composed of three steps: image registration, morphological reconstruction and image inpainting. Image registration, also known as image fusion, can be defined as the process of aligning two or more images in order to find the optimal transformation that best aligns the structures of interest in the input images [17-19]. Image registration is applied to the neighboring CT slices (images), one that contains a metal artifact and one without such an artifact. In order to have more similar anatomical structures on both slices, these two slices are chosen to be as close as possible. As a result of the registration, a marker image is obtained. This marker image is used to perform the morphological reconstruction. Morphological reconstruction based on images with artifacts and marker image performs detection of the artifact caused by metal

This work has been supported by the Federal Ministry of Education and Science of Bosnia and Herzegovina.

objects, leaving anatomical structures such as teeth and bones. In this way, the image area with the artifacts will be isolated and reconstructed using the undamaged parts of the image by applying the inpainting method.

A. Hierarchical image registration

First step of the proposed MAR method represent the hierarchical approach for elastic medical image registration using approximating splines [20, 21]. Hierarchical image registration framework decomposes the non-rigid registration problem into numerous local rigid transformations. By progressively subdividing the floating image, an image pyramid is constructed, comprising at every level of a growing number of sub-images whose centers and their related transformation parameters constitute the control points of the deformation field. The initial registration parameters of every sub-image are inherited from the previous hierarchical level. The new partitioned sub-images are then checked by Moran's information consistency test and only sufficiently structured image patches are passed to the registration stage. For all the sub-images that fail this consistency test, the hierarchical splitting is stopped. Maximization of mutual information (MI) is used to find the best individual registration parameters that locally match every sub-image to the reference image. After each level, outlier sub-images are detected using heuristic consistency tests. The stopping criterion checks whether any sub-images are left that can be further subdivided according to Moran's test. Otherwise, the dense deformation field of the final result is calculated using the Gaussian elastic body splines (GEBS) with an anisotropic landmark errors and rotational information [22], densely interpolating the local transformations of the sub-images over the whole image domain. The set of landmarks for the GEBS is formed by the centers and the rotation information of all terminating sub-images. Since the GEBS approach is based on a physical model in form of analytical solutions of the Navier equation it can very well cope with the local as well as global deformations present in the images by varying the standard deviation of the Gaussian forces.

B. Morphological reconstruction

Morphological grayscale reconstruction based on geodesic dilation employs two input images. These two images are called marker and mask images. Both images must have the same size and the mask image must have intensity values greater or equal to the marker image. A morphological transformation is applied to the marker image and the gray levels of the resulting image are forced to remain above or below the mask [16].

Let F and G be the mask and the marker images, respectively, defined in the same domain. The unitary geodesic λ-dilation of the marker G with respect to the mask F, $\delta^{(1)}_{F,\lambda}(G)$ can be defined as the point-wise minimum between the mask and the unitary non-flat dilation, $\delta^{(1)}_{\lambda}$, of the marker, which is:

$$\delta^{(1)}_{F,\lambda}(G) = \delta^{(1)}_{\lambda}(G(\mathbf{x})) \wedge \mathrm{F}(\mathbf{x}). \tag{1}$$

The unitary λ-dilation, $\delta^{(1)}_{\lambda}$, represents dilation with a unitary non-flat structuring function, $b(\mathbf{x})$, defined as a weightening function:

$$b(\mathbf{x}) = \begin{cases} -\lambda & \mathbf{x} \in B \\ -\infty & \mathbf{x} \notin B \end{cases} \tag{2}$$

where B represents a structuring element. Therefore, $\delta^{(1)}_{\lambda}$ can be defined as:

$$\delta^{(1)}_{\lambda}(F(\mathbf{x})) = \left\{ F(\mathbf{y}) : F(\mathbf{y}) = \sup\left[F(\mathbf{z}) - \lambda \right], \mathbf{z} \in B_{\mathbf{x}} \right\} \vee F(\mathbf{x}). \tag{3}$$

The geodesic λ-dilation of size n of the marker G with respect to the mask F can be obtained by performing n successive geodesic λ-dilation of G with respect to F:

$$\delta^{(n)}_{F,\lambda}(G) = \delta^{(1)}_{F,\lambda}\left[\delta^{(\mathrm{n}-1)}_{F,\lambda}(G) \right] \tag{4}$$

with $\delta^{(0)}_{F,\lambda}(G) = \mathrm{G}$. The λ-reconstruction of the mask image F from the marker image G can be defined as the geodesic λ-dilation of G with respect to F until stability:

$$\gamma^{r}_{\lambda}(F(\mathbf{x}), G(\mathbf{x})) = \delta^{(k)}_{F,\lambda}(G(\mathbf{x})), \tag{5}$$

where k is $\delta^{(k)}_{F,\lambda}(G) = \delta^{(k+1)}_{F,\lambda}(G)$.

Using the λ-reconstruction, only the pixels of the mask image that match up the marker are reconstructed with the maximum value of the mask. The rest of the connected matched areas are reconstructed with an intensity level that decrease with slope equal to λ.

C. Image inpainting method

The aim of image inpainting is to reconstruct the artifacted portions of the image, in order to restore its unity [23]. Inpainting algorithm used in this paper is partial derivative equations (PDE) based algorithm in which the information is propagated in the occluded area, through isophote lines that cross the edges. In order to obtain the restored image, it is necessary to interleave inpainting steps with a number of anisotropic diffusion steps. The algorithm operates iteratively and creates a family of images, each image representing an improved version of the previous one.

III. RESULTS

To evaluate the performance of the proposed method, three experiments were conducted. In the first experiment, two clinical dental CT scan obtained using different CT scanners were used. In the second and third experiment, a CT chest image and CT hip images, respectively, were used. The data have been saved into DICOM file format. The original slices with artifacts are remarkably affected by noise and strong artifacts such as beam hardening and streaking.

For all conducted experiments, the following registration parameter settings were used: (a) the 2D histogram was generated using 256 bins (256×256 normalized gray values); (b) for the Gaussian forces, a

standard deviation $\sigma_f = 5$ pixel was chosen, and the Poisson ratio was set to $\nu = 0.49$; (c) the regularization parameter was set to 0.01. Following the recommendation proposed in [16], the λ parameter of the morphological reconstruction has been set to 15 in all cases.

Figure 1 and 2 show the performance of the proposed method. The figures depict: (a) slice with an artifacts, (b) a neighboring slice without artifacts, (c) a registered image used as a marker for the morphological reconstruction and (d) a restored image. Slice with artifacts on Fig. 2(a) contains two metal objects positioned closely to each other, so the artifacts are more severe than to those on Fig. 1(a). They obscure more information about the anatomical structures. Restored images (Fig. 1(d) and Fig. 2(d)) show how the proposed method reduces significantly streaking artifacts and noise, preserving the anatomical structures such as teeth, bones and cavities. However, the performance of the method for the second case is slightly decreased and black regions between teeth can be noticed. These black regions are results of beam hardening surrounding the metallic object, and they are not significantly reduces, since the size of structuring element, used in the morphological reconstruction to reduce the streaking artifacts, is not large enough to remove these effects. Increase of the structuring element size would lead to the increase of the image smoothing.

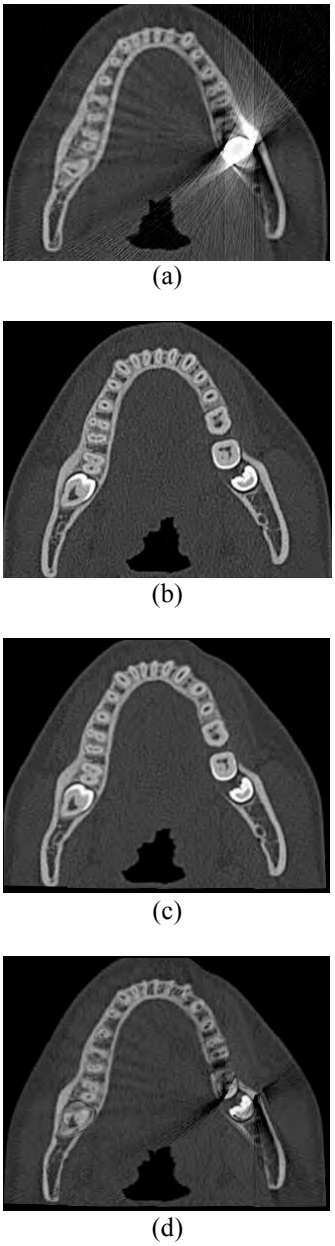
Figure 3(a) depicts the chest image with artifacts caused by metallic suture material. Method provides a substantial reduction of the metallic artifacts and the resulting image (Fig. 3(c)) appear smoother with less accentuated artifacts.

Figure 4 shows the results of metal artifacts reduction for the patient with a hip implant. As can be seen from the Fig. 4(c), significant part of the metal artifact has been reduced, especially the bright artifacts around the implant, but some newly introduced artifacts can be observed. These artifacts are related to the choice of the neighboring slice without artifact: since the metal artifacts caused by the hip implant are propagated through several slices, first neighboring slice without artifacts can be farther away from the slice of interest and can contain somewhat different anatomical strucures.

As in the case of dental CT images, beam hardening is less reduced in the chest image and the hip image, since it presents a wider pattern with a very low values.

IV. CONCLUSION

In this paper an image-based metal artifact reduction method from CT images has been proposed. The method is based on medical image registration and morphological reconstruction. It makes use of information from the neighboring slice of 3D CT images to compensate for data corruption and reveal diagnostic information from the data contaminated by multiple metal objects in the CT scanner field of view. The experimental results show that the method is practical and efficient in reducing the streaking artifacts. Future work will be focused on automatization of the method regarding the neighboring slice selection.



(a) (b) (c) (d)

Figure 1. (a) Slice with an artifacts; (b) neighboring slice without artifacts; (c) registered image used as a marker for the morphological reconstruction; and (d) restored image.

REFERENCES

[1] W.A. Kalender, R. Hebel, J. Ebersberger, Reduction of CT artifacts caused by metallic implants. Radiology vol. 164, pp. 576–577, 1987

[2] A. Mouton, N. Megherbi, K. Van Slambrouck, J. Nuyts, T.P. Breckon , An experimental survey of metal artefact reduction in computed tomography. J Xray Sci Technol vol. 21, pp. 193–226, 2013

[3] A.H. Mahnken, R. Raupach, J.E. Wildberger, B. Jung, N. Heussen, T.G. Flohr TG, R.W. Gunther, S. Schaller, A new algorithm for metal artifact reduction in computed tomography: in vitro and in vivo evaluation after total hip replacement. Invest Radiol vol. 38, pp. 769–775, 2003

[4] P.T. Liu, W.P. Pavlicek, M.B. Peter, M.J. Spangehl, C.C. Roberts, R.G. Paden, Metal artifact reduction image reconstruction algorithm for CT of implanted metal orthopedic devices: a work in progress. Skeletal Radiology, vol. 38, pp. 797–802, 2009



(a)



(b)



(c)



(d)

Figure 2. (a) Slice with an artifacts; (b) neighboring slice without artifacts; (c) registered image used as a marker for the morphological reconstruction; and (d) restored image.

[5] L. Yu, H. Li, J. Mueller, J.M. Kofler, X. Liu, A.N. Primak, J.G. Fletcher, L.S. Guimaraes, T. Macedo, C.H. McCollough, Metal artifact reduction from reformatted projections for hip prostheses in multislice helical computed tomography: techniques and initial clinical results. Invest Radiol vol 44, pp. 691–696, 2009

[6] E. Meyer, R. Raupach, M. Lell, B. Schmidt, M. Kachelriess, Normalized metal artifact reduction (NMAR) in computed tomography. Med Phys vol. 37, pp. 5482–5493, 2010



(a)



(b)



(c)

Figure 3. CT chest image: (a) Slice with an artifacts; (b) neighboring slice without artifacts; and (c) restored image.

[7] E. Meyer, R. Raupach, M. Lell, B. Schmidt, M. Kachelriess, Frequency split metal artifact reduction (FSMAR) in computed tomography. Med Phys, vol. 39, pp. 1904–1916, 2012

[8] M.Bal, L. Spies, Metal artifact reduction in CT using tissue-class modeling and adaptive prefiltering. Med Phys, vol. 33, pp. 2852–2859, 2006

[9] M. Kachelriess, O. Watzke, W.A.Kalender, Generalized multi-dimensional adaptive filtering for conventional and spiral single-slice, multi-slice, and cone-beam CT. Med Phys, vol. 28, pp. 475–490, 2001

[10] (a) Slice with an artifacts; (b) neighboring slice without artifacts; (c) registered image used as a marker for the morphological reconstruction; and (d) restored image.

[11] G. Wang, T. Frei, M.W. Vannier, Fast iterative algorithm for metal artifact reduction in X-ray CT. Acad Radiol 2000, vol. 7, pp. 607–614, 2000

[12] F. Morsbach, M. Wurnig, D.M. Kunz, A. Krauss, B. Schmidt, S.S. Kollias, H. Alkadhi, Metal artefact reduction from dental hardware in carotid CT angiography using iterative reconstructions. Eur Radiol , vol. 23, pp. 2687–2694, 2013

[13] F.. Morsbach, S. Bickelhaupt, G.A. Wanner, A. Krauss, B. Schmidt, H. Alkadhi, Reduction of metal artifacts from hip prostheses on CT images of the pelvis: value of iterative reconstructions. Radiology, vol. 268, pp. 237–244, 2013

[14] Y. Zhang, H. Yan, X. Jia, J. Yang, S.B. Jiang, X. Mou, A hybrid metal artifact reduction algorithm for x-ray CT. Med Phys, vol. 40, no.4, p. 041910, 2013

[15] O. Watzke, W. Kallender, A pragmatic approach to metal artifact reduction in CT: merging of metal artifact reduced images, European Radiology, vol.14, pp. 849-856, 2004

[16] M. Abdoli, M.R. Ay, A. Ahmadian, N. Sahba, H. Zaidi, A Novel Approach for Reducing Dental Filling Artifact in CT-Based Attenuation Correction of PET Data, 4th European Conference of the International Federation for Medical and Biological Engineering, J. Vander Sloten, P. Verdonck, M. Nyssen, J. Hauseisen (Eds.), IFMBE Proceedings 22, pp. 492-495, Belgium 2008

[17] V. Naranjo, R. Llorens, M. Alcaniz, R. Verdu-Mondero, J. Larrey-Ruiz, J. Morales-Sanchez, A new 3D paradigm for metal artifact reduction in dental CT, 18th International Conference on Image Processing (ICIP), pp. 461-464, Brussels, 2011

[18] J.B.A. Maintz, M. A. Viergever, A survey of medical image registration. Med. Im. Anal., 2(1):1:36, 1998

[19] J. Modersitzki, Numerical Methods for Image Registration.Oxford University Press, 2004

[20] W.R. Crum, T. Hartkens, D.L.G. Hill, Non-rigid image registration: Theory and practice. Brit. J. of Radiology, 77(special issue):S140:S153, 2004

[21] A. Andronache, P. Cattin, G. Szekely, Adaptive subdivision for hierarchical non-rigid registration of multimodal images using mutual information. Med Image Comput Comput Assist Interv, 90, 976:983, 2005

[22] A. Serifovic-Trbalic, D. Demirovic, P.C. Cattin, Intensity-based hierarchical elastic registration using approximating splines. International Journal of Computer Assisted Radiology and Surgery (JCARS), 9(1), 463:468, 2014

[23] S. Worz, K. Rohr, New Approximating Gaussian Elastic Body Splines for Lanmdark-Based Registration of Medical Images. Bildverarbeitung fur die Medizin, 206:210, 2006

[24] M. Bertalmío, G. Sapiro, V. Caselles and C. Ballester, “Image Inpainting”, SIGGRAPH '00, 2000

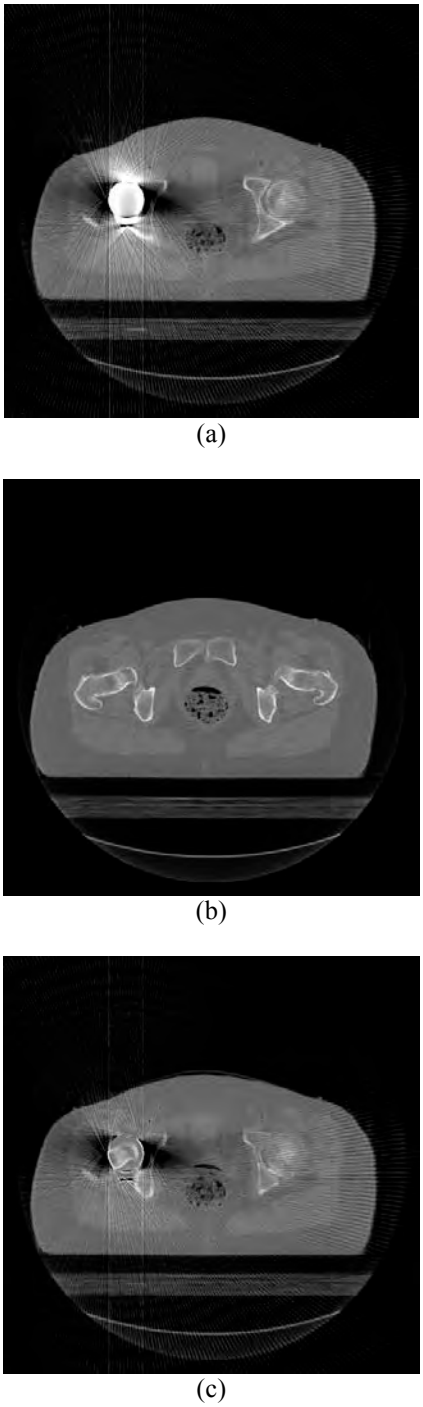


Figure 4. CT hip image: (a) Slice with an artifacts; (b) neighboring slice without artifacts; and (c) restored image.

New algorithm for automatic determination of systolic and diastolic blood pressures in oscillometric measurements

Vojko Jazbinšek
Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
Email: vojko.jazbinsek@imfm.si

***Abstract*—Most automated non-invasive blood pressure measuring devices are based on some empirically derived criteria applied to the oscillometric index, which is defined as a certain characteristic physical property of pressure pulses measured by an inflated cuff wrapped around the upper arm during cuff pressure deflation. In this study, different algorithms for automatic determination of the systolic and diastolic pressures are compared. In addition to the measured pressure pulses, which are one of a typical physical property used for the oscillometric index, some other properties are applied such as a time derivative and a powered short time variance of the pressure data, as well as the audible part of the data measured by a microphone implanted in the cuff (Korotkoff sounds). Beside known empirical algorithms based on characteristic ratios of the oscillometric pulse amplitude, and the maximum and minimum slope of the oscillometric envelope, a new algorithm is introduced based on the presence and absence of oscillometric activity. This algorithm can be applied either to the powered short time variance of the pressure changes in the cuff or to the Korotkoff sounds.**

I. Introduction

Almost all automated non-invasive blood pressure (NIBP) measuring devices use the oscillometric technique [1], [2], which is based on arterial pressure pulses, called oscillometric pulses, that are generated by the arterial blood pressure (BP) pulsation in the cuff that occludes the artery during cuff pressure deflation [3]. Algorithms for automatic determination of the systolic (SP) and diastolic (DP) pressure values are based on some empirically derived criteria applied to the so-called oscillometric index [2], which is defined as a certain characteristic physical property plotted against the baseline cuff pressure. Empirically based algorithms include those based on characteristic ratios of the oscillometric pulse amplitude, and the maximum and minimum slope of the oscillometric envelope [2]. However, the algorithms used for detecting SP and DP are different from one device to another and are not revealed by the manufacturers [1].

The aim of this work was to find a possibly better method of oscillometric data analysis in NIBP measuring devices. A new presentation of an enhanced oscillometric index obtained by a powered short time variance (STV) of the oscillometric data is introduced. This presentation shows significant activity only below SP and above DP values, which simplifies the criteria for the automatic determination of blood pressure values, as in the case of the auscultatory technique where only the presence and absence of pulses has to be detected. A new algorithm for automatic detection of SP and DP based on the presence and absence of activity using the proposed STV oscillometric index is introduced and compared with known algorithms applied to oscillometric index based on peak-to-peak pressure pulses presented in the related study [4].

II. Methods

A. Measurements

Measurements were performed on healthy volunteers with the cuff mounted on the upper arm using the same device (LODE, Groningen, Netherlands) described in the related paper [4]. Briefly, the device has a transducer for recording pressure changes in the cuff and a microphone implanted in the cuff for recording Korotkoff sounds. We used two modes of gradual cuff deflation, linear and exponential, both with an averaged deflation rate of 3 mm Hg/s. For each volunteer we recorded BP at least twice for both deflation modes. Simultaneously, we always measured the BP with a commercial OSZ4 device (Welch Allyn) and the electrocardiogram (ECG) with the HP 78353A device (Hewlett Packard).

For the evaluation of different algorithms we used the same database as in [4] consisted of 92 recordings. The average heart rate was 69±9 (range from 56 to 94) beats per minute and the average measured SP and DP values were 113±10 (89 to 139) and 73±8 (55 to 97) mm Hg, respectively. Measurements were performed on 23 volunteers of age from 20 to 66 (32± 15), 12 males of age from 22 to 66 (39±17) and 11 females of age from 20 to 42 (25±6). For 3 males (all older than 55) we found specific ECG signals (arrhythmia, extra systoles). The remaining 9 males were from 22 to 55 years old (mean age 31±11).

B. Oscillometric index

The oscillometric index is in general defined as an envelope of a certain characteristic physical property of data obtained with the NIBP measuring device. In the previous studies [4], [5], we performed an extensive study of the signals, i.e. pressure changes in the cuff and responses of a microphone implanted in the cuff, which can be recorded during NIBP measurements. We applied several methods, including segmentation into heart beats, fitting, Fourier digital filtering, and a combination of numerical derivative and anti-derivative procedures, to separate the cuff deflation from the arterial pressure pulses. Using

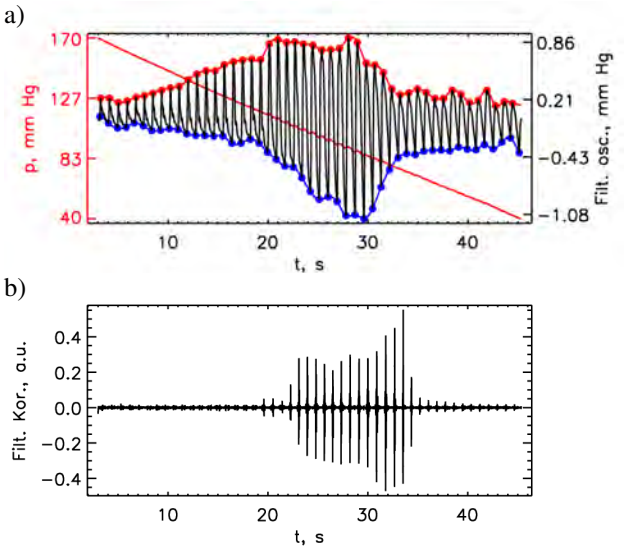


Fig. 1. a) Measured pressure data (left scale) and the arterial pressure oscillations obtained by digital filtering (right scale). Envelopes are defined with maximal (MAX) and minimal (MIN) pressure values for each pulse and are denoted by red and blue bullets, respectively. b) High frequency - audible part of microphone data.

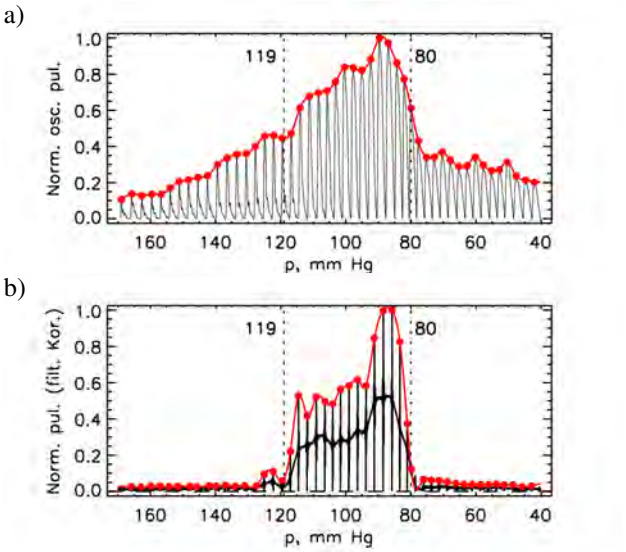


Fig. 2. Normalized oscillometric index obtained from a) the arterial pressure oscillations (Fig. 1a), and b) E_{kp} – the audible part of microphone data (Fig. 1b). Envelope of peak-to-peak oscillation values, obtained by subtraction of MIN envelope from MAX envelope, are displayed as a function of cuff pressure (deflation) level. Vertical dashed lines denote SP (119 mmHg) and DP (119 mmHg) values measured with OS4 device.

digital filtering we separated the microphone data into an audible part (Korotkoff sounds) and a low frequency part, which we found to exhibit a similar property as the time derivative of pressure pulses. We derived several presentations of the oscillometric index using measured signals in our NIBP device.

Using digital filtering, we separated deflation and arterial pressure pulses from the measured pressure data (Fig. 1a) and we extracted a high frequency audible part of the measured microphone data (Fig. 1b), which can be related to the Korotkoff sounds observed in an auscultatory BP measurements [6]. Figs. 2a,b show two types of the oscillometric index defined as envelopes of peak-to-peak oscillation amplitudes from Figs. 1a,b plotted vs. the pressure level in the cuff (deflation).

In this study, another method to represent oscillometric activity is proposed: power enhanced short time variance (STV) defined as the powered normalized variance of the measured data $p(t)$ (see, Fig. 1a) at time subintervals of length Δt

$$V^{\alpha}(t_i) = V_i^{\alpha} = \left\| \sum_k [p(t_k) - \bar{p}_i]^2 \right\|^{\alpha}, \qquad (1)$$

$$t_k \in [t_i - \Delta t/2, t_i + \Delta t/2], \quad \bar{p}_i = \frac{1}{N} \sum_k p(t_k),$$

where N is he number of time points in Δt ($N = \Delta t \cdot f_{\mathrm{s}}$, f_{s} is sampling frequency), t_k is a time point in the i-th Δt and $\bar{p}_i$ is a mean value of measured pressure in that Δt. Fig. 3 shows the influence of the two parameters, Δt and the power α, on the shape of the oscillometric index obtained from (1) (E_{vp} – envelope of normalized peak-to-peak values of V_i^{α}). In the limit case ($\Delta t \to 1/f_{\mathrm{s}}$, $\alpha = 1/2$), V_i^{α} is equivalent to the absolute value of the time derivative of the measured data. For larger Δt, STV gives a measure of the rate of arterial pulse changes in a given time interval Δt, or in other words, how much energy is transferred in Δt between the arterial pulsation in the limb and the cuff wrapped around the limb. Using the heuristic method, we chose one half of an average heartbeat duration (t_{hb}) as the most appropriate value of parameter Δt. The curves in Fig. 3a show that i) above SP and below DP not much power is transferred, ii) that a rapid change of transferred power is observed around SP and DP, and iii) that most of the power is transferred between SP and DP. Using power parameter α one could enhance the oscillometric index. As is demonstrated on Fig. 3b, significant activity is observed only in the interval between SP and DP for $\alpha \geq 4$.

C. Algorithms

Two general types of criteria are known in the literature [2]: one is based on the amplitude reading and the other on the derivative reading. They are called the height-based (HB) and the slope-based (SB) method.

In the HB method, the SP and DP values are determined as a certain height of the normalized index when the pressure level in the cuff is equal to SP and DP. These values are different for each type of oscillometric index [4]. In our data base of 92 recordings, where the reference SP and DP were measured with OSZ4 device, we obtained for the oscillometric index constructed from the arterial pressure pulses averaged height ratios of 0.45±0.07 and 0.70±0.10 for the SP and DP values, respectively. Using these height ratios, we obtained for the case shown in Fig. 1a (121/81) mmHg for (SP/DP) values.

In the SB method, we first found the cuff pressure p_m where E_{xy} reached its maximum height. SP and DP were then defined as the maximum slope of the envelope in the increasing and decreasing part of the envelope. The slope is defined by a derivative of the oscillometric index. However, the oscillometric index represents a non-monotonic

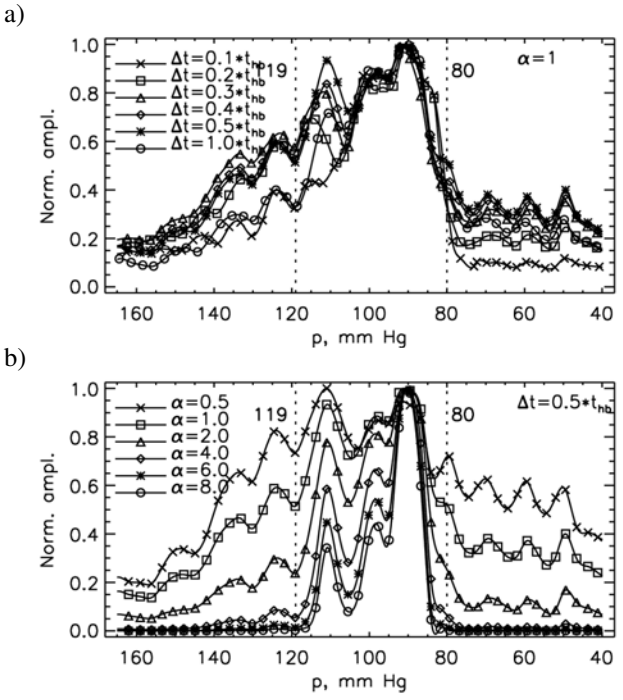


Fig. 3. Influence of a) Δt and b) α on power enhanced STV index E_{vp}) defined as envelope of V^{α} in (1). Vertical dashed lines denote measured SP and DP values in mm Hg.

function due to the beat-to-beat variability of the pulse amplitude, artefacts, etc. Therefore the derivative can have many local extrema. To improve the prognostic value of the SB method, we introduced in [4] some constraints for estimation of SP and DP. Using SB method, we obtained for the case shown in Fig. 1a (114/80) mm Hg for (SP/DP) values.

For the power enhanced STV oscillometric index defined (E_{vp}) defined as an envelope of V^{α} (1), we introduced a new algorithm based on the presence and absence of signal activity. We called it a presence/absence (PA) algorithm. It consisted of two steps. In the first step, threshold values T_{s} and T_{d} for SP and DP were determined from average background activities A_{s} above SP and A_{d} below DP as

$$T_{\mathrm{s}} = (\alpha+1)\cdot A_{\mathrm{s}} \text{ and } T_{\mathrm{d}} = (\alpha+1)\cdot A_{\mathrm{d}}\,, \qquad (2)$$

where α denotes a power in the enhanced normalized index V^{α} (1). Average activities A_{s} and A_{d} were defined as the mean value of E_{vp} for the first N_{b} heartbeats and the last N_{b} heartbeats, respectively. N_{b} was usually set to 3. From T_{s} an onset pressure P_{on} was found as the deflation level at which the envelope reaches T_{s}, and from T_{d} an offset pressure P_{off} was found as the deflation level at which the envelope falls to T_{d}. P_{on} and P_{off} were the first estimates of SP and DP. In the second step, we found maximum slopes S_{s} and S_{d} of the envelope just after P_{on} and just before P_{off}, respectively. A final estimate of SP was defined as the first pressure $p(t)$ level in the cuff below P_{on} that fulfills the following condition:

$$E_{\mathrm{vp}}(t+t_{\mathrm{hb}}) - E_{\mathrm{vp}}(t) > S_{\mathrm{s}}/3\,. \qquad (3)$$

Similarly, DP was determined as the first pressure $p(t)$ above P_{off} that fulfill:

$$E_{\mathrm{vp}}(t) - E_{\mathrm{vp}}(t+t_{\mathrm{hb}}) > S_{\mathrm{d}}/3\,. \qquad (4)$$

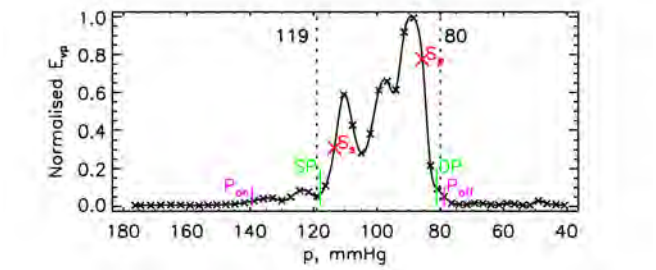


Fig. 4. Demonstration of PA method on E_{vp} envelope. Dashed vertical lines represent measured (PS/PD) values (119/80) mm Hg, vertical magenta lines denote P_{on}/P_{off} (139.4/78.8) mm Hg, red × mark maximum slopes S_{s} and S_{d}, and green vertical lines denote (PS/PD) values (118,81) mm Hg obtained by PA method.

Demonstration of PA method applied to the E_{vp} ($\alpha = 4$, $\Delta t = 0.5\cdot t_{\mathrm{hb}}$) is displayed on Fig. 4. With this method we obtained (118/81) and (120/78) mm Hg for the E_{vp} shown in Fig. 3b and for the E_{kp} constructed from Korotkoff sounds in Fig. 2b, respectively.

D. Evaluation protocol

Two standard protocols for evaluating the accuracy of NIBP devices have been widely used. One was published by the British Hypertension Society (BHS) [7] and the other by the American Association for the Advancement of Medical Instrumentation (AAMI) [8]. The criteria for fulfilling the BHS protocol are that devices must achieve at least grade B (50% of readings falling within 5 mm Hg, 75% within 10 mm Hg and 90% within 15 mm Hg of the mercury standard). The criteria for fulfilling the AAMI protocol are that the test device must not differ from the mercury standard by a mean absolute difference ($|\Delta p|$) > 5 mm Hg or a standard deviation (SD) > 8 mm Hg. The mercury standard refers to SP and DP values obtained by the auscultatory method performed by a trained person using a stethoscope for listening to the Korotkoff sounds and a mercury sphygmomanometer to measure the pressure level in the cuff.

Like in the related paper [4], we used a modified combination of BHS [7] and AAMI [8] protocols for the evaluation of different methods for automatic determination of the SP and DP values. For the oscillometric index E_{pp} based on the arterial pressure pulses (Fig. 2a), we applied the HB and SB algorithms to estimate SP and DP values. For the power enhanced STV oscillometric index E_{vp} ($\alpha = 4$, $\Delta t = 0.5\cdot t_{\mathrm{hb}}$), see Fig. 3b), and for the Korotkoff envelopes (see Fig. 2b), we applied PA algorithm to estimate SP and DP values. As in the BHS protocol we used grading criteria, but instead of using a certain percentage of readings falling within 5, 10 and 15 mm Hg of the reference values, we found the total number of cases falling within 1, 3, 5, 7 and 10 mm Hg of the reference values. Results were classified into 6 grades according to the rule given in Table I. For all estimations of SP and DP, the classification values $\mathrm{V_{SP}}$ and $\mathrm{V_{DP}}$ were found and a combined classification value $\mathrm{V_{SP+DP}}$ was calculated using the following formula biased to the worse of the two classification values:

$$\mathrm{V_{SP+DP}} = \mathrm{round}\left(\frac{\min(\mathrm{V_{SP}},\mathrm{V_{DP}}) + 4\max(\mathrm{V_{SP}},\mathrm{V_{DP}})}{5}\right). \qquad (5)$$

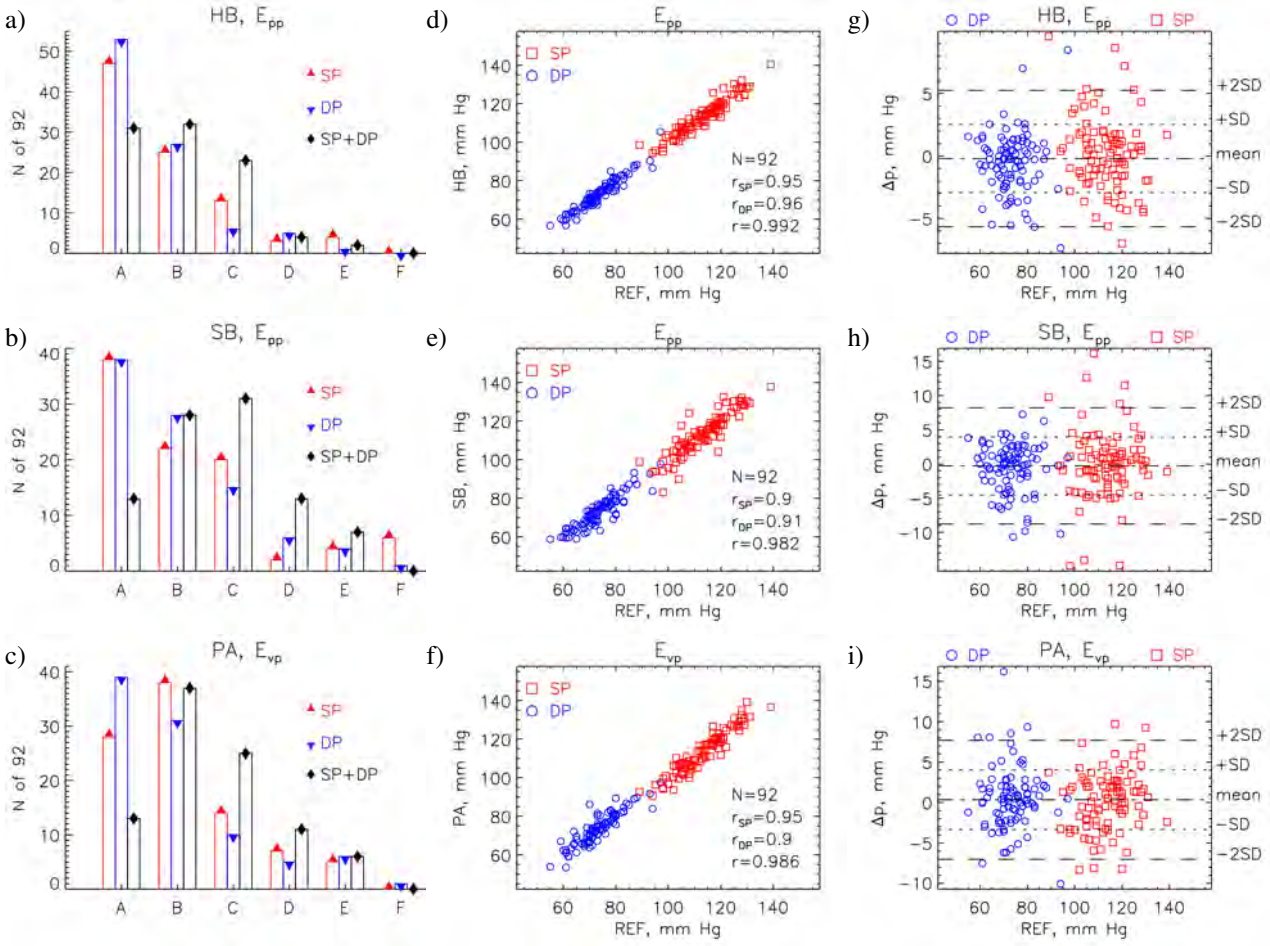


Fig. 5. Bar charts of classification results for the a) HB and b) SB methods applied to E_{pp}, and c) PA method applied to $E_{vp}(\alpha = 4)$. Each bar in the plot corresponds to the total number of cases classified in a given group (A, B, ..., F) defined in Table I. Results for the SP, DP and combined SP+DP are denoted by an red upward triangle, blue downward triangle and black diamond, respectively. Comparison between the SP (labeled by red squares) and the DP (labeled by blue circles) values measured with the OSZ4 device and corresponding SP and DP values obtained by different methods. Scatter plots for d) HB, e) SB, and f) PA methods are displayed, respectively. Linear regression correlation coefficient is denoted by r_{SP}, r_{DP} and r for the SD, DP and combination of both SP and DP. Bland-Altman plots for g) HB, h) SB, and i) PA methods are displayed, respectively. Dash-dotted, dotted and dashed line styles denote mean, mean $\pm$ SD and mean $\pm$ 2SD values for the combination of both SP and DP (see, Table II for values of shown quantities).

TABLE I. CLASSIFICATION OF RESULTS

Grade	Value	description	absolute difference
A	1	excellent	0 or 1 mm Hg
B	2	very good	2 or 3 mm Hg
C	3	good	4 to 5 mm Hg
D	4	approximate	6 to 7 mm Hg
E	5	bad	8 to 10 mm Hg
F	6	fail	more than 10 mm Hg

As in the AAMI protocol, we calculated $\Delta p \pm SD$ but we also included a calculation of the mean pressure difference Δp, which gives the average shift of the estimation regarding the measured reference. In addition, we performed a linear regression fit of estimated vs. measured pressure values, which gives the linear correlation (r).

III. RESULTS AND DISCUSSION

Figs. 5a-c shows bar charts of classification results obtained by the HB and SB methods applied to E_{pp} envelopes and by the PA method applied to E_{vp}. Table II displays some quantitative evaluation parameters for the cases shown in Fig. 5, such as C - the total number of cases classified as good (up to 5 mmHg difference between the estimated and the measured SP and DP), F - the number of failed cases (a difference above 10 mmHg), a – offset, b – slope and r – correlation of the linear regression, $\Delta p \pm$SD – the mean difference and its standard deviation in mmHg, $|\Delta p| \pm$SD – the mean absolute difference and its SD in mmHg, and Δp_m – the maximum difference in mmHg. Figs. 5d-i show comparisons between the SP and the DP values measured with the OSZ4 device and the corresponding SP and DP values obtained by the HB, SB and PA methods. Scatter plots (Figs.5d-f) show that for all these methods we obtained a high correlation ($r > 0.9$) with the measured SP and DP obtained by the OSZ4 device. Bland-Altman plots [9] (Figs.5g-i-f) show that the mean difference Δp between the measured and the estimated SP and DP is below 1 mm Hg for all these methods and most of the results are within $\pm$SD values, which were (see Table II) 2.7, 4.3 and 3.7 mm Hg for the HB, SB and PA methods, respectively. The mean absolute differences were all below 3 mm Hg, i.e. 2.0 ± 1.8, 3.0 ± 3.1 and 2.7 ± 2.5 mm Hg for the HB, SB and PA methods, respectively.

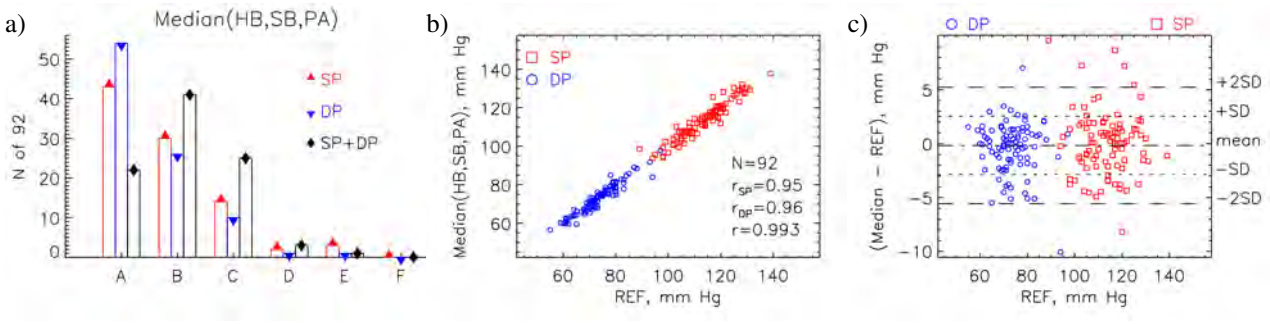


Fig. 6. a) bar chart, b) scatter and c) Bland-Altman plots for median evaluation results, where median values of SP and DP obtained by the HB, SB and PA are compared with corresponding measured SP and DP. See Fig.5 for an explanation of graphs and Table II for values of shown quantities.

TABLE II. SUMMARY OF CLASSIFICATION RESULTS

Method	BP	C	F	a	b	r	$\Delta p \pm$SD	$\lvert\Delta p\rvert \pm$SD	Δp_m
HB	SP	87	0	7.4	0.94	0.952	0.13±3.0	2.2±1.9	10
applied to	DP	87	0	-0.7	1	0.955	-0.54±2.5	1.8±1.7	8
E_{pp}	SP+DP	86	0	-1	1	0.992	-0.21±2.7	2.0±1.8	10
SB	SP	80	6	-2.5	1	0.9	-0.02±4.8	3.3±3.6	16
applied to	DP	81	1	2.6	0.96	0.91	-0.55±3.6	2.6±2.4	-11
E_{pp}	SP+DP	72	0	-1.3	1	0.982	-0.28±4.3	3.0±3.1	16
PA	SP	80	0	-8.9	1.1	0.95	0.18±3.6	2.9±2.2	10
applied to	DP	80	1	1.6	0.98	0.9	0.50±3.8	2.6±2.7	16
E_{vp}	SP+DP	75	0	0.3	1	0.986	0.34±3.7	2.7±2.5	16
median	SP	87	0	2	0.98	0.954	0.05±3.0	2.2±2.0	10
values of	DP	90	0	1.8	0.97	0.955	-0.38±2.4	1.7±1.7	-10
HB,SB,PA	SP+DP	88	0	-0.7	1	0.993	-0.17±2.7	2.0±1.9	-10

C – sum of cases with grades A, B and C, F – number of cases with grade F
a, b, r – offset, slope and correlation of the linear regression, respectively

Results show that all methods fulfill criteria for the BHS protocol [7] (at least 50% of readings falling within 5, 75% within 10 and 90% within 15 mmHg of the reference), as well as for the AAMI protocol [8] (tested device must not differ from the reference by a mean absolute difference ($|\Delta p|$) $>$ 5 mmHg or a standard deviation (SD) $>$ 8 mmHg).

Finally, the median values of SP and DP values obtained by all three methods were calculated. Results displayed in Fig. 6 and bottom part of Table II show that the median values of SP and DP obtained by the HB, SP and PA methods give a best match with the corresponding measured SP and DP. However, the median classification results are only slightly better than the results of the HB method applied to E_{pp} (Fig. 5 and Table II). Detailed statistics of the median values show that in total 39, 34 and 19 cases for SP, and 51, 26, 15 cases for DP were selected from the results obtained by the HB, SB and PA methods, respectively. The algorithm used in the OS4 device is not known to us, but the above results suggest that it is most likely that the HB algorithm is used in this device.

IV. CONCLUSION

We proposed a new presentation of the oscillometric index based on power enhanced STV of measured pressure data, which like in the case of Korotkoff sound in conventional auscultatory method showed significant signal activity only in the region below SP and above DP values. We developed the PA algorithm for automatic determination of SP and DP values. Evaluation study on 92 recordings measured on 23 healthy volunteers showed that the proposed PA algorithm applied to power enhanced STV envelopes gave results comparable to the two known algorithms, HB and SB applied to pressure pulses envelopes. For the evaluation we used a modified protocol which is a combination of the standard AAMI and BHS protocols adapted to our measurements where the measured SP and DP values obtained by the OSZ4 device were used as a reference. Median values of SP and DP estimated by the HB, SB and PA methods gave the best match with the corresponding measured SP and DP.

REFERENCES

[1] Pickering, T. G., Hall, J. E., Appel, L. J., Falkner, B. E., Graves, J., Hill M. N., Jones, D. W., Kurtz, T., Sheps, S. G., and Rocella, E. J., "Recommendations for blood pressure measurement in humans and experimental animals, part 1: blood pressure measurement in humans: a statement for professionals from the subcommittee of professional and public education of the american heart association council on high blood pressure research," *Hypertension*, vol. 45, pp. 142–161, 2005.

[2] Ng, K-G. and Small, C. F., "Survey of automated non-invasive blood pressure monitors," *Journal of Clinical Engineering*, vol. 19, pp. 452–475, 1994.

[3] Marey, E. J., *Pressione et vitesse du sang*. Masson, Paris, France, 1876.

[4] Jazbinsek, V., Luznik, J., Mieke, S., and Trontelj, Z., "Influence of different presentations of oscillometric data on automatic determination of systolic and diastolic pressures," *Anal. Biomed. Eng.*, vol. 38, pp. 774–787, 2010.

[5] Jazbinsek, V., Luznik, J., and Trontelj, Z., "Non-invasive blood pressure measurements: separation of the arterial pressure oscillometric waveform from the deflation using digital filtering," *IFBME proceedings of EMBEC'05*, pp. 3171–3174, 2005.

[6] Korotkoff, N. C., "On the subject of methods of determining blood pressure." *Bull. Imperial Mil. Med. Acad. (St. Petersburg)*, pp. 365–367, 1905.

[7] O'Brien, E., Petrie, J., Littler, W.A., de Swiet, M., Padfield, P.L., and Altman, D., "The British Hypertension Society protocol for the evaluation of blood pressure measuring devices," *J. Hypertens.*, vol. 11(suppl 2), pp. S43–63, 1993.

[8] A. for the Advancement of Medical Instrumentation, *American national standard. Electronic or automated sphygmomanometers, ANSI/AAMI SP 10171992*. Arlington, VA: AAMI, 1993.

[9] Bland, J. M. and Altman, D. G., "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 1, pp. 307–310, 1986.

TGTP-DB – a database for extracting genome, transcriptome and proteome data using taxonomy

K. Križanović [1*#], M. Marinović [2#], A. Bulović [3], R. Vaser [1] and M. Šikić [1,3]
[1] Department of Electronic Systems and Information Processing, University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
[2] Atos IT Solutions and Services d.o.o. Zagreb, Croatia
[3] Humboldt-Universität zu Berlin
[4] Bioinformatics Institute, Singapore 138671, Singapore
# contributed equally
* to whom correspondence should be addressed

Abstract - An enormous potential of metagenomics has caused a significantly increased research activity in recent years. However, discovering pathogens present in a metagenomics sample is still inadequately accurate and very time consuming. Narrowing down a list of potential pathogens would significantly decrease the referent database size and thus improve both accuracy and speed. While publicly available databases enable free access to all their data, the interface is generally unwieldy and impractical for producing concrete reference datasets quickly, making the process either very time consuming or requiring significant programming skills. Relatively simple and sufficiently powerful tools allowing the users to quickly extract desired reference datasets would make metagenomic tools readily available to a larger number of researchers. In this paper we present a set of scripts for downloading publicly available data on taxonomy, genomes, transcriptomes and proteomes and loading it into a relational database. This enables the user to use SQL and other standard relational database tools to search and extract specific datasets. The power and ease of use of the TGTP-DB is demonstrated with a search engine that allows the user to search the taxonomy tree for a desired group of organisms, and download files containing genomes, transcriptomes or proteomes of the organisms in the group.

Availability: The scripts for creating and updating the database are available at:
https://bitbucket.org/marin/mpipe_refdb.
A webpage with a search engine is available at:
http://tgtpdb.zesoi.fer.hr.

I. INTRODUCTION

A field of metagenomics has a large potential in various industrial and medicinal applications. Its basic task is to determine all organisms presents in a metagenomics sample and, if possible, their abundances. This insight can be useful such as in determining microorganisms in soil that facilitate the growth of certain plants, to control fast food preparation processes or to locate an origin of a disease outbreak. However, time to complete a metagenomics analysis can be relatively long because of sequencing times and computationally demanding bioinformatics methods which make the whole process too cumbersome for real-life situations.

Metagenomic analysis tools have been available for several years, such as Pathoscope [1], Kraken [2] and more recently MetaPhlan2 [3] and MetaScope [4]. In many applications, where time factor is critical, it would be very useful to perform sequencing and analysis of a metagenomics sample in the field. The MinION sequencer [5], presented by Oxford Nanopore Technologies (ONT) has already been successfully used in such a way [6]. Recently, a complete workflow has been developed that uses MinION sequencer to obtain an unprocessed sample composition within a few hours [7].

One of the most computationally demanding steps in analyzing a metagenomics sample is mapping reads onto a reference database (a set of reference genomes), in order to determine the composition and the abundance of organisms in the sample. Depending whether the sample contains DNA, RNA or protein reads, the reference database should contain DNA, RNA or protein sequences.

Overly large reference database may cause false positive results due to similar regions in many related or nonrelated organisms. The size of reference database also significantly impacts the duration and in some cases trying to match the sample to a too large reference database might make analyzing the sample prohibitively long. A priori information about the sample could be used to reduce the set of possible organisms and thus reduce the reference database to a manageable size.

Publicly available databases offer everything needed to make a reference sequence database, regardless if sequences analyzed are DNA, RNA or protein. Most of the data is available on NCBI [8] [9]. Bulk downloads are available on the NCBI FTP site ftp://ftp.ncbi.nlm.nih.gov. Ribosome RNA is available for download from the SILVA project [10].

However, even if such additional information is available, the construction of a reference database (or several databases) might prove to be neither easy nor fast. While publicly available databases and search engines allow the user to find and download the data, extraction of a specific set of data is not simple. The interface is often unwieldy and allows downloading sequences one at the time, or in groups unsuitable for the problem at hand.

Automating the downloading process, on the other hand, requires significant programming skill.

To quickly and consistently produce multiple datasets using a complex search mechanism, publicly available data should be structured and loaded into a relational database. Data could then be searched and extracted using tried and tested technologies such as SQL.

To produce relevant reference databases, searches using taxonomic relationships would be essential. First, taxonomy nodes describe hierarchical relationships between organisms, i.e. which organisms are related to each other, and how strong that relationship is. Second, taxonomy names contain various official and unofficial names of organisms. Allowing searching the database using unofficial organism names makes the process more user friendly and doesn't require such precise knowledge from the user.

For example, if a prior knowledge is available that some strain of E. Coli is present in a sample, it would be useful to quickly extract all known strains of E. Coli and then test the sample only against them, thus making the process much quicker and more reliable.

The NCBI taxonomy browser: (www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) allows the user to search the database using taxonomic relationships [11], but it allows downloading only single pieces of information (a single sequence) or a large database containing a huge amount of information (i.e. sequences of all viral organisms).

II. DATABASE

To enable complex searches over publicly available data, the data should be structured and loaded into a relational database. It could then be searched and extracted using tried and tested technologies such as SQL. Thus constructed database could be used, among other thing, to generate datasets for metagenomics analyses. However, the database should be limited to data required for metagenomics analyses and data required to search the database more efficiently (namely taxonomy data).

In this paper, we present a database containing Taxonomy, Genome, Transcriptome and Proteome data, which is called TGTP-DB for short.

The TGTP-DB database uses PostgreSQL (http://www.postgresql.org), an open source Database Management System (DBMS). The data is downloaded in raw format from the NCBI ftp site and, in case of rRNA data, from the SILVA database project [10]. The data is then parsed, structured and loaded into the database.

The set of scripts used to create the database is available at: https://bitbucket.org/marin/mpipe_refdb.

A web search engine demonstrating the use of the database is available at: http://tgtpdb.zesoi.fer.hr.

A. *Datasets*

TGTP-DB database contains the following datasets:

(1) Taxonomy - The NCBI Taxonomy database contains the up to date taxonomy and their mapping to organism names/ranks. This database is available from the NCBI ftp site ftp://ftp.ncbi.nih.gov/pub/taxonomy and is contained in the file named taxdump.tar. TGTP-DB database is loaded with data about taxonomy nodes and data about taxonomy names.

(2) GI – taxonomy mapping – The taxonomy mapping data is downloaded from the NCBI ftp site: ftp://ftp.ncbi.nih.gov/pub/taxonomy. TGTP-DB database is loaded with data about mapping for both nucleotide and proteins sequences.

Together with the taxonomy data, this allows the user to query the database using taxonomic relationships between organisms or groups of organisms, and to search the database using unofficial names from the taxonomy database.

(3) Bacterial, Viral, Fungi and Protozoa genomes and annotations – downloaded from the NCBI ftp site ftp://ftp.ncbi.nlm.nih.gov/genomes/. Contains annotations and CDS sequences for bacterial, viral, fungi and protozoa organisms. Due to the size, CDS sequences are not loaded into relational tables in the database but are kept in a FASTA file in the operating system.

CDS (Coding DNA sequence) represent RNA sequences (translated into DNA) that are transcribed into proteins. They are used for analyzing RNA samples.

(4) Ribosomal RNA – Downloaded from the SILVA project: http://www.arb-silva.de/no_cache/download/. TGTP-DB is loaded with data on small (16S/18S) and large (23S/28S) subunits and with taxonomy mapping data. Due to the size, sequence data is not loaded into relational tables in the database but are kept in a FASTA file in the operating sys-tem.

Ribosomal RNA is most often used for species determination.

(5) NR database - Contains the non-redundant protein sequence database with entries from GenPept, Swissprot, PIR, PDF, PDB, and NCBI RefSeq. Downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/. Due to the size, sequence data is not loaded into relational tables in the database but are kept in a FASTA file in the operating system.

(6) NT database - Contains the nucleotide sequence database, with entries from all traditional divisions of GenBank, EMBL, and DDBJ excluding bulk divisions. Downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/. Due to the size, sequence data is not loaded into relational tables in the database but are kept in a FASTA file in the operating system.

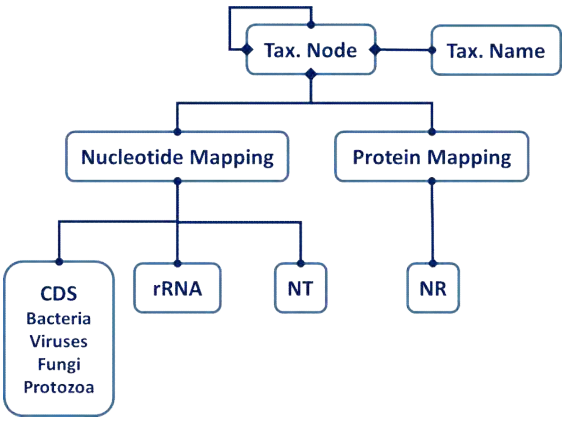


Figure 1: Simplified TGTP-DB database schema

Fig. 1 displays the simplified database schema. Taxonomy nodes represent taxonomic hierarchy and captures taxonomic relationships. Each taxonomy unit can have one or more names. Taxonomic units are connected to various nucleotide sequences (CDS, rRNA, NT databases) using nucleotide mapping and are connected to protein sequences (NR) using protein mapping.

Actual nucleotide and protein sequences are not stored in the relational database, but are left in the FASTA files on the disk. The reason for this is that, to the best of our knowledge, modern DBMSs do not provide sufficient and concrete support for handling sequence data or storing them efficiently. In the event that the appropriate support for bioinformatics data and methods becomes available, sequences data could then be moved to the relational database and the option to query sequences by similarity could be provided to the user.

TABLE I. THE CONTENTS OF RELATIONAL TABLES (AND CORRESPONDING FILES) AFTER INITIAL LOADING

Dataset	***# rows***	***Total sequence length*** *(in number of base pairs)*
Taxonomy	Nodes: 1.4M Names: 2M	-
GI-Taxonomy mapping	Nucleotide: 544M Protein: 248M	-
Bacteria	CDS: 8.7M	8.3Gbp
Virus	CDS: 203k	207Mbp
Fungi	CDS: 5.8M	7.9Gbp
Protozoa	CDS: 1.7M	2.5Gbp
rRNA	Sequence: 1.8M	2.8Gbp
NR	Sequence: 235M	29G
NT	Sequence: 34M	111Gbp

Tab. 1 displays the contents of the TGTP-DB database after initial data loading. The table also displays the total sequence length for each table. Even though the sequence data is currently not stored in the database, scripts available at (https://bitbucket.org/marin/mpipe_refdb) also create relational tables for sequence data.

B. Updating the database

A set of scripts used to create the database is also used for updating it. Since there is not enough information for incremental updates, the whole TGTP-DB database is created from scratch each update. Datasets are updated from the same sources they were created. To make the system usable during update, datasets are updated one by one. While one dataset is updated, others can be used. I.e. while rRNA dataset is being updated, CDS, NT and NR datasets can be queried using the taxonomy dataset.

Tab. 2 displays database loading times for different datasets. Of all CDS datasets, bacterial one is the largest and is given as representative.

Currently, the database is updated arbitrarily, when the administrator estimates that the original datasets have sufficiently changed since the previous update. In the future, automated updating process will be implemented. To keep the data relatively current and at the same time keep the downtime within reasonable limits, we propose that each dataset is updated on monthly bases.

TABLE II. DATASET DOWNLOAD AND DATABASE LOADING TIMES

Dataset	***# rows***
Taxonomy	2min 51sec
GI-Taxonomy mapping	5h 11min
Bacterial	2h 5min
rRNA	12min 30sec
NR	17h 12min
NT	4h 25min

C. TGTP search engine

The use of TGTP-DB database is demonstrated through a simple web search engine. It allows the user to search the TGTP-DB database using taxonomy and to specify which sequence databases to search. It then retrieves the sequences that satisfy the search parameters, stores them in one or more files, and offers them to the user for download. **Fig. 2** shows the workflow used by the TGTP web search engine.

The first step is searching the database using taxonomy. The user can specify the search text, taxonomy rank and search depth. The text is then compared against taxonomy names, looking not only at official scientific names, but also at unofficial names. The taxonomy nodes are searched recursively using a PostgreSQL recursive query. The taxonomy search produces a list of taxonomy units, each with its own taxonomy identifier (tax_id), which are presented to the user.

In the second step the user selects sequence databases that will be searched. The search engine requires a list of taxonomy identifiers which are either copied from the first

step or supplied by the user. The user can thus jump right away to the sequence database search if he has saved a list of taxonomy identifiers from a previous search. Selected sequence databases are then searched and a list of sequence identifiers is produced.

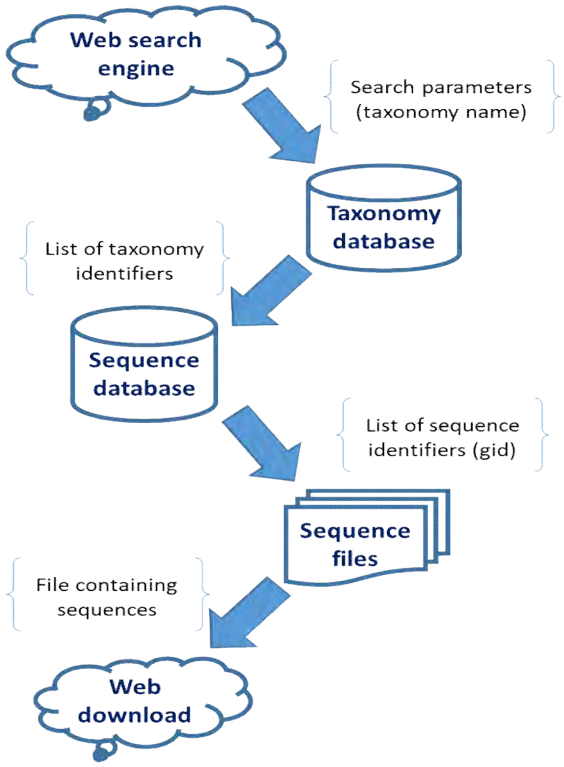


Figure 2: Searching the TGTP-DB database and downloading the resulting sequence files

In the third step, a python script searches the FASTA files corresponding to the selected sequence databases, extracts sequences with given sequence identifiers, compiles them into FASTA files and offers them to the user for download.

Sequences are grouped thematically. One file for genomes (NT dataset), one file for transcriptome (CDS and rRNA datasets) and one file for proteome (NR dataset).

III. CONLUSION

The speed and accuracy of metagenomics sample analysis can be significantly improved if prior knowledge of sample composition is available. Such knowledge can be used to reduce the size of the reference database the metagenomics sample is mapped against.

This paper presents a set of scripts that download a chosen subset of publicly available DNA, RNA and protein sequence data together with taxonomy data, extracts it and loads it into a relational database. The taxonomy data can be used to search the database and extract a desired set of sequences that could later be used, for example, as a reduced reference dataset. The practical use of this approach is demonstrated through a simple web search engine.

While the database and web search engine described in this paper demonstrate the validity of such an approach, a number of issues remain to be addressed in the future.

The compilation of a large dataset still takes a significant amount of time (measured in hours). This time could be shortened by optimizing the database. For example, the tables containing rRNA, NT and NR datasets could be segmented similar to CDS data. Since taxonomy search will usually return related organisms, this will allow us to look only at a part of sequence data in steps 2 and 3. Other methods of optimization will also be considered.

Placing sequence data within the database will be given additional consideration, as will exploring the possibility of introducing some measure of support for sequence data in PostgreSQL.

Also, automating the database updating process will be completed.

FUNDING

This work has been supported in part by Croatian Science Foundation under the project UIP-11-2013-7353 "Algorithms for Genome Sequence Analysis".

REFERENCES

[1] O. E. Francis, M. Bendall, S. Manimaran, C. Hong, N. L. Clement, E. Castro-Nallar, Q. Snell, G. B. Schaalje, M. J. Clement, K. A. Crandall, and W. E. Johnson, "Pathoscope: species identification and strain attribution with unassembled sequencing data.," *Genome Res.*, vol. 23, no. 10, pp. 1721–9, 2013.

[2] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments.," *Genome Biol.*, vol. 15, no. 3, p. R46, 2014.

[3] D. T. Truong, E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, and N. Segata, "MetaPhlAn2 for enhanced metagenomic taxonomic profiling," *Nat. Methods*, vol. 12, no. 10, pp. 902–903, 2015.

[4] B. Buchfink, D. H. Huson, and C. Xie, "MetaScope - Fast and accurate identification of microbes in metagenomic sequencing data," pp. 1–12, 2015.

[5] C. L. C. Ip, M. Loose, J. R. Tyson, M. de Cesare, B. L. Brown, M. Jain, R. M. Leggett, D. A. Eccles, V. Zalunin, J. M. Urban, P. Piazza, R. J. Bowden, B. Paten, S. Mwaigwisya, E. M. Batty, J. T. Simpson, T. P. Snutch, E. Birney, D. Buck, S. Goodwin, H. J. Jansen, J. O'Grady, and H. E. Olsen, "MinION Analysis and Reference Consortium: Phase 1 data release and analysis," *F1000Research*, vol. 4, Oct. 2015.

[6] A. L. Greninger, S. N. Naccache, S. Federman, G. Yu, P. Mbala, V. Bres, D. Stryke, J. Bouquet, S. Somasekar, J. M. Linnen, R. Dodd, P. Mulembakani, B. S. Schneider, J.-J. Muyembe-Tamfum, S. L. Stramer, and C. Y. Chiu, "Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis.," *Genome Med.*, vol. 7, no. 1, p. 99, 2015.

[7] S. Juul, F. Izquierdo, A. Hurst, X. Dai, A. Wright, E. Kulesha,

R. Pettett, and D. J. Turner, “What’s in my pot? Real-time species identification on the MinION,” *bioRxiv*, p. 030742, 2015.

[8] N. R. Coordinators, G. E. Om-, M. Viewer, and S. Read, “Database resources of the National Center for,” *Nucleic Acids Res.*, vol. 43, no. November 2014, pp. 1–12, 2014.

[9] D. a Benson, I. Karsch-Mizrachi, K. Clark, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank.,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D48–53, Jan. 2012.

[10] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, “The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools,” *Nucleic Acids Res.*, vol. 41, no. November 2012, pp. 590–596, 2013.

[11] S. Federhen, “The NCBI Taxonomy database.,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D136–43, 2012.

Development and Perspectives of Biomedical Engineering in South East European Countries

Almir Badnjević[1,2,3,4], Lejla Gurbeta[1,2]
[1]Verlab Ltd. Sarajevo
[2]International Burch University Sarajevo
[3]Faculty of Electrical Engineering University of Sarajevo
[4]Faculty of Medicine University of Sarajevo,
Bosnia and Herzegovina
almir@verlab.ba , lejla@verlab.ba

Abstract - In respect with growing population, chronical disease management and aging population, the medicine and health care have drastically changed over the past decades. Engineering became more involved in medicine resulting in development of new discipline, Biomedical Engineering.
As new engineering solution for problems in medicine regarding therapy, diagnosis and treatment have emerged, the need for new interdisciplinary educational curriculums evolved.
Today, South East Countries are focused on developing new educational curriculums in Biomedical Engineering, following the models established in other European and United States Countries, and recognizing the new discipline – Biomedical Engineering as a professional discipline.
International Federation for Medical and Biological Engineering, supported by United Nations (UN), gives support for national societies that are focused on Biomedical Engineering.
In this article, overview of development and perspectives of Biomedical Engineering in South East European (SEE) countries is given, with focus on Bosnia and Herzegovina.

I. INTRODUCTION

Even though Biomedical Engineering (BME) is relatively new profession, researchers have discovered evidence that people have combined engineering methods for solving problems in medicine centuries, even thousands years ago. In 2000, German archeologists uncovered a 3,000-year-old mummy from Thebes with a wooden prosthetic tied to its foot to serve as a big toe [1], suggesting that it could be the oldest known limb prosthesis. The first idea of medical devices emerged in ancient Egypt as hollow reeds were used for looking and listening to the internal goings on of the human anatomy [2]. In 1816, Rene Laennec set the base for invention that led to today's ubiquitous stethoscope [2]. From there, medical devices have developed in accordance to technological advancements and needs of growing population. Miniaturization of diagnostic and therapeutic devices, their lower power consumption, portability and appearance of microprocessor enabled their applications in practically all branches of medicine. This led to miniaturization of all devices from X-ray imaging, cardiac pacemaker, antibiotic production technology, artificial kidney, computerized tomography and magnetic resonance imaging devices, up to genomic sequencing and micro-arrays, positron emission tomography and image-guided surgery from the last decades, telemedicine and telemetry application based on information – communication technologies (ICT).

In the modern health care system, engineering is needed for overcoming challenges in human health protection, disease prevention, treatment and rehabilitation of patients through introducing sophisticated technology in everyday use, covering a wide range of products, from simple bandages to the most sophisticated life - support equipment and prosthesis. The needs of population changed over time and people became more dependent on technology. To meet the demands of newly established lifestyle in respect to health and wellbeing of population, a progressive shift occurred in industry, so different specialty areas within the field of biomedical engineering were established, such as bioinstrumentation, biomaterials, biomechanics, systems physiology, clinical engineering, rehabilitation engineering and others.

Beside the sustainable and renewable energy sources, biomedical engineering is one of the most rapidly growing branches of industry in the developed world [3]. In 2012, Biomedical Engineer was ranked as top one profession according to the CNN analysis with average growth rate of 6 – 20% [4]. The U.S. Bureau of Labor Statistics (BLS) counted about 20,080 BME jobs in 2014, and projected a 27% increase in positions through 2022 [5]. In Europe, BME is a key sector for European competitiveness represented with €100 billion-market size, with over 20,000 companies or 575,000 jobs [6]. Up to 2015, the development of this profession followed the predicted rate and the BME sector is becoming strategic interest to Europe and other developed countries.

As for education, the earliest academic curriculums began to take shape in the 1950s. There are many undergraduate degree courses in conventional engineering subjects – particularly Electrical, Mechanical and Chemical – which contain BME options or electives. Biomedical applications appear to be particularly popular in Electrical Engineering, with an emphasis on Biomedical Electronics, instrumentation, and signal and image analysis.

Today, in the world, there are also specialized education curriculums focused on BME application, providing their student's theoretical knowledge accompanied with practical experience and raising

interdisciplinary professionals that can work in different working environments, from health care institutions, research centers, quality assurance, industry and development. The profession of biomedical engineer is a registered profession in many countries and in those where it is not a registered profession, biomedical engineers are considered to work in an interdisciplinary area.

II. BIOMEDICAL ENGINEERING IN SOUTH EAST EUROPEAN COUNTRIES – EDUCATION AND RESEARCH

The background for development of Biomedical Engineering in South East European countries, like in the rest of the world, were faculties of engineering and medicine, and later on the industry, recognizing the advantages of interdisciplinary professionals for applied research and development.

The framework for introducing Biomedical Engineering in all of these countries is the same. First, an excellent individual researcher made a breakthrough in this area motivating the development of research groups and laboratories gathering engineers and medical professionals. Then, specific courses were established as part of current educational curriculums in engineering and medicine. Later, research center, professional organizations and societies for medical and biological engineering were established with a goal to create new educational programs in Biomedical Engineering, to recognize the Biomedical Engineering as a professional discipline and to connect the educational institutions and industry.

BME in SEE countries started developing in late 1950s. One can say that pioneer of Biomedical Engineering in SEE countries is Croatian scientist and researcher, Prof. Ante Šantić, who was the first European BME researcher to receive the IEEE EMBS Career Achievement Award in 2003 for his contributions and leadership in creating Biomedical Engineering courses in Europe [7-8]. He was the Head of Electronics Laboratory at the Institute of Electrical Engineering in Zagreb, specialized in electronic instrumentation and medical instrumentation which lead the Institute to become the leading manufacturer of electroencephalographs in Central Europe [3].

The educational curriculum in Biomedical Engineering in Croatia started at the University of Zagreb, Faculty of Electrical Engineering in early 1950s. In the following 1960s, teaching activities in area of Biomedical Engineering started in Italy at University of Padova, University of Naples and Polytechnic of Milan. Since then many other courses and educational curriculums have started in other Italian Universities, so in 2000 about 150 undergraduate courses in Biomedical Engineering were active at 19 Italian Universities, while PhD courses in Bioengineering and postgraduate courses were available in 9 universities [9]. As for Croatia, curriculums in BME leading to a university degree are offered by University of Zagreb, University of Split and University of Osijek, on graduate and postgraduate level. However, these do not result in a BME degree, but a diploma in Electrical Engineering [9].

First educational programs in Biomedical Engineering in Slovenia were introduced at the Faculty of Electrical Engineering, University of Ljubljana. The foundation of its Laboratory of Medical Electronics and Biocybernetics in the 1960s represents the beginning of research and higher-education activities in BME [10]. The pioneer of BME in Slovenia is Prof. Lojze Vodovnik. With his research on functional electrical stimulation he was able to join efforts of medical doctors and engineers to work on problems related to rehabilitation of disabled patients [11]. Extended stand-alone program of BME in Slovenia was established in the study year 2012/2013. Today, Biomedical Engineering in Slovenia is well recognized and relatively well funded on a national level with a number of excellent research groups and good perspective as the governmental institution has recognized the potential of economic development of country through BME. With goal to establish Slovenia on the global map as an incubator of highly innovative companies that build the most advanced biomedical devices in the world, EU regional Competency Center for Biomedical Engineering (CC BME) was established. It is coordinated by the Laser and Health Academy, and supported by the European Regional Development Fund and Slovenia's Ministry of Higher Education, Science and Technology [12].

Among countries in Balkan, Serbia is the only one offering their students an educational curriculums that leads to diploma in BME. The first one is BME curriculum at the Technical University of Novi Sad which offers 3 cycle system from BSc to PhD level. The national society for BME organizes many activities in order to raise awareness and interest to this area among young scientist and researchers.

In Montenegro, Prof. Radovan Stojanović, founder of Laboratory for Medical Electronics at the Faculty for Electrical Engineering, University of Montenegro introduced this area. He was coordinator of several EU projects dealing which this topic including Tempus project BioEMIS [13], which resulted in first model for education of biomedical engineer in Montenegro. Former and outgoing TEMPUS projects in curriculum development significantly contributed to the development of BME and Medical Informatics (MI) in Western Balkan Region. Several study curriculums, modules and courses are developed as well as laboratories.

The situation in other SEE countries, that are still in transition (Bosnia and Herzegovina, Albania, Kosovo) is the same because of poorly developed highly sophisticated industry which raises the need for BME by itself. Also, healthcare systems are orientated toward classical approach instead of usage of technology for improvement of quality of medical care and life support.

In SEE countries, national societies are formed with goal to organize national and international scientific events in BME, establish studies in the field of BME, publish a professional journal and create models for implementing health reforms a BME in industry. The first of these national societies was established in late 1967, in Italy. Associazione Italiana di Ingegneria Medica e

Biologica (AIIMB) exist today and it is a member of International Federation for Medical and Biological Engineering (IFMBE).

After the fall of the Federal Republic of Yugoslavia and Yugoslavian Society for Medical and Biological Engineering, Slovenian Society for Medical and Biological Engineering, The Croatian Medical and Biological Engineering Society were established. These societies are now the members of the IFMBE.

The Society of Biomedical Engineering and Medical Physics of Serbia was founded in late 1980s as a voluntary scientific - professional national organization consisting of qualified biomedical engineers and medical physicists with university degrees and working in alliance with medical staff in hospitals, universities, or research institutes [14].

III. BIOMEDICAL ENGINEERING IN BOSNIA AND HERZEGOVINA

Field of BME in Bosnia and Herzegovina is very young and still in the first stage of development. Investment in BME has been recognized as a base for progress in many fields, from cutting health care expenses, better quality and accuracy in diagnosis and patient treatment, increase of medical therapies reliability and, finally, improved populations' life quality [15].

In 2014, group of young engineers and doctors, led by Almir Badnjević, established Bosnia and Herzegovina Medical and Biological Engineering Society (DMBIUBIH). This society brings together a large number of university professors, doctors, engineers and technicians in the field of medical, biological, electrical, mechanical engineering, pharmacy and other related science fields. The main goal of this society is to support development of Biomedical Engineering in Bosnia and Herzegovina by helping the government in determining the strategic direction related to Biomedical Engineering and supporting researches in this area [16]. In order to raise the awareness and to bring this area closer to all other researchers the Society members organized the 1st Conference of Medical and Biological Engineering in Bosnia and Herzegovina (CMBEBIH 2015) in Sarajevo, in March 2015 [17]. The conference raised tremendous interest of the scientific and wider community in Bosnia and Herzegovina and in the region. As first ever BME Conference in the region, the CMBEBIH was endorsed by the International Federation for Medical and Biological Engineering (IFMBE), the largest society based biomedical engineering organization in the world which proves that development of this area in our county has great support.

The education in Biomedical Engineering is mostly provided through specific and elective courses in conventional engineering and medical curriculums on Public Universities. The previously mentioned group of people developed on International Burch University department for Genetic and Bioengineering with specific scientific curricula that is unique in SEE Countries.

Today about 50% of all treatment and diagnostic methods did not exist 10 years ago and worldwide preventive care of medical devices for health care industry creates high costs. On suggestion of the members of DMBIUBIH and to set a base for management of technology in healthcare system and to reduce the percentage of deaths, as well as injuries caused by defective medical devices, National Metrology Institute (NMI) in Bosnia and Herzegovina [18] introduced 10 groups of medical devices in legal metrology [19]. In this way, independent inspection body specialized in medical devices provides technical support for healthcare institutions and professional staff, and ensure the reliability and safety of the Medical Devices, prevents bad practice, where manufacturer and distributors supervised their own equipment. For the purpose of independent inspection of medical devices unique laboratory Verlab was established [19]. In the first year of their work more than 50% of all public and private healthcare institutions verificated their medical devices.

IV. STRATEGIC FRAMEWORK FOR DEVELOPMENT OF BIOMEDICAL ENGINEERING IN SEE COUNTRIES

Since BME is interdisciplinary field which involves integration of social sciences and technological innovation, from concept to production, the key factors of strategic framework for BME development are governmental, academic and industrial support and cooperation.

Strategic framework relay on developing BME educational curriculums should be based on intellectual integration of medicine, electrical engineering, mechanical engineering, IT, natural sciences at the undergraduate, graduate and doctoral level. In order to develop quality educational curriculums national societies for biomedical engineering can serve as support body. These societies are places where academic staff meet professionals so they can recognize the needs and provide solutions that can help in shaping the educational curriculums by the industry, healthcare needs, etc. Staff and students mobility at the prestigious world BME universities and institutes should be very important because they can result in domestic BME engineers with wider knowledge and cultural horizons and can ensure long term modernization of healthcare system. Funding's for these study curriculums can be achieved through funding's from international grants and projects. Effective cooperation between national societies in SEE countries would solve the problem of inferiority of BME in this region through organizing scientific events and international meetings, organizing exchange for researchers in SEE countries, raising interest for publishing activities and working on proposals and implementation of educational curriculums.

At the same time with development of the educational curriculums, BME industry in SEE Countries must develop. Only harmonization between university and industry can position BME on proper way. For that purpose well-coordinated, government's strategic orientation is essential. Systematization of working places and coordination with WHO guidelines are necessary in order to recognize biomedical engineer as professional orientation and to establish a place for them in work market.

Previously mentioned good practices from Bosnia and Herzegovina in introducing BME in the law and healthcare system, can be helpful to all other SEE Countries. Members of BME National Societies should recognize the needs in healthcare sector and suggest the solutions to the government.

V. CONCLUSION

BME is a prosperous field what has been proven with continuous development rate in recent 10 years, in United States of America as well as in Europe and in the rest of the world.

In SEE Countries, it is necessary to use suggested unique strategic framework or to create a new one, that will ensure favorable business environment for BME developments. Firstly, new educational curriculums should be developed, that will raise interdisciplinary professionals. To connect educational institutions with labor market appropriate legislative framework is a must.

The advantages of SEE Countries, their common language and background, can be used in capacity building for educational curriculums which can provide teachers training and retraining, student exchanges, empowerment of market in Biomedical Engineering and raising competitiveness in respect to other, more developed European countries.

For each SEE country, investing in BME can attract foreign investments and develop new capacities such as production of new drugs, new laboratories and other medical and engineering facilities as well as modernize existing facilities.

The most urgent issue in SEE that influences the development of BME is the recognition of this discipline within national qualification frameworks. Namely, the countries which recognize BME as profession are very rare. It is one obstacle for young student to choose BME as their future occupation. Thus, this problem should be addressed as soon as possible by joint efforts.

LITERATURE

[1] J. H. Lienhard, "A 3000-year-old toe", The Engines of Our Ingenuity, available at: http://www.uh.edu/engines/epi1705.htm

[2] EAMBES, European Alliance for Medical and Biological Engineering, "Engineering for Health: A partner in the development of a knowledge-based society for the benefit of European healthcare".

[3] R. Magjarević, I. Lacković, "Biomedical Engineering – past, present, future", Automatika, 2011, pp. 5-11.

[4] CNN Money statistics, 2012, available at: http://money.cnn.com/pf/best-jobs/2012/snapshots

[5] U.S. Bureau of Labor Statistics (BLS), available at: http://www.bls.gov/

[6] C. James, L. Peccia, "Profile: The Impact of engineering in biology and medicine: the biomedical engineer (BME)", Adjacent Government, 2015.

[7] M. R. Neuman, "Ante Šantic: A Memory", IEEE Eng in Med and Biol, 28, 2009, pp 91.

[8] E. Tonkovic, "In Memoriam Prof. Emer. Ante Šantic", Au- ´ tomatika, 2008, pp. 101-104.

[9] Biomedical Engineering education in Europe – Status Reports, IFMBE, 2005.

[10] D. Miklavčič, S. Reberšek, A. Kralj. Lojze Vodovnik kot človek, učitelj in znanstvenik. Predgovor v (D. Miklavčič, T. Kotnik, G. Serša, ur.) *Lojze Vodovnik – Zbrana dela*. Ljubljana, Univerza v Ljubljani, Fakulteta za elektrotehniko, 2003.

[11] Biomedical Engineering in Slovenia D. Miklavcic, University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia

[12] Competency Center for Biomedical Engineering Slovenia, available at: http://www.bmecenter.com/en/

[13] BioEMIS - Studies in Bioengineering and Medical Informatics, Tempus projects, availbale at: http://projects.tempus.ac.rs/en/project/813

[14] "Launching IFMBE into the 21st Century: 50 Years and Counting", edited by H. Voigt, R. Magjarevic, Springer, 2014.

[15] S. Mustoo, L. Gurbeta, "Turning the Challenge into Opportunity – A Strategic Framework for the Biomedical Engineering Development in Bosnia and Herzegovina", Folia Med. Fac. Med. Univ. Saraeviensis 2015, pp. 18-22.

[16] Bosnia and Herzegovina Medical and Biological Engineering Society (DMBIUBIH), available at: http://dmbiubih.org/

[17] The the 1st Conference of Medical and Biological Engineering in Bosnia and Herzegovina (CMBEBIH 2015), available at: http://cmbebih2015.dmbiubih.org/

[18] Badnjević, Gurbeta, Džemić, Bošković, "Measurement in medicine – Past, present, future". Folia Medica; 2015: 50(1).

[19] Badnjević, Gurbeta, Bošković, Džemić, "Medical devices in Legal Metrology". 4th Mediterranean Conference on Embedded Computing MECO 2015, Budva, Montenegro, Volume: ISBN 978-1-4799-8999-7

Clustering of Heartbeats from ECG Recordings Obtained with Wireless Body Sensors

A. Rashkovska*, D. Kocev ** and R. Trobec*
* Department of Communication Systems, Jožef Stefan Institute, Ljubljana, Slovenia
** Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia
{aleksandra.rashkovska, dragi.kocev, roman.trobec}@ijs.si

Abstract - Long-term electrocardiographic (ECG) recordings can be beneficial for detection and diagnosis of heart diseases, in particular arrhythmias. A wireless multi-function biosensor that measures a potential difference between two proximal electrodes on the skin enables monitoring of vital functions, like heart rate, respiration and muscular activity. It can thus make long-term ECG measurements while users are performing their everyday duties and activities. These measurements are significantly longer and heterogeneous than the measurements performed in a controlled hospital environment. Consequently, their inspection for identification of different groups/clusters of heartbeats, either manual or computer supported, is obligatory. In this paper, we propose a method for automatic clustering of heartbeats from an ECG obtained with a wireless body sensor. We use state-of-the-art data mining methods for time series clustering - hierarchical agglomerative clustering in conjunction with dynamic time warping distance. The results show that the proposed methodology is robust and comparable to the classical Holter algorithms and therefore worth to be further evaluated.

I. INTRODUCTION

Long-term electrocardiographic (ECG) recordings are intended to help in detection or diagnosis of heart diseases. These measurements are significantly longer and heterogeneous than the measurements performed at a controlled hospital environment. Consequently, manual inspection of these recordings in order to identify different groups/clusters of heartbeats (that can be used for better describing the health status of the subject) is a tedious, hard and expensive job. An alternative is to use computational techniques for automatic clustering, like neural networks [1], data mining methods for clustering of ECG features [2] or time series clustering methods [3]. In this paper, we address the task of automatic heartbeat clustering using data mining methods for time series clustering.

Today the most standard ECG device used in medicine is the well-known 12-lead ECG, where wires are connected to the electrodes placed on 10 locations of the body. More long-term ECG recordings are usually acquired by a Holter monitor where reduced number of electrodes are connected with wires to a small portable recorder that acquires continuous ECG measurement throughout several days. On this type of equipment, more electrodes are used to obtain the signal, in which one of them serves as a reference for the others. Usually one to three wired leads are utilized for automatic tasks of ECG analysis. Related studies for ECG clustering utilize this type of measurements. However, to align with future trends in e-health, in this paper, we utilizes a unique ECG data acquisition that promises better future for long-term ECG recordings – unobtrusive measurement with a single ECG lead without wires. A wireless multi-function biosensor that measures a potential difference between two proximal electrodes on the skin enables monitoring of vital functions - heart activity and respiration [4].

In this paper, we propose a method for automatic clustering of heartbeats from an ECG obtained with a wireless body sensor. We exploit state-of-the-art data mining methods for time series clustering, namely hierarchical agglomerative clustering in conjunction with dynamic time warping distance. The obtained clusters will then be analyzed and used to derive meaningful heartbeat categories.

The remaining of this paper is organized as follows. Section II introduces the fundamental steps in heartbeat clustering and describes the selected methods for each step; Section III presents and discusses the results; and finally, Section IV concludes the paper.

II. METHOD

An automatic system for clustering of heartbeats from ECG recordings can be divided into four steps (see Fig. 1) as follows: 1) ECG data acquisition; 2) ECG signal preprocessing; 3) heartbeat segmentation; and 4) heartbeat clustering. The selection of methods for each of the four steps can have a crucial implication on the final result – the heartbeat clusters and consequently, the medical interpretation of such results.

A. ECG Data Acquisition

The data analyzed in this work has been obtained from unobtrusive ECG body sensor (dimensions: 2x9cm, weight: 14 g) with two electrodes at the distance of 8 cm. The placement of electrodes on the chest can be easily

The authors Aleksandra Rashkovska and Roman Trobec acknowledge the financial support from the Slovenian Research Agency under the grant P2-0095. Dragi Kocev is supported by the European Commission through the project MAESTRA - Learning from Massive, Incompletely annotated, and Structured Data (Grant number ICT-2013-612944).

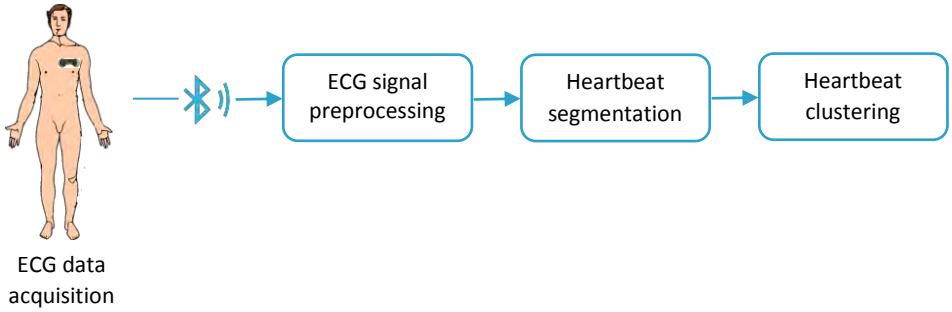


Figure 1. Diagram of the heartbeat clustering system.

fine-tuned to maximize the quality of the ECG recording [5]. Note that besides ECG, other features can be extracted from the measured potential, e.g., muscle activity and respiration [6]. In addition to the ECG, the sensor can also sense the information about the measurement conditions, e.g., movements and temperature, thus providing information that allow for ambient intelligence [7, 8]. A moderate sampling rate of 125 Hz with 10 bit analogue/digital converter is used as an optimum between medical value and amount of generated data. The sensor has a long autonomy (up to 7 days) [9], a low power wireless connection (BT4) to a Smartphone or other personal device, and a corresponding software for visualization and interpretation of measurements.

The ECG measurements obtained with the sensor are suitable for medical use, e.g., screening of patients with potential heart rhythm disturbances, reconstruction of the standard 12-chanell ECG [10], Holter-like investigations for long-term monitoring of patients after cardiac surgery, monitoring in cardio oncology [11], etc. The device can support solutions to every-day problems of the medical personal in hospitals, health clinics, homes for the elderly and health resorts. Its exceptionally lightweight design allows for unobtrusive use also during sports activities or during exhaustive physical work. Currently, there are only a few similar devices, but none has competing services on such advanced level [12-15]. A prototype of the wireless multifunctional body sensor is shown in Figure 2.



Figure 2. Prototype of the wireless multifunctional body sensor

B. *ECG Signal Preprocessing*

The preprocessing step usually concerns the noise removal from the ECG signal. The simplest and most widely used technique for reducing noise in ECG signals is the application of finite impulse response (FIR) digital filters [16]. Such filters perform well for the attenuation of known frequency bands, such as the noise coming from the electrical network (50 Hz or 60 Hz). However, filters must be applied with caution, since they can significantly distort the morphology of the ECG signal and make it unusable for cardiac diagnostic. Therefore, most state-of-the-art methods for ECG signal analysis do not even apply preprocessing on the signal.

In order to investigate the potential influence of the preprocessing task on the final clustering results, we examine scenarios with preprocessed and not preprocessed (raw) ECG signal. The applied preprocessing technique on the ECG signal is a low pass FIR filter with cutoff frequency set to 50 Hz.

C. *Heartbeat segmentation*

The heartbeat segmentation step refers to the detection of R peaks and generation of heartbeat segments from the ECG signal. Heartbeat segmentation methods have been studied for more than three decades [17, 18, 19, 20] and vary in complexity, i.e., from very simple methods to more elaborated ones. The method used for R peak detection in this work has been described and evaluated in [21]. The cut-off time for the segments is 0.2 *s* before the R peak time.

D. *Heartbeat Clustering*

Clustering is concerned with grouping objects into classes of similar objects [22]. Given a set of examples (object descriptions), the task of clustering is to partition these examples into subsets, called clusters. The goal of clustering is to achieve high similarity between examples within individual clusters (intra-cluster similarity) and low similarity between examples that belong to different clusters (inter-cluster similarity). Consequently, the notion of similarity (and conversely distance) is of a crucial importance in clustering.

A cluster is typically represented with a prototype, such as the mean/centroid or the medoid of the examples belonging in the cluster. The aim is to obtain compact clusters that have a low (intra-cluster) variance. Methods like k-means clustering or hierarchical agglomerative clustering (HAC) can be used to find sets of clusters with low intra-cluster variance and low inter-cluster similarity.

Here, we focus on HAC as a clustering method. HAC builds a hierarchy of clusters in a “bottom-up” approach: at the lowest level each example belongs in its own cluster and then pairs of clusters are merged together as one moves up the hierarchy. Typically, merging of the examples is done in a greedy manner [23].

The result of hierarchical clustering can be visually presented using a dendrogram (see Fig. 3). The dendrogram allows for cutting the hierarchy in two major

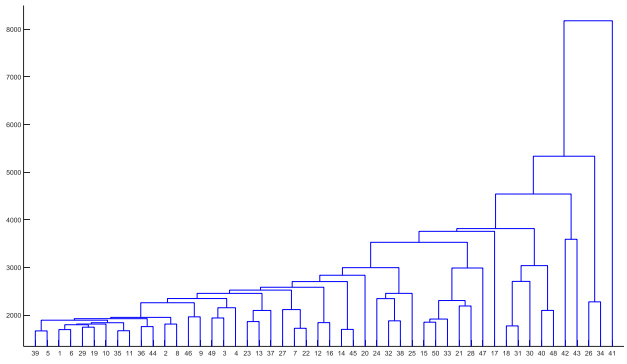


Figure 3. Dendrogram of hierarchical cluster tree

ways: one could specify the desired number of clusters or the maximal distance between two clusters to allow merging. This allows for closer inspection of the clusterings observed in the hierarchy of clusters.

A very important part of the HAC method is the definition and calculation of the distance between two clusters, i.e., the linkage function. There are several types of linkage that could be selected specifically for a given task. These different linkage types include maximum or complete linkage (maximum distance between any two examples from these clusters), minimum or single linkage (minimum distance between any two examples from the clusters), mean or average linkage (average distance between the examples of the two clusters), centroid linkage (the distance between the centroids/prototypes of the two clusters) and the decrease in variance for the cluster being merged (Ward's criterion). Here, we used average linkage for constructing the hierarchy of clusters.

In this work, we consider the heartbeat signals as time series, hence, for use methods for HAC clustering of time series data. More specifically, we use distance-based clustering methods that rely on the definition of an appropriate distance measure between two time series. There are several distance measures for time series that could be applied in this context. When selecting a distance, we should consider the properties of the data at hand and the properties of the specific distances.

For example, if the time series have equal length, then one can apply some standard distance measures such as the Euclidean distance or the Manhattan distance. These measures could, however, not be appropriate in this context because (1) they assume that the time series are synchronized, and (2) they mainly capture the difference in scale and baseline. Considering the properties of the heartbeat signal analyzed here, use of the above distances is not recommended: the duration of the heartbeats varies and the main interest is not as much the amplitude of the signal as its shape.

Considering the specific properties of the heartbeat signals (different length, shape over amplitude), we need to resort to more sophisticated distances that are able to capture the desired dynamics of the heartbeat signals. A distance that supports the required flexibility is the Dynamic Time Warping (DTW) distance [24]. The DTW distance "stretches" the time axis to obtain better matching between two time series. It accomplishes this by assigning multiple values of one of the time series to a single value of the other. This means that a time series that is delayed with respect to another one or otherwise temporally stretched (but maintains the approximate magnitude) will still be considered similar to the original series. If there is no stretching of the time axis, the DTW distance behaves as the Euclidean distance. Finally, the DTW distance can be used for time series of different length and not synchronized, which makes it a very flexible distance function.

In this work, we instantiate the parameters for the clustering algorithms as follows. As mentioned, we used HAC with average linkage implemented in Matlab. The DTW distance was calculated using the guidelines provided by Ratanamahatana and Keogh [25]: when calculating the distance between two time series, the window size is set to 10% of the length of the longer time series. For efficient calculation of the distance matrix between all of the heartbeats, we used the FastDTW implementation provided by Salvador and Chan [26].

III. RESULTS

A. Experimental setup

The ECG data analyzed in this work consists of four 30-minute measurements from a single volunteer, acquired during different activities. Therefore, each measurement is clustered separately. The statistics of the measurements are given in Table 1.

The dendrogram from the HAC clustering was cut to obtain 50 clusters. This high value is a precaution measure since the clustering method will cluster outliers in separate clusters (with very few examples), leaving the number of meaningful clusters reasonable. A cluster is considered meaningful if it contains more than 3 examples. The number of 3 was selected considering the length of the investigated measurements.

TABLE I. ECG MEASUREMENTS STATISTICS
Sinus – Normal sinus beat; SVES – Supra Ventricular Extra Systole; AF – Atrial Fibrilation

ID	Activity	Mean BPM	# Beats	# Clusters [a]		Clustered beats [%]		Type of rhythm
				Raw ECG	*Filtered ECG*	*Raw ECG*	*Filtered ECG*	
M1	walking	91.0	2727	21	23	98.57	98.72	**Sinus (< 1% SVES)**
M2	sitting	82.1	2462	20	19	98.62	98.33	**87%AF + 12%Sinus + 1%SVES**
M3	laying	57.3	1717	11	11	97.50	97.44	**82 % Sinus + 18 % SVES**
M4	sitting	64.1	1917	19	21	98.07	98.33	**Sinus (< 0.1% SVES)**

a. With more than 3 examples

B. *Raw vs. Filtered ECG*

The results showed that the numbers of meaningful clusters from raw and filtered ECG measurements do not differ much (see Table 1). Also, the percentage of clustered heartbeats from raw and filtered ECG does not differ much for each measurement. The percentage of clustered beats is around 98% for all measurements, regardless of the preprocessing. Visual inspection of the clusters showed that similar clusters are often generated from the raw and filtered ECG. Therefore, it can be concluded that the used method works well on the ECG signal without any preprocessing done beforehand. However, clusters from filtered ECG tend to be more coherent.

C. *Clustered ECG interpretation*

For visual inspection of the results, the clusters obtained from the filtered ECG measurements M1, M2, M3 and M4 are given in Fig. 4, 5, 6 and 7, respectively. The clusters are order by their size (number of examples). Only the clusters containing more than 3 examples are presented. For each cluster, the medoid heartbeat of the cluster is represented with black curve, accompanied with maximum 20 examples/heartbeats closest to the medoid represented with gray curves. The medoid is an example of the cluster whose average distance to the other examples is the smallest or, in other words, it is the example that is the closest to the "average" example in the cluster. The medoid is considered as a cluster representative/prototype in applications where mean is not attainable, which is the case in heartbeats clustering with DTW.

Results show that the clustering method identified different cardiac events in the measurements, even in the measurement acquired while walking. In all cases, the clusters with the highest number of examples clearly correspond to a normal ECG beat, except for the measurement M2. Namely, the P wave is not present in the ECG beats from the biggest cluster from M2, indicating atrial fibrillation. In M2, also some other smaller clusters, e.g., C11 and C5, can be identified as atrial fibrillation. The rest of the clusters (12%) from M2 belong to normal sinus beats, except for C6, C12 and C24, which indicate SVES (Supra Ventricular Extra Systole). The results for measurement M3 clearly identify clusters with SVES beats (around 18% - C2, C3, C16 and C25), namely clusters with the QRS looking very different from normal (the QRS has spikes). A small percent of SVES beats (< 1 %) is also present in M1, in particular, identified by C30, C20, C17, C38 and C11. Measurement M4 is clearly sinus ECG rhythm – all the clusters indicate normal ECG beat, except for a very small number of artifacts/outliers (C29 and C14).

The cardiac events identified in each measurement are also summarized in Table 1. For the ultimate medical interpretation of the clusters, a medical professional should be consulted.

IV. CONCLUSION

In this paper, we propose a method for automatic clustering of heartbeats. Unlike most related studies for ECG clustering which utilize long-term ECG measurements from 12-lead ECG or Holter monitor, here we utilize a unique ECG data acquisition - noninvasive measurement with a single channel of bipolar ECG without wires. After appropriate preprocessing and segmentation of the ECG measurements, we perform hierarchical agglomerative clustering using Dynamic Time Warping (DTW) as distance measure.

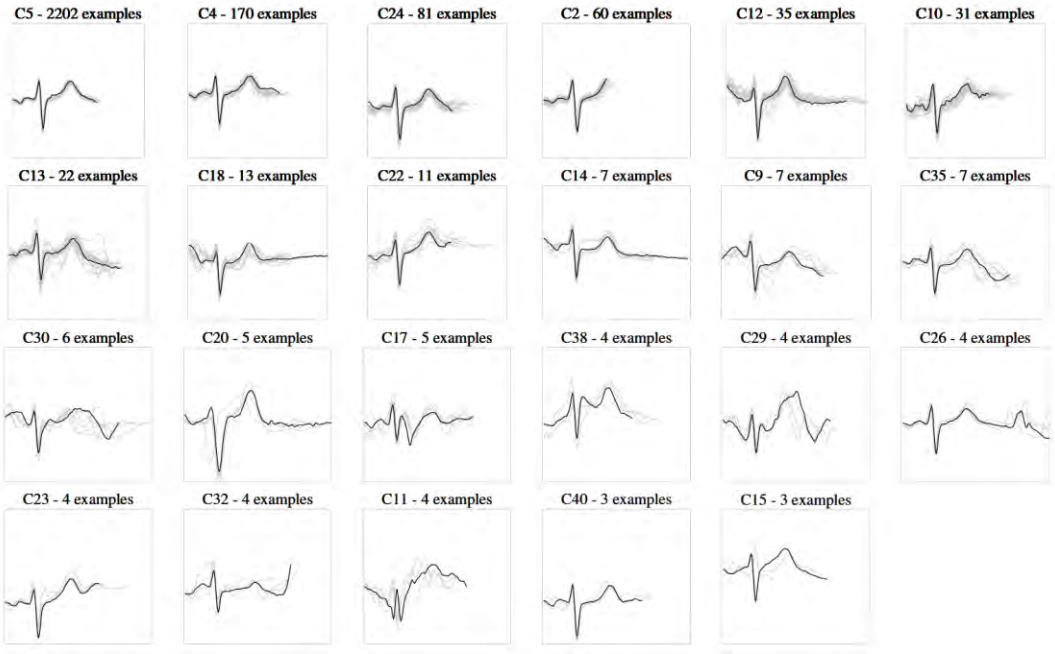


Figure 4. Clusters from filtered measurement M1. Black line represents medoid, gray lines represent maximum 20 closest examples to the medoid.

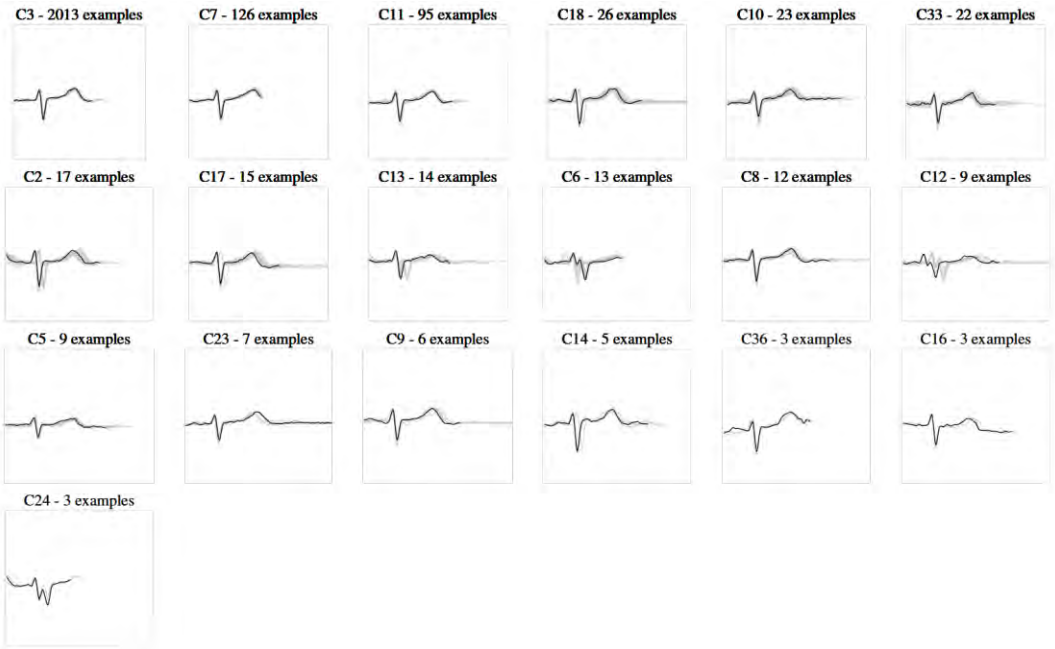


Figure 5. Clusters from filtered measurement M2. Black line represents medoid, gray lines represent maximum 20 closest examples to the medoid.

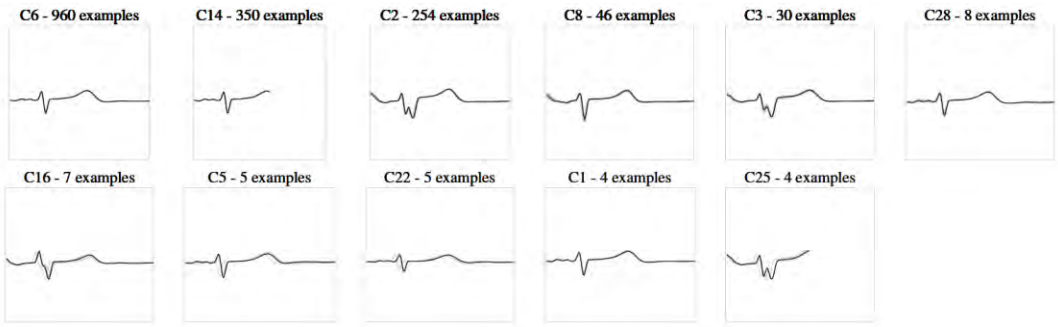


Figure 6. Clusters from filtered measurement M3. Black line represents medoid, gray lines represent maximum 20 closest examples to the medoid.

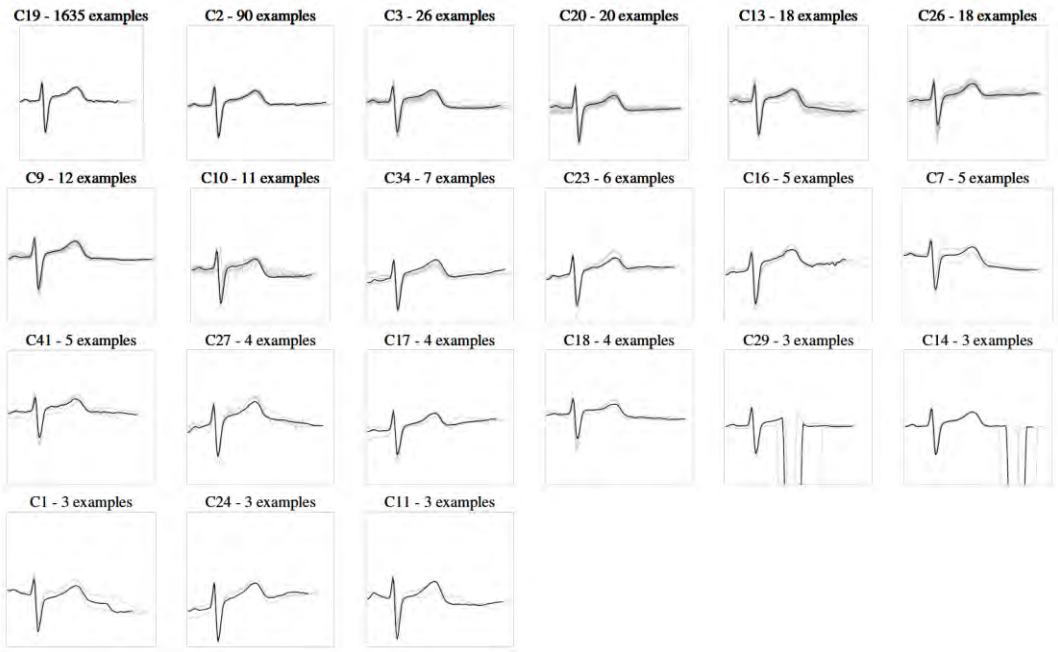


Figure 7. Clusters from filtered measurement M4. Black line represents medoid, gray lines represent maximum 20 closest examples to the medoid.

The results show that we were able to identify meaningful clusters, confirming that the clustering from a differential ECG can provide an interesting insight into the medical condition based on the measurements. However, these methods should be accounted only as assistance for the medical professional and should not be trusted blindly.

In further work, we plan to investigate different segmentation techniques and their potential influence on the final clustering results. Furthermore, other distances over time series considering the specifics of heartbeat signals will be designed and explored, as well as other clustering methods, including k-medoids and predictive clustering trees. The obtained clustering results will then form the base for annotation of the ECG beasts as a pre-step for the future task of ECG beats classification.

REFERENCES

[1] M. Lagerholm, C. Peterson, G. Braccini; L. Edenbrandt; and L. Sornmo, "Clustering ECG Complexes Using Hermite Functions and Self-Organizing Maps," IEEE Transactions on Biomedical Engineering, vol. 47, pp. 838–848, August 2002.

[2] A.Vágner, L. Farkas, and I. Juhász, "Clustering and Visualization of ECG Signals," Advances in Intelligent and Soft Computing, vol. 101, Editors: D. Dicheva, Z. Markov, E. Stefanova, Eds. Springer, 2011, pp. 47-51.

[3] J. R. Annam, S. S. Mittapalli, and R. S. Bapi, "Time series Clustering and Analysis of ECG heart-beats using Dynamic Time Warping," Proceedings of INDICON 2011, Annual IEEE India Conference, December 16-18, 2011, pp. 1-3.

[4] A. Rashkovska, I. Tomašić, K. Bregar and R. Trobec, "Remote Monitoring of Vital Functions - Proof-of concept System," Proceedings of MIPRO 2012, 35th International Convention, May 21-25, 2012, Opatija, Croatia, pp. 446–450.

[5] I. Tomašić, S. Frljak, and R. Trobec, "Estimating the universal positions of wireless body electrodes for measuring cardiac electrical activity," IEEE Transactions on Bio-medical Engineering, vol. 60, pp. 3368–3374, December 2013.

[6] R. Trobec, A. Rashkovska, and V. Avbelj, "Two proximal skin electrodes - a respiration rate body sensor," Sensors, vol. 12, pp. 13813–13828, October 2012.

[7] R. Trobec, V. Avbelj, and A. Rashkovska, "Multi-functionality of wireless body sensors," The IPSI BgD transactions on internet research, vol. 10, pp. 23–27, January 2014.

[8] H. Gjoreski, A. Rashkovska, S. Kozina, M. Luštrek, M. Gams, "Telehealth using ECG sensor and accelerometer," Proceedings of MIPRO 2014, 37th International Convention, May 26-30, 2014, Opatija, Croatia, pp. 270–274.

[9] K. Bregar and V. Avbelj, "Multi-Functional Wireless Body Sensor Analysis of Autonomy," Proceedings of MIPRO 2013, 36th International Convention, May 20-24, 2013, Opatija, Croatia, pp. 322–325.

[10] R. Trobec and I. Tomašić, "Synthesis of the 12-lead electrocardiogram from differential leads," IEEE Transactions on Information Technology in Biomedicine, vol. 15, pp. 615–621, July 2011.

[11] J. M. Kališnik et al., "Mobile health monitoring pilot systems," Proceedings of IS 2015, 18th International Multiconference Information Society, October 9-12, 2015, Ljubljana, Slovenia, pp. 62–65.

[12] E. Fung et al., "Electrocardiographic patch devices and contemporary wireless cardiac monitoring," Front Physiol., vol. 6, May 2016, doi: 10.3389/fphys.2015.00149.

[13] M. Hernandez-Silveira et al. "Assessment of the feasibility of an ultra-low power, wireless digital patch for the continuous ambulatory monitoring of vital signs," BMJ Open, vol. 5, May 2015, doi:10.1136/bmjopen-2014-006606.

[14] M. Etemadi et al., "A Wearable Patch to Enable Long-Term Monitoring of Environmental, Activity and Hemodynamics Variables," IEEE Transactions on Biomedical Circuits and Systems, in press.

[15] I. H. Hansen, K. Hoppe, A. Gjerde, J. K. Kanters, and H. B. Sorensen, "Comparing Twelve-lead Electrocardiography with Close-To-Heart Patch Based Electrocardiography," Proceedings of IEEE EMBC 2015, 37th Annual International Conference, August 25-29, 2015, Milan, Italy, pp. 330–333.

[16] P. Lynn, "Recursive digital filters for biological signals," Medical and Biological Engineering and Computing, vol. 9, pp. 37–43, 1979.

[17] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," IEEE Transactions on Biomedical Engineering, vol. 32, pp. 230–236, March 1985.

[18] V. X. Afonso, W. J. Tompkins, T. Q. Nguyen, and S. Luo, "ECG beat detection using filter banks," IEEE Transactions on Biomedical Engineering, vol. 46, pp. 192–202, February 1999.

[19] . C. Yeh and W. J. Wang, "QRS complexes detection for ECG signal: The difference operation method," Computer Methods and Programs in Biomedicine, vol. 91, pp. 245–254, September 2008.

[20] O. Sayadi and M. B. Shamsollahi, "A model-based bayesian framework for ECG beat segmentation," Physiological Measurement, vol. 30, pp. 335–352, March 2009.

[21] V. Avbelj, R. Trobec, and B. Gersak, "Beat-to-beat repolarisation variability in body surface electrocardiograms," Med. Biol. Eng. Comput., vol. 41, pp. 556–560, September 2003.

[22] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis," 1st Edition, Wiley-Interscience, 1990.

[23] L. Rokach and O. Maimon, "Clustering methods," in Data mining and knowledge discovery handbook, Springer US, 2005, pp. 321–352.

[24] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition, " in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 1, pp. 43-49, 1978.

[25] C. A. Ratanamahatana and E. Keogh, "Three Myths about Dynamic Time Warping," Proceedings of SDM '05, SIAM International Conference on Data Mining, April 21-23, 2005, Newport Beach, CA, pp. 506–510.

[26] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," Intelligent Data Analysis, vol. 11, pp. 561–580, October 2007.

Heart rate analysis with NevroEkg

Miha Mohorčič* and Matjaž Depolli**
*student at Jožef Stefan Insitute, Department of Communication Systems, Ljubljana, Slovenia
**Jožef Stefan Insitute, Department of Communication Systems, Ljubljana, Slovenia
mohorcic.miha@gmail.com, matjaz.depolli@ijs.si

Abstract – NevroEkg is a computer application for analysis of ECG and related bio-signals, such as breathing and blood pressure. It was made in collaboration between computer scientists, engineers and neurocardiologists. Recently, it has been modified to also support the unconventional measurements of differential ECG, made on wireless wearable gadgets. These wearable gadgets measure ECG a bit differently - with lower resolution, lower sampling frequency, and more noise. These features require modified and additional processing of the ECG signal, which is not required for standard 12-channel ECGs. A novel algorithm is proposed to help the human operator handle beat-detection in novel ECG measurements.

I. Introduction

NevroEkg is a toolset for computer visualization and analysis of multi-channel ECG and related bio-signal measurements for neurocardiologists [1]. It was implemented in a way that makes analysing short and noise-free measurements that were made in controlled conditions relatively easy and automated. Novel wearable battery-powered gadgets that measure ECG [2] are becoming affordable for health monitoring, and are able to produce so far unseen amounts of ECG data per subject. Wearable gadgets presents a novel means of taking ECG measurements though, which differs from the well-established means in many aspects, and produce different measurements. The main characteristics of ECG measurements produced by wearable gadgets are listed below:

- Measurements are performed in **an uncontrolled environment**, on subjects on their everyday lives, performing their routine tasks or jobs.
- Current technology enables **over 3 days of continuous ECG measuring**. In the future, the continuous measurement time will likely extend for over a week [4].
- Since the measurements are performed on active subjects, they are far **noisier** than the measurements made on resting subjects.
- The wearable gadgets measure **single-channel** ECG, sacrificing the benefits of multi-channel measurements for the benefits of simplicity and unobtrusiveness [5].
- Measuring ECG against the common potential is not possible, since the wearable device only uses two electrodes, therefore **the differential ECG** is measured as the difference between voltage on those two electrodes.
- The aim for long battery life imposes limitations on measurement precision. Relatively **low sampling frequency and low sampling resolution** are usually used on wearable gadgets, compared to modern electrocardiography devices. In this work, gadget that samples ECG with 10 bits and 125 Hz is used.
- Allowing for **lossy data transmission** further lowers power requirements of the wearable gadgets and simplifies their hardware and is therefore justifiable. It is reasoned, that the sheer length of measurement more than compensates for the small number of missing samples.

To accommodate processing of large amount of ECG data with the presented characteristics, new algorithms and processing procedures were required. NevroEkg was modified to accommodate the new requirements and provide some automatic and some semi-automatic processing of the measurements. The semi-automated beat detection was tested and evaluated on the Physionet MIT-BIH Arrhythmia Database [3].

Section II describes the process of handling data and extracting information from them. Subsections A, B and C describe process in more detail. In section III the tests and results of the proposed algorithm are presented. Section IV summarises and concludes the presented work.

II. Algorithms and methods

Processing of measurement data is divided into three steps: an automatic conversion of input measurement files into a file format recognisable by NevroEKG, a semi-automatic detection of heartbeats and a manual analysis of the observed irregularities in heart rhythm. First two steps must take the characteristics of the acquired data into account, to make the input to the last step – the heart rhythm – as error-free as possible for the human operator that is observing it. The first step comprises the analysis of the measurement input file, i.e, the estimation of sampling frequency, the missing data detection, and error detection and handling. This is described in details in subsection IIA

Semi automatic detection helps the human operator to quickly process long measurements containing thousands of beats. This work is normally done manually, and to achieve precise results this is still the preferred method. However for long measurements, where relatively large percentage of beats are regular, a quick method of generalized beat detection is a valuable addition. It is presented in subsection IIB

As the last step, human operator can detect grouped

anomalies in the measurement when given an overview of the detected beats. This step still includes human knowledge and intuition. It is important to be done manually, particularly since the subject is mobile during the measurement, and automatic detection could misinterpret the normal changes in the rhythm as an anomaly. We discuss the method of visual analysis in section IIC

A. Data converter

The collected data is encapsulated in a simple text file format, as a stream of packets, where each packet comprises timestamp in nanoseconds, sample counter, and sequence of integer-valued samples with constant size. Currently all packets contain exactly 14 samples. Text file also contains meta information, such as the measurement start time, the identifier of the gadget that performed measurement, the sampling frequency, the multiplier and offset for translating sample values, the patient-specific metadata, and user comments. The NevroEKG proprietary nekg file format is able to include all the given meta information, as well as multiple measurement channels, and multiple event channels, which are derived from the measurement channels. Both NevroEKG and its file format are not, however, designed to hold information about the missing data.

The converter processing can be summed up in the following steps:

1. Reading of all contents of the input file into a structured form – separating timestamps, counters, samples, and metadata.
2. Checking for consistency of timestamps and counters, with automated fixing of errors that could occur during the packet transmission. Packets with nonconformant timestamps are discarded while the deviating sample counters with small errors in value (expected value differs by 1 or two) are fixed to match expected values. Missing data is also logged at this step.
3. Sampling frequency is estimated from the counter samples and timestamps by weighted average of the observed sample rates of all continuous sequences (sequences that do not include any missing data). Linear weighting is used that gives each sequence the weight equal to the length of the sequence.
4. Output in nekg file format is written, using the gathered metadata and samples, with the detected missing samples assigned value 0.
5. Problems encountered in the conversion process are also logged as a comment in the nekg file.

B. Beat detection algorithm

Algorithm is based on numerous first derivative based QRS detectors [6], and is probably most similar to the simplified QRS detection algorithm by Pan and Tompkins [7]. The algorithm is provided an input signal in a time series, called *signal channel*. The proposed automated part of the beat detection consists of the following steps.

1) Low-pass filtering

The signal channel is filtered with a low pass filter, with a user defined cut-off frequency. This filter is implemented with simple triangular convolution vector whose width corresponds to the cut-off frequency. This step removes most of the high-frequency noise emitting from electrical appliances. This step also ensures that the missing data is handled properly, that is, the values that are used to indicate missing data are not used for calculation, and are not smoothed out. On Fig. 1, the input signal processing is shown with the input signal in red and the filtered signal in blue.

2) Derivation

Filtered signal is numerically derived. During the derivation, the algorithm again checks for missing data, and ensures that transitions between measured and missing data are not detected as high derivatives. The derivative of missing data is therefore set to zero, since events cannot be detected in such areas. To simplify the algorithm and to ensure that ECGs of all orientations are handled properly, absolute value of the derivative is used for the beat detection. Thus, there is no need for detection of ECG orientation and signal flipping. A sample absolute value of the derivative is shown in green on Fig. 1.

3) Amplitude analysis

Possible beat locations are then detected using amplitude analysis algorithm. This algorithm searches for the areas of the derivative higher than a given threshold. Such peaks are very likely the peaks in QRS complex, more exactly on the slopes between R and S waves. The distance between such peaks represents a close approximate to the RRI (beat-by-beat R to R interval) value.

At this stage, all the peaks are declared heartbeats and inserted into a newly created event channel. Event channel contains events as ¡*time, value*¿ pairs. For heartbeat events, *time* is defined as the absolute time since the measurement start, and *value* is defined as the time passed since the previous event (RRI).

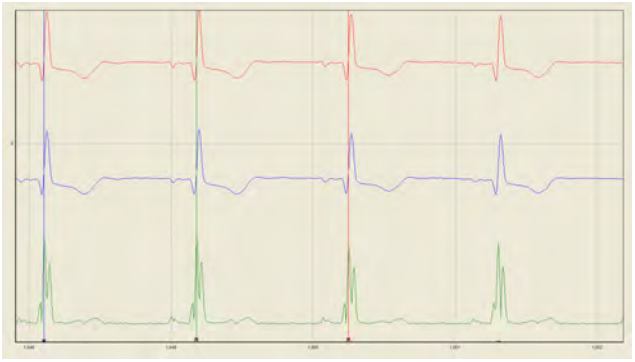


Figure 1. Three steps of the algorithm are shown: red is the input signal, blue is the filtered signal, and green is the absolute value of the first derivative.

4) Extreme events removal

The last stage of the algorithm prunes the extreme events. Possible erroneously detected beats include peaks from noisy areas, and peaks from P and T waves. Physical limitations limit the heart beat frequency, therefore, to remove peaks that are less likely to correspond to a QRS complex, peaks that are not within margins of normal heart beat frequency are removed. This is done by imposing

minimum and maximum thresholds on the value of heartbeat events. Heartbeat events are removed if their RRI is lower than $0.15s$, or higher than $2.5s$. Maximum threshold thus removes the correctly detected heartbeat events, which occur after long intervals of inactivity in the signal, but correctly removes the obviously bad RRI values contained within those heart beat events. Intervals of long inactivity are expected, since they easily arise from simple problems in the measuring procedure and removing a few good heartbeats from the measurement is a good trade-off for removing bad RRI values.

5) Automatic threshold detection

The proposed beat-detection algorithm has only one parameter to tune - the threshold for peak detection from the signal derivative. Since the algorithm is used on very diverse measurements, the threshold should be defined per measurement. To aid the human operator in selecting an appropriate threshold value, the following approach is used.

Process is employing the previously defined steps. Signal is filtered with the low pass filter of 50Hz, and the filter output is derived. Then the program tries to determine the best value of the threshold for the given derived filter output in a loop over a predefined set of candidate values. From experiences gathered by analysing such signals manually, lower bound for candidate values is defined as 1000, and upper as 15500, the values up to 2000 increment by 250, later by 500. For every candidate value, the amplitude analysis and extreme values removal are performed, followed by calculation of statistics. Statistics are gathered from beats: total number of beats, mean heart rate and standard deviation of heart rate.

To accelerate compute times and disregard thresholds that cause the algorithm to find no useful data, computation may stop before reaching the upper bound. When the algorithm detects a very low average heart rate using a certain candidate value for threshold – 15 beats per minute or less – it searches no further, since higher thresholds only further worsen the calculated statistics and higher values are thus deemed invalid by the algorithm itself. The average heart rate is calculated only from the correctly sampled data, with missing samples disregarded.

After the statistics are gathered for all the candidate values in a loop, they are presented to the human operator. A window is shown, with the statistics plotted on two graphs (see Fig. 2 for an example), to help the operator choose the best threshold value. The first graph shows the standard deviation and average heart beat rate, while the second graph shows the number of heart beats detected.

The candidate value, which results in the lowest standard deviation in heart rate if used as the threshold, is offered to the operator as the best candidate (see the vertical lines on the graphs on Fig. 2). After the operator confirms the selection of the threshold value, the detected beats are converted into an event channel, to be displayed underneath the input signal on screen, as seen on Fig. 3.

As can be seen from the added vertical lines on the figure, the detection of beats on the used ECG measurement latched on the slope of Q wave. This is not the same for all the measurements, the detection latches to the point of highest absolute derivative, the point of which is generally within the QRS complex but may vary within it, due to different possible placements of the differential electrodes on the chest.

The automated approach works well for signals, where the subject has normal heart rhythm most of the time, and there is no constant noise of fixed frequency. For such cases, the operator only clicks OK after reviewing the two graphs of statistics.

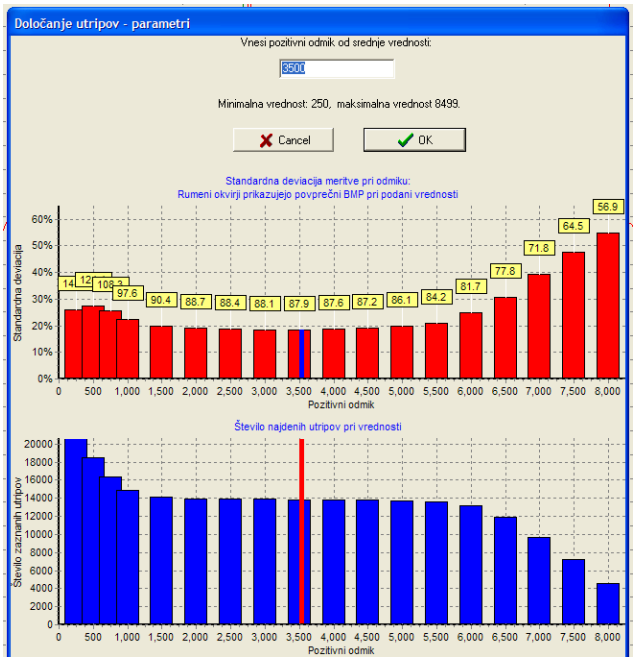


Figure 2. Graph shown to user during beat detection

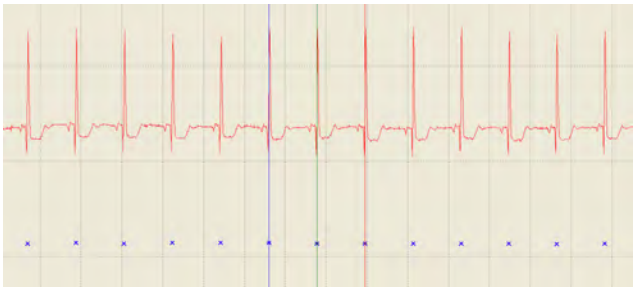


Figure 3. Result of beat detection.

6) Problems of automatic threshold detection

The operator should visually inspect the graphs, focusing on the threshold values on and near to the proposed candidate value. The standard deviation around the proposed value should be low and heart rate in appropriate bounds, while the number of samples should not change drastically around the proposed value. If these conditions are satisfied, one click on the OK button accepts the threshold and shows the beats on the graph (as seen on Fig. 3). However, some heart beat anomalies can cause the analysis to show multiple areas that satisfy those conditions. Fig. 4 is an example where the analysis returned two areas that seem appropriate for the threshold values. On the Fig. 4, the two mentioned areas are encircled in yellow. This particular analysis occurred on a measurement, where almost every second heart beat was irregular and produced very high peaks in the sig-

nal derivative. While the analysis did suggest the value in the proper area in this case, it could also very easily been wrong and suggested a value from the other area. This is why the human operator is included into the algorithm, with the possibility to visually analyse the statistics, to override the suggested threshold and even to repeat the whole beat-detection if the beats are later found to be miss-identified too frequently. The results of the correct threshold selection for this case are shown on Fig. 5.

C. Visual analysis

The presented algorithm is relatively robust, fast, and has good enough accuracy to help the physicians make a basic overview over the collected measurement.

One of the usual methodologies of further ECG analysis is visual inspection of heart rhythm and visualization of problematic sections. Problematic sections are sections of ECG, where rapid change in heart rhythm is seen or a pattern in the heart rhythm changes. An example of whole measurement visualization can be seen on Fig. 6 and of a visualization of a small problematic section on Fig. 7).

Visualisation of the whole measurement provides a global overview of subject's activity and heart rhythm. Several observations can be made from the visual analysis of the measurement overview (Fig. 6).

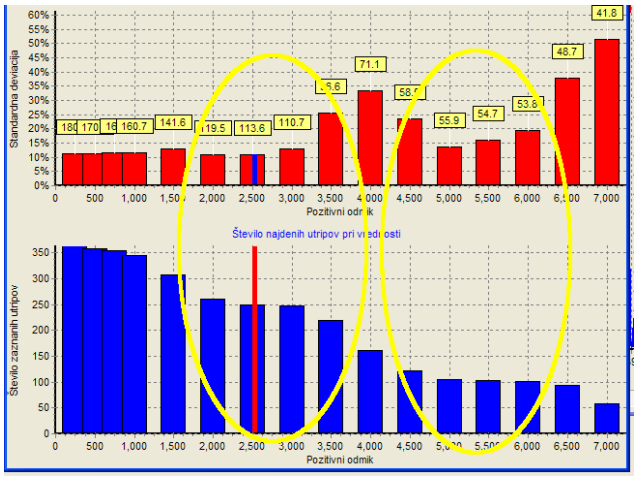


Figure 4. Case of a graph with abnormal heart beat

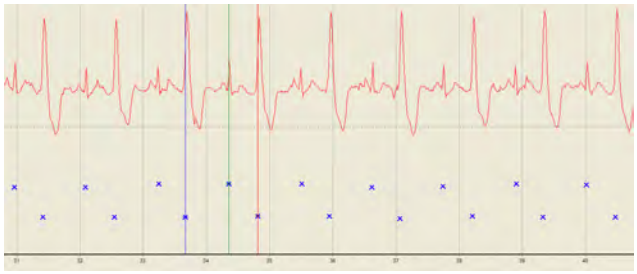


Figure 5. Result of beat detection on abnormal signal

- At the beginning (from 20 to 500 seconds in the measurement), there were no samples gathered. This is an example of gadget losing a connection. During that time there is a flat line in the signal (represented with red line) and no events (blue crosses) are detected.
- Following that period the subject had an unsteady heart rate (area between 500 and 1000 seconds on x axis).
- A steady heart rhythm follows (area between 1000 and 2500 seconds on the x axis). Dispersed events around the relatively steady line indicate irregular heart beats. This is one of the possible indices of problematic areas, allowing the human operator to detect such areas quickly and focus on them.
- On the gathered sample measurement, the subject was involved in a sports activity that periodically raised his heartbeat. This period can be seen quite clearly on the RRI events channel from 3000 seconds to the end of the measurement. Within this periodic activity, there are also numerous outliers in the heart beat events. Those occur mostly in the areas of higher RRI (lower heart rate). Again, these are the areas where the operator should focus on.

Focusing on a smaller area, the signal can be visually inspected in more detail, visualising individual transition in RRI values. Shown on Fig. 7 is a sample of detected transition from irregular to regular heart rhythm. Human operator can inspect such transitions and classify irregularities that are present in the measurement.

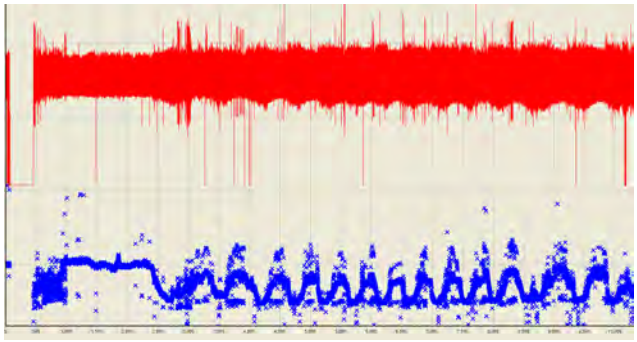


Figure 6. Three hour measurement overview.

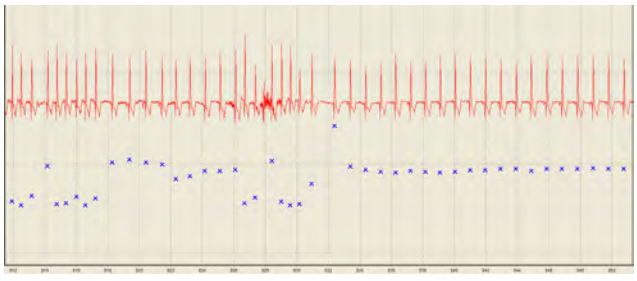


Figure 7. Visual indication of change in heartbeat rhythm.

III. Results

The proposed algorithm was tested on the ECG samples from the PhysioNet's MIT-BIH Arrhythmia Database [3], using MLII channel of each measurement. The detected heart beats were compared to the heart beats annotations from the database. The tests started with the signal being imported into the NevroEkg program where an automatic

beat detection was executed (using the threshold value suggested by the algorithm). The resulting RRI event channel was then compared to the beat annotations given by the database. Only the annotations that represent QRS complex were used and all other annotations were disregarded.

The results are presented in Tab. 1, which contains true positive values (tp; correctly detected beats), false positives (fp; falsely detected beats) and false negatives (fn; beats that were not detected). The table are contains the calculated values for precision, defined as $tp/(tp + fp)$, and recall, defined as $tp/(tp + fn)$. The values of precision and recall that are lower than 90% are coloured red in the table.

On the whole database, the execution of the algorithm produced 97.7% precision and 98.9% recall. By design, the presented algorithm always discards the first heartbeat after a long period without detected beats, since its previous heartbeat is not measured and the RRI cannot be determined. Therefore the 100% recall is rarely achievable on arrhythmic measurements. Closer inspection of the table reveals that the variation in results is quite high. There are only three measurements that have precision or recall lower than 95% though, and out of these, the measurement numbered 219 is the worst with precision only 51.2%.

In measurements numbered 108 and 207, the signal to noise ration was the highest. A relatively high threshold was proposed by the algorithm for those two measurements, which helped reduce the number of false positives in beat identification but also caused a lot of false negatives. Precision therefore was high, while recall was low. Measurements 207 also contains periods of ventricular fluttering[8]. Algorithm has not been adapted to cope with such extreme events, and does not distinguish them from noise. In cases with such low recall, the operator should be able to notice the unidentified beats and manually annotate them after a close review of the signal.

To see weather the threshold values could be set more optimally than the one suggested by the algorithm, the measurements numbered 108, 207, and 219 were processed again with human operator in loop. In Tab. 2 the threshold values were manually adjusted until the results were the most satisfactory for the human operator. Detection on measurement numbered 106 shows large improvement, with both precision and recall rising over 90%. Measurement numbered 207, however, includes too many extreme events, such as the previously mentioned ventricular fluttering. Lowering the threshold value manually did improve recall but for the cost of the reduced precision.

In measurement numbered 219, the algorithm with automatically determined threshold latched on both R and the extraordinarily high P waves. Manually setting the threshold lower on this measurement improved precision to 98.3% while recall remained the same, as shown in Tab. 2.

Even without exclusion of record 207, results are comparable to known robust detectors [7, 9, 10], that are also able to detect beats with both precision and recall in 85-99% range.

TABLE 1. RESULTS OF AUTOMATIC BEAT DETECTION ON MIT-BIH ARRHYTHMIA DATABASE

File	tp	fp	fn	Precision [%]	Recall [%]
100	2272	0	1	100.000	99.956
101	1862	5	3	99.732	99.839
102	2180	6	7	99.726	99.680
103	2083	0	1	100.000	99.952
104	2213	37	16	98.356	99.282
105	2551	58	21	97.777	99.184
106	2018	3	9	99.852	99.556
107	2132	3	5	99.859	99.766
108	1227	0	536	100.000	69.597
109	2530	2	2	99.921	99.921
111	2122	0	2	100.000	99.906
112	2538	0	1	100.000	99.961
113	1793	0	2	100.000	99.889
114	1878	4	1	99.787	99.947
115	1953	0	0	100.000	100.000
116	2390	2	22	99.916	99.088
117	1534	0	1	100.000	99.935
118	2277	5	1	99.781	99.956
119	1987	3	0	99.849	100.000
121	1860	3	3	99.839	99.839
122	2475	1	1	99.960	99.960
123	1517	0	1	100.000	99.934
124	1618	0	1	100.000	99.938
200	2596	28	5	98.933	99.808
201	1953	2	10	99.898	99.491
202	2134	1	2	99.953	99.906
203	2921	145	59	95.271	98.020
205	2640	0	16	100.000	99.398
207	2035	51	297	97.555	87.264
208	2917	7	38	99.761	98.714
209	3004	3	1	99.900	99.967
210	2638	24	12	99.098	99.547
212	2747	0	1	100.000	99.964
213	3249	1	2	99.969	99.938
214	2255	5	7	99.779	99.691
215	3362	1	1	99.970	99.970
217	2199	2	9	99.909	99.592
219	2151	2052	3	51.178	99.861
220	2047	0	1	100.000	99.951
221	2427	0	0	100.000	100.000
222	2478	5	5	99.799	99.799
223	2600	0	5	100.000	99.808
228	2028	48	25	97.688	98.782
230	2255	3	1	99.867	99.956
231	1570	0	1	100.000	99.936
232	1701	6	79	99.649	95.562
233	3074	1	5	99.967	99.838
234	2752	0	1	100.000	99.964
$\sum$	108743	2517	1223	97.738	98.888

TABLE 2. TABLE OF RESULTS FROM MIT-BIH ARRHYTHMIA DATABASE, WITH MANUAL THRESHOLD CORRECTION

File	tp	fp	fn	Precision [%]	Recall [%]
108	1654	55	109	96.782	93.817
207	2098	137	234	93.870	89.966
219	2151	37	3	98.309	99.861

IV. CONCLUSION

During the development of the proposed algorithm, input from multiple areas of expertise was considered. Most of the focus was gathered in the area of enabling quick visual analysis of the signal.

The gathered low resolution noisy samples with missing data have been successfully handled with the proposed semi-automatic algorithm. The algorithm and manual work-flow do require some practice for the human operator to perfect, but are easily understood and can be used after a short introduction. Operators have confirmed that such presentation and the proposed algorithm are sufficient for general overview of long measurements.

The goals of reducing human operator workload, and speeding up and simplifying the process were fulfilled, even though further improvements are still possible. There is no standardized method of processing long and noisy ECG measurements yet.

While awaiting further feedback from human operators that are testing the program, some areas of the process are already recognized as improvable. For example, some speed up of the process could be achieved by using parallel processing on the signals. Post-processing of the result would benefit from improved accuracy and could also expand by adding beat classification. Building upon the improved post-processing, the detected problematic areas or individual beats could also be reported to the human operator in a concise manner, accompanied by written explanation.

ACKNOWLEDGMENT

The authors acknowledge the financial support from the state budget by the Slovenian Research Agency under Grant P2-0095

REFERENCES

[1] R. Trobec, V. Avbelj, M. Šterk, B. Meglič, V. Švigelj. Neurological data measuring and analysis software based on object oriented design. In proceedings of The 7th European Federation of Autonomic Societies (EFAS) Meeting, Bled, Slovenia, May 18-21, 2005. B. Meglič (ed.), pp. 69.

[2] A. Rashkovska, I. Tomašić, R. Trobec. A telemedicine application : ECG data from wireless body sensors on a smartphone. In proceedings of: MIPRO 2011, 34th International Convention, May 23 - 27, 2011, Opatija, Croatia. P. Biljanović, K. Skala (ed.), pp. 293-296.

[3] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, H. E. Stanley, Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals, Circulation 101(23)b 2000b pp. 215–220. Database and tools available at: http://www.physionet.org/.

[4] S. S. Lobodzinski, M. M. Laks. New devices for very long-term ECG monitoring. Cardiology Journal 2012 Mar 1;19(2), pp. 210-214.

[5] R. Trobec, V. Avbelj, A. Rashkovska. Multi-functionality of wireless body sensors. The IPSI BgD transactions on internet research, 2014, 10(1), pp. 23-27.

[6] Arzeno, N. M., Deng, Z. D. and Poon, C. S., Analysis of first-derivative based QRS detection algorithms. IEEE Transactions on Biomedical Engineering, 2008, 55(2), str. 478-484.

[7] J. Pan, W. J. Tompkins. A real-time QRS detection algorithm. IEEE Transactions on Biomedical Engineering, 1985 Mar 3, pp. 230-236.

[8] Goldman, Lee, and Andrew I. Schafer. Goldman's Cecil medicine. Elsevier Health Sciences, 2011, pp. 350.

[9] Hamilton, P.S., Tompkins, W.J. Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database. Biomedical Engineering, IEEE Transactions 1986. (12), pp.1157-1165.

[10] Llamedo, M., Martínez, J.P. Qrs detectors performance comparison in public databases. In Computing in Cardiology Conference (CinC), 2014 September (pp. 357-360). IEEE.

CTI

International Conference on

TELECOMMUNICATIONS & INFORMATION

Steering Committee

Chairs: Tihana Galinac Grbac, University of Rijeka, Croatia
Branko Mikac, University of Zagreb, Croatia
Stjepan Golubić, Zagreb, Croatia
Dragan Jevtić, University of Zagreb, Croatia
Renato Filjar, Ericsson Nikola Tesla, Zagreb, Croatia

Members: Slaviša Aleksić, Vienna University of Technology, Austria
Ţeljka Car, University of Zagreb, Croatia
Saša Dešić, Ericsson Nikola Tesla, Zagreb, Croatia
Vlasta Hudek, Rezultanta, Zagreb, Croatia
Darko Huljenić, Ericsson Nikola Tesla, Zagreb, Croatia
Robert Inkret, Ericsson Nikola Tesla, Zagreb, Croatia
Admela Jukan, Institute of Computer and Network Engineering, Germany
Zdravko Jukić, Croatian Regulatory Authority for Network Industries, Zagreb, Croatia
Marko Lacković, IBM Croatia, Croatia
Erich Leitgeb, Graz University of Technology, Austria
Draţen Lučić, Croatian Regulatory Authority for Network Industries, Zagreb, Croatia
Davor Šarić, Erste Group Card Processor, Zagreb, Croatia
Dina Šimunić, University of Zagreb, Croatia
Velimir Švedek, Croatian Regulatory Authority for Network Industries, Zagreb, Croatia
Antonio Teixeira, University of Aveiro, Portugal

FNS

Special Session on

FUTURE NETWORKS AND SERVICES

Preface

This volume contains papers presented at the special session on Future Networks and Services (FNS 2016) that was organized within MIPRO 2016 conference. The FNS 2016 special session was held on May 31, 2016, in Opatija, Croatia. The session was collocated event to the summer school „*Autonomous Control for Reliable Future Networks and Services'* that was organized by COST Action IC1304 "*Autonomous Control for a Reliable Internet of Services*" *(ACROSS).* Aim of the special session is to provide the opportunity to summer school participants, mainly for PhD students and early-career researchers working in the ACROSS project, to discuss their thesis project idea and early results. We thank for financial support by the EU COST Action ACROSS and to the research project "*Evolving Software Systems: Analysis and Innovative Approaches for Smart Management" (EVOSOFT)* funded by Croatian Science Foundation. We are also grateful to the MIPRO organization for huge support in organizing this event but also preparing proceedings. We also thank to Hotel Admiral Croatia for supporting the special session by providing the lecture hall and equipment.

Tihana Galinac Grbac







PAPERS

A Survey of IoT Cloud Providers

T. Pflanzner, A. Kertesz
University of Szeged
H-6720 Szeged, Dugonics ter 13, Hungary
Email: {tampfla,keratt}@inf.u-szeged.hu

***Abstract*—Cloud Computing enables flexible resource provisions that has become hugely popular for many businesses to take advantage of responding quickly to new demands from customers. Many cloud providers follow recent trends in cloud application development, and offer federation capabilities at the platform level, thus creating Platform-as-a-Service solutions. There is a growing number of providers offering IoT-specific services, since cloud computing has the potential to satisfy IoT needs such as hiding data generation, processing and visualization tasks. While each cloud provider offers its own set of features, two critical features they all have in common are the ability to connect devices and to store the data generated by those devices. In this paper we investigate the basic properties of IoT cloud providers by presenting a survey and a classification of them. This work can be useful for application developers planning to exploit IoT capabilities of PaaS cloud providers.**

***Index Terms*—cloud computing, internet of things, cloud platform, mobile devices**

1. Introduction

Cloud providers can be classified to three main categories according to the level of services they offer: Infrastructure as a Service (IaaS) – at this level infrastructure providers manage and offer virtualized resources managed by hypervisors. IaaS clouds generally have marketplaces with pre-installed virtual machine images, block storages and other file or object storages and virtual network management functionalities; Platform as a Service (PaaS) – providers in this category deliver a virtualized computing platform, including an operating system, programming language execution environment, database and web server; finally at the highest level Software as a Service (SaaS) – these services operate application software in the cloud. Users of these services do not manage the infrastructure nor the platform on which their application runs.

Cloud computing [1] is already a part of our everyday life, and there is so much data produced by humans and machines used by them. As the technology evolves, new kind of innovative use cases can be invented. The Cluster of European Research Projects on the Internet of Things [4] defined the Internet of Things (IoT) as a dynamic global network infrastructure with self configuring capabilities based on standard and interoperable communication protocols. Things in this network interact and communicate among themselves and with the environment by exchanging sensor data, and react autonomously to events and influence them by triggering actions with or without direct human intervention. According to recent reports (e.g. [2]) there will be 30 billion devices always online and more than 200 billion devices discontinuously online by 2020. Such estimations call for an ecosystem that provides means to interconnect and control these devices. With the help of cloud solutions, user data can be stored in a remote location, and can be accessed from anywhere. Cloud providers try to address these needs, and offer services for the IoT community at all three levels, but the PaaS level is the most suitable for IoT application developers to use. Gubbi et al. [3] have identified that to support the IoT vision, the current computing paradigm needs to go beyond traditional mobile computing scenarios and cloud computing has the potential to address these needs as it is able to hide data generation, processing and visualization tasks. Cloud federations can also be applied to serve the needs of IoT applications, as discussed by Celesti et al. [5].

The increasing competition between the leading vendors in the cloud market, such as Amazon, Microsoft, Google and SalesForce, each of which promotes its own, mostly incompatible cloud standards and formats [6], prevents them from agreeing on a widely accepted, standardized way to utilize cloud details and specifications, which trend also appeared in the IoT field. However, in the field of resource abstraction for IoT, good efforts have been made towards the description and implementation of languages and frameworks for efficient representation, annotation and processing of sensed data. The integration of IoT and clouds has been envisioned by Botta et al. [7] by summarizing their main properties, features, underlying technologies, and open issues. A solution for merging IoT and clouds is proposed by Nastic et al. [8]. They argue that system designers and operations managers face numerous challenges to realize IoT cloud systems in practice, due to the complexity and diversity of their requirements in terms of IoT resources consumption, customization and runtime governance. These related works also serve as a motivation to our research by raising the need for managing a large number of protocols and data formats by means of simulation.

The aim of this paper is to investigate the basic properties of IoT cloud providers by presenting a survey and a classification of them. It is not easy to objectively compare cloud providers offering IoT services, as they all provide a slightly different set of services, since providers can advertise such qualities that may not be available or measurable compared to others. Nevertheless this work can be useful for application developers planning to exploit IoT capabilities of PaaS cloud providers.

The remainder of this paper is as follows: Section 2 overviews and categorises the most relevant IoT cloud providers, while Section 3 presents future work towards enhancing these solutions with the help of an IoT device simulator. Finally, the contributions are summarized in Section 4.

2. Survey of Cloud Providers with IoT support

The motivation behind our research is that more and more PaaS cloud providers have started to offer IoT specific services to ease the development of IoT cloud applications, but cases where many heterogeneous things need to be managed are hard to realize and examine. This section introduces and compares IoT capabilities of current PaaS cloud providers.

2.1. PaaS IoT comparison categories

We used the following properties to compare the overviewed providers:

- Provider: The name of the actual cloud provider.
- Open source: Denotes if it is an open source project or not.
- Hosting: If the user can deside the place of the deployment it is open, otherwise if the deployment is restricted to the providers servers it is closed.
- Protocols: The supported IoT device protocols to connect to the cloud.
- Server language: The supported programming languages to be used to write custom server side code.
- Client language: The supported programming languages to be used to write custom client side code.
- Mobile SDK: The offered development kits for the supported mobile platforms.
- Data store: Yes, if the system can store data from devices.
- BLOB: Yes, if the system can store binary data, like images.
- GEO: Yes, if the system supports storing and querying geolocation coordinates.
- Push not.: Yes, if the system capable of handling push notifications.
- Trigger: Yes, if the received data from the devices can be analyzed and can trigger an action.
- Visualization: Yes, if the system has tools to help visualize the data coming from the IoT environment.

The IBM Bluemix Platform [9] is an IoT-enabled PaaS solution offered by IBM. It can be used for quick development of cloud-based applications that take advantage of the data generated by the sensors and devices. Products of several major device manufacturers are supported, such as ARM, the Electronics B&B, Intel, Multi-Tech Systems and Texas Instruments, but other individual cases can also be solved on the platform. Data generated by the equipment is sent by the popular and lightweight MQTT protocol to the cloud. The service allows the users to configure, manage the devices, and to store the history of generated data or stream real-time data to the application. The data transfer can be done through secure APIs.

To illustrate the inner workings of the platform, a real-time data visualization demo is also provided. To use it, first a data provider should be configured, which is in the simplest case a smartphone, but it is possible to use a TI SensorTag, ARM Mbed, Raspberry Pi, Intel and other devices. The opened browser page on the smartphone can send real-time data of the phone's movement to the cloud application. The framework also provides a pre-defined web-based sensor simulator [10] that is able to act as three simulated sensors (sending temperature and humidity values through websockets).

The Bluemix platform offers several specialized services to support the development of cloud applications. Some examples of these services are: Push for messaging, Cloudant NoSQL DB to manage NoSQL databases, Geospatial Analytics for location tracking, and IBM Analytics for Hadoop for Hadoop computations. The supported languages for application development are Java, JavaScript, GO, PHP, Python and Ruby. In terms of costs, a price calculator helps to determine a monthly fee for a 30-day trial period. 20 devices can be connected and 100MB of data can be sent to the devices for free, which is enough for about 50,000 messages. 1GB storage space can also be used in this period.

The Parse [18] PaaS platform developed by Facebook also has IoT support. This platform promises quick and easy application development with the support for mobile devices (through its MBaaS service). In addition to the C SDK for Linux (Raspberry Pi) and real time systems (RTOS) (TI CC3200), there are some more special versions offered such as Arduino, and a number of partners have SDKs, such as Atmel, Broadcom, Intel and Texas Instruments. A great benefit of the platform is that all of their SDKs are open source. These SDKs allow data sending and the so-called push notifications, and they can also take advantage of the Parse cloud services. Many sample applications are available to demonstrate its usage, including farming, music and cooking scenarios. JavaScript applications are supported, and the Parse Webhook service makes it able to link applications from remote clouds. The supported mobile platforms are the iOS, Windows Phone, Android, Unity and Xamarin. Web and desktop SDKs are for OSX, Windows, JavaScript, Unity, PHP, and .NET. The SDK supports offline data storing, social media connections can be properly handled. It is possible to schedule jobs, as well as a high quality Dashboard is also available that supports data modification, statistical

TABLE 1. PAAS GENERAL FEATURES.

Provider	Open source	Hosting	Server	Client languages	Mobile SDK languages
Bluemix	no	closed	many	Java, JS	Android, iOS
Parse	yes - SDK	closed	JS	Java, JS, C	Android, iOS
Google	no	closed	many	Java, Python	Android, iOS
Amazon	no	closed	many	C, JS	Android, iOS
Azure	no	closed	many	C, Java, JS	Android, iOS, WP
Heroku	no	closed	many	-	-
CloudFoundry	no	closed	many	N/A	N/A
Kinvey	no	open	JS	JS	Android, iOS
DreamFactory	yes	open	JS	JS	Android, iOS

TABLE 2. PAAS IOT FEATURES.

Provider	Protocols	Data store	BLOB	GEO	Push not.	Trigger	Visualization
Bluemix	MQTT	yes	no	yes	yes	yes	yes
Parse	REST	yes	yes	yes	yes	yes	no
Google	REST	yes	yes	yes	yes	yes	yes
Amazon	MQTT, REST	yes	yes	yes	yes	yes	yes
Azure	MQTT, AMQP, REST	yes	yes	yes	yes	yes	N/A
Heroku	MQTT	yes	yes	N/A	no	no	N/A
CloudFoundry	REST	yes	yes	N/A	no	no	no
Kinvey	REST	yes	yes	N/A	yes	yes	N/A
DreamFactory	REST	yes	yes	N/A	yes	yes	N/A

testing, push notification, and storing logs. Objects can be managed in the system after specifying their schema. Their schema supports many field formats, e.g. arbitrary objects, array, blobs and geopoints. Their dashboard provides no tools for data visualization. Blobs can be imported from a public URL, even directly from the Base64 encoded Data URI format.

The Google IoT solution [11] is part of the Google Cloud Platform, which includes various Google services. The scalability is an excellent feature in this platform. It allows devices to be connected, and it collects data and visualizes them. The data sent from the devices are received by the Google Load Balancer and forwards to instances of the AppEngine applications. In general, the main part of the application is the AppEngine, which may use other services. Compute intensive tasks are supported by the Compute Engine. The Cloud Storage and the Cloud SQL manages data. It is possible to send data with streams to the BigQuery service, which is ideal if we want to work with real-time data. In IoT systems the visualization is an important feature, it is supported in real time using the Google Charts. Google is also strong in managing a large amount of data processing, which is important, since there are many devices generating huge amount of data in IoT systems. Google FireBase plays an important role in the management of the devices. It was originally designed to assist mobile devices (like MBaaS). It provides synchronized real-time database, authentication and capable of offline operations.

Amazon Web Services [12] is a collection of services that make up a cloud computing platform, which are based on 11 geographical regions across the world. The most central and well-known services are Amazon EC2 (Elastic Compute Cloud) and Amazon S3 (Simple Storage Service). The products are offered to large and small companies as a service to provide large computing capacity faster and cheaper than the client company building and maintaining an actual physical server farm. AWS automatically handles the details such as resource provisioning, load balancing, scaling and monitoring. One can create applications in PHP, Java, Python, Ruby, node.js, .NET, Go or in a Docker container that runs on an application server with a database. An environment using the default settings will run a single Amazon EC2 micro instance and an Elastic Load Balancer. Additional instances will be added if needed, to handle any peaks in workload or traffic. Each Amazon EC2 instance is built from an Amazon Machine Image which can be an Amazon Linux AMI or an Amazon Windows Server 2008 R2 AMI by default. Amazon is a PaaS provider, and they have many ”blocks” to build applications with. This allows for more general usage, but not so many details, that could make the developer’s job easier. With the three main components (Cognito for user management, Mobile Analytics and Simple Notification Service) the mobile solution is a valuable part for the whole Amazon cloud offering. This is still not mature enough for enterprise usage, because the lack of integration and security.

Amazon IoT connects devices to services and other devices with a secure way. The device state is synchronized, so messages can be sent even if the device is offline. The Rule Engine helps to convert the data for services.

Azure [13] is a Cloud computing platform, which allows developers to publish web applications running on different frameworks, written in different programming languages such as any .NET language, node.js, php, Python and Java. Azure Web Sites supports a website creation wizard that can be used to create a blank site or use one of the several

pre-configured sites. Developers can add or modify content of the website via multiple deployment methods: TFS, FTP, CodePlex, GitHub, Dropbox, Bitbucket, Mercurial or git. Developers can select the place where their website will be hosted from several Microsoft data centers around the globe. Azure Traffic Manager routes traffic manually or automatically between websites in different regions. Web sites are hosted on IIS 8.0, running on a custom version of Windows Server 2012. The component regarding IoT called IoT Hub, which can communicate with devices with protocols like MQTT, AMQP and HTTP, but it's possible to implement other protocols too.

Heroku [14] has been in development since 2007, starting with support for Ruby, and adding support for many languages through the years, such as Java, Node.js, Scala, Clojure, Python, PHP and Perl. Heroku was acquired by Salesforce.com in 2010, as a subsidiary. Herokus services run on the Amazon cloud systems. From the Developer Experience point-of-view, Herokus interface is well-polished, intuitive and easy to use. Many times Heroku has been seen as an example by other PaaS providers, for their ease of use, features and reliability. One such example is Deis, which uses a Heroku inspired buildpack system in their deployments. The basic units of computing power in the Heroku ecosystem are the Dynos. A Dyno is a lightweight, isolated container that runs an instance of the application. The IoT capability of Heroku is only an MQTT broker, which can be added as a 3rd party add-on, called CloudMQTT.

CloudFoundry [15] is an open source PaaS service, originally developed by VMware, later owned by Pivotal Software, primarily written in Ruby and Go. CloudFoundry is available in three flavors, Cloud Foundry OSS, an open source project available to everybody, which uses the developers own infrastructure and the BOSH shell to interact with it, Pivotal Cloud Foundry, a commercial product from Pivotal, which includes extra tools for installation and administration, and Pivotal Web services, which is an instance of Pivotal Cloud Foundry hosted on Amazon Web Services. Applications deployed to CloudFoundry access external resources via Services. All external dependencies such as databases, messaging systems, file system, etc. are Services. When releasing an application, the developer must specify the Services it should use. Many pre-defined services are available via an administration console, such as MySQL, PostgreSQL, MongoDB, etc. as database services, RabbitMQ as a messaging service, Jenkins for continuous integration, and API Gateway, Data Sync, Push Notifications for mobile development.

DreamFactory [16] has many install guides for IaaS providers (Docker, Amazon Web Services, Microsoft Azure, Google Cloud Platform, VMware Marketplace, Bitnami Cloud Hosting), for PaaS providers (Red Hat OpenShift, Pivotal Web Services, IBM Bluemix, Heroku) for Desktop Computer (LAMP or WAMP) (Linux, OS X, Windows) IBM SoftLayer, Rackspace Marketplace. For push notification it uses the Amazon SNS. The SDK is moderately good, too general, requires too much coding. On the free hosted version the login window pops up for almost every click.

The Kinvey [17] platform has 3 main parts, the client, the context and the aggregation part. The client part called the MOBILE CLIENT TIER is a consistent set of mobile client features like offline caching and encryption that stay current and enable any use case. The context part called the MOBILE CONTEXT TIER is on-demand services for data, push notifications, location, analytics and business logic to support any use case, versioning and deployment of new apps and backend infrastructure for multiple projects. The aggregation part called the MOBILE AGGREGATION TIER is responsible for the consistent and secure data and identity integrations to deliver the right information for the mobile experience.

The main IoT-related properties of these cloud providers are shown in Table 1 and 2. Summarizing the comparison tables, Google, Ammazon, Azure and Bluemix has the highest variety of IoT-related services. The MQTT (or other IoT protocol) should be a basic functionality to an IoT cloud platform, but many provider has just a REST interface. Parse has its high quality MBaaS service, while its IoT feature is a limited version of it. IoT applications in Google can be composed of many connected services, which makes it complex providing more freedom for the developers. They are also very good at scaling and performance, and this complexity is compensated by the simplicity of FireBase.

From this survey we can see that the most popular cloud providers have already realized the need for IoT support, and most of them provide reasonably good solutions for IoT application development. Nevertheless interoperability issues still exist, and applications managing a large number of different IoT devices are hard to develop and evaluate. Bluemix has also identified the need for a sensor or device simulator, but its tool is meant to serve simple demonstration purposes. Our aim is to design a generic solution to this problem.

3. Requirements for a Cloud-enabled IoT Device Simulator

After examinig the most popular IoT Cloud providers, we identified the following challenges related to IoT networks by examining the currently available solutions:

- IoT devices are battery powered;
- They communicate using low-power wireless technologies (e.g., IEEE 802.11, IEEE 802.15.4, Bluetooth);
- There are different resource constraints of devices (e.g., on CPU, memory, connectivity);
- IoT networks are very dynamic as network conditions can change rapidly;
- They are heterogeneous: there is a large spread on device capabilities (e.g., powerful cameras, low cost temperature sensors), additionally there are sources (sensors) and sinks of information (actuators);
- They are very dynamic: the networking environment in a IoT environment is largely unstructured can vary rapidly.

In order to enhance the design, development and testing processes of IoT systems and networks, an IoT device simulator would be useful that can emulate real devices and sensors instead of real resources. Our ongoing and future work addresses these needs, therefore we highlight the basic requirements of such a simulator in this section.

The requirements for basic functionalities of an IoT device simulator are to send and receive messages, generate sensor data (for one or more devices), and react to received messages. These capabilities are sufficient to use the simulator in IoT system analysis. Requirements for advanced functionalities such as simulating network errors, recording and replaying concrete simulation cases, and connecting real IoT devices to the simulator can contribute to the analysis of more realistic system.

There are different kind of IoT environments, hence their static or dynamic properties, and the number of utilized devices can affect the design of such a simulator. For example, a connected house can be regarded as a static environment, because its devices are usually in one place, possibly with wired connection, providing reliable network stability. The dynamic environment is more complex to simulate, in such cases we would like to simulate a broader part of the environment considering WiFi interference, battery lifetime and locations of the devices.

In our future work we plan to support the basic functionalities with the following settings:

- a simulated device should have an ID, or tokens for authentication;
- the generated sensor data should be made available in binary, plain text, or JSON format with metadata like date, time, and device state;
- finally MQTT or REST communication protocols should be supported.

Based on these planned functionalities, we propose a mobile IoT device simulator called MobIoTSim, which can be used to help cloud application developers to learn IoT device handling without buying real sensors, and to test and demonstrate IoT applications utilizing multiple devices. The structure of the application lets users create IoT environment simulations in a fast and efficient way with the options for custom settings.

We also plan to support larger, scalable experiments in the future. For this purpose we will use a scripting language to be able to specify the devices and their properties, and to schedule their activities. For cloud application testing, a great feature would be to record sensor data and networking events, and later it could be replayed again many times, with exactly the same scenario.

4. Conclusion

In this paper we have shown that there is a growing number of cloud providers offering IoT-specific services, since cloud computing has the potential to satisfy IoT needs such as hiding data generation, processing and visualization tasks. While each cloud provider offers its own set of features, two critical features they all have in common are the ability to connect devices and to store the data generated by those devices.

We investigated the basic properties of IoT Cloud providers by presenting a survey and a classification of them. This work can be useful for application developers planning to exploit IoT capabilities of PaaS cloud providers.

Our future work will address the development of a mobile IoT device simulator to aid IoT application development with the surveyed providers. Such a simulator can be used to help cloud application developers to learn IoT device handling without buying real sensors.

Acknowledgments

The research leading to these results has received funding from the European COST programme under Action identifier IC1304 (ACROSS).

References

[1] B. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, I. and Brandic, Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems, vol. 25, no. 6, pp. 599-616, June 2009.

[2] J. Mahoney and H. LeHong, The Internet of Things is coming, Gartner report. Online: https://www.gartner.com/doc/1799626/internet-things-coming, September 2011.

[3] J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami. Internet of Things (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems, Volume 29, Issue 7, pp. 1645-1660, September 2013.

[4] H. Sundmaeker, P. Guillemin, P. Friess, S. Woelffle. Vision and challenges for realising the Internet of Things. CERP IoT - Cluster of European Research Projects on the Internet of Things, CN: KK-31-10-323-EN-C, March 2010.

[5] A. Celesti, M. Fazio, M. Giacobbe, A. Puliafito, and M. Villari, Characterizing Cloud Federation in IoT. IEEE 30th International Conference on Advanced Information Networking and Applications Workshops - Workshop on Cloud Computing Project and Initiatives, 2016.

[6] G. Sperb Machado, D. Hausheer, and B. Stiller, Considerations on the Interoperability of and between Cloud Computing Standards. In 27th Open Grid Forum (OGF27), G2CNet Workshop: From Grid to Cloud Networks, Banff, Canada, 2009.

[7] A. Botta, W. de Donato, V. Persico, A. Pescape, On the Integration of Cloud Computing and Internet of Things. The 2nd International Conference on Future Internet of Things and Cloud (FiCloud-2014), August 2014.

[8] S. Nastic, S. Sehic, D. Le, H. Truong, and S. Dustdar, Provisioning Software-defined IoT Cloud Systems. The 2nd International Conference on Future Internet of Things and Cloud (FiCloud-2014), August 2014.

[9] IBM Bluemix Platform. Online: https://console.ng.bluemix.net/. Accessed in January, 2016.

[10] IBM Bluemix IoT Sensor. Online: https://developer.ibm.com/recipes/tutorials/use-the-simulated-device-to-experience-the-iot-foundation/. Accessed in March, 2016.

[11] Google Cloud Platform. Online: https://cloud.google.com/solutions/iot/. Accessed in January, 2016.

[12] Amazon Web Services. Online: http://aws.amazon.com/. Accessed in January, 2016.

[13] Microsoft Azure. Online:https://azure.microsoft.com/. Accessed in January, 2016.

[14] Heroku. Online:https://www.heroku.com/. Accessed in January, 2016.

[15] CloudFoundry. Online:http://cloudinary.com/. Accessed in January, 2016.

[16] DreamFactory. Online:http://www.dreamfactory.com/. Accessed in January, 2016.

[17] Kinvey. Online:http://www.kinvey.com/. Accessed in March, 2016.

[18] Parse. Online: https://parse.com/products/iot. Accessed in March, 2016.

[19] IBM IoT Foundation message format. Online: https://docs.internetofthings.ibmcloud.com/gateways/mqtt.html#/managed-gateways#managed-gateways. Accessed in January, 2016.

[20] IBM IoT Foundation Quickstart application. Online: https://quickstart.internetofthings.ibmcloud.com. Accessed in January, 2016.

[21] IOT Visualization application. Online: https://github.com/ibm-messaging/iot-visualization/. Accessed in January, 2016.

QoS-aware Deployment of Data Streaming Applications over Distributed Infrastructures

Matteo Nardelli
Department of Civil Engineering and Computer Science Engineering
University of Rome Tor Vergata, Italy
Email: nardelli@ing.uniroma2.it

***Abstract*—The advent of the Big Data era and the diffusion of the Cloud computing paradigm have renewed the interest in Data Stream Processing (DSP) applications, which can timely extract valuable information from an increasing number of data sources (e.g., sensing devices, social networks). The distribution of data sources, the huge and unpredictable volumes of produced data, and the diffusion of heterogeneous computing devices, often available on demand, foster new strategies for an efficient and network-aware usage of the underlying infrastructure.**

The doctorate work, presented in this paper, investigates the deployment of DSP applications with Quality of Service (QoS) requirements over a large-scale distributed infrastructure of heterogeneous computing and networking resources. Specifically, to support our study, we extend an open-source DSP system, Apache Storm, by providing mechanisms for executing distributed QoS-aware scheduling policies and self-adaptation. We provide a general formulation of the optimal placement problem for DSP application (for short, ODP), taking into account the heterogeneity of computing and networking resources.

The ongoing research aims at investigating the following directions. First, we will design heuristics able to determine the best placement in a feasible amount of time. Second, we will investigate runtime adaptation strategies and online placement algorithms. Third, to prove the generality of our approach, we will customize the designed models and algorithms for similar problems (e.g., service selection, container deployment).

I. Introduction

Today our environments are permeated with an increasing number of networked computing and sensing devices, which continuously produce streams of data that can be collected and analyzed by distributed data stream processing (DSP) applications, to timely extract valuable information about many fundamental aspects of our life, such as urban mobility, healthcare, energy management. The key feature of DSP applications is the ability of processing data on-the-fly (i.e., without storing them), moving them from an operator to the next one, before reaching the final consumers of the information. Each operator can be seen as a black-box and self-contained processing element that receives incoming streams, applies a user-defined code, and finally generates new outgoing streams. The user-defined code can be a predefined operation (e.g., filtering, aggregation, merging) or something more complex (e.g., POS-tagging). A DSP application can be described with a directed acyclic graph (DAG), where the vertices represent the application operators and the edges represent the logical links between operators, through which the data streams flow.

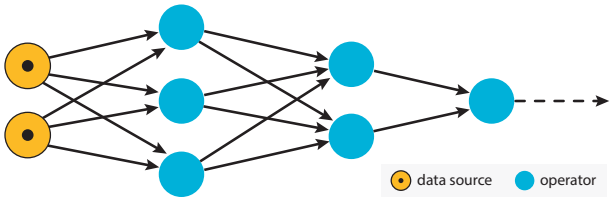


Fig. 1: An example of a DSP application

Figure 1 shows a DSP application, represented through its DAG.

The classic approach deploys DSP applications on large-scale and centralized data centers that are often distant from data sources. However, as data increase in size, pushing them towards the Internet core could cause excessive stress for the network infrastructure and also introduce excessive delays. To increase scalability and reduce latency, a possible solution lies in taking advantage of the ever increasing presence of near-edge computing resources (i.e., Fog Computing [1]) and decentralizing the DSP application, by moving the computation close to data sources. Nevertheless, the deployment on a diffused infrastructure poses new challenges that include geographic distribution, system and network heterogeneity, and non-negligible network latencies. This latter aspect may have a strong impact on DSP applications for latency-sensitive domains (e.g., [2]).

To deploy an application, the DSP system needs to solve the *operator placement problem*, which determines, within a set of available distributed computing nodes, the nodes that should host and execute each operator with the goal of optimizing the Quality of Service (QoS) attributes of the application. Being a special case of the partitioning problem, the operator placement problem is NP-hard [3]. The features of DSP applications further exacerbate the challenges of their deployment; indeed, the computational characteristics of the operators are usually not known a-priori, the properties of the input streams change continuously, and the load imposed has to be sustained for long provisioning times. All these features call for run-time adaptation capabilities of DSP systems, which should change the operators placement in order to efficiently satisfy the QoS requirements. The adaptation capabilities can be expressed along different directions, which comprise the runtime relocation of the application operators and the elastic

scaling of the number of operator instances (i.e., replicas, for increasing fault tolerance, or shards, for exploiting parallelism). Elasticity and adaptation allow to fully exploit the Cloud computing principles, which promote the elastic usage of on-demand resources by acquiring and releasing them as needed.

This paper outlines my PhD work that investigates the deployment of DSP applications with QoS requirements over a large-scale distributed infrastructure of heterogeneous computing and networking resources. Although the placement problem is well known in literature (e.g., [4], [5]) and several heuristics have been proposed (e.g., [6], [7], [8], [9], [10]), the emerging context, made of strict requirements and heterogeneous resources, demands for a further investigation of the problem, to better exploit the potential of the available infrastructures.

II. Research Challenges

In a distributed system, applications often consist of small building blocks (e.g., components, operators, services) that exchange information using the network. Therefore, an efficient placement is beneficial for different contexts (e.g., DSP, service-oriented architectures, container development, VM allocation). Even focusing only on DSP applications, different research directions can be identified due to the great potentialities offered by the underlying computing infrastructure. The research questions that motivate our work are as follows.

- RQ1 **Decentralized Computation:** What are the key requirements of DSP applications? What are the challenges of deploying DSP applications in a distributed environment? How do network latencies affect the application performances?
- RQ2 **QoS-awareness and Optimal Placement:** Considering both QoS requirements of DSP applications and heterogeneity of computing and networking resources, what is the role of QoS attributes in determining the applications placement? How can we model the optimal placement problem?
- RQ3 **Adaptation Capabilities:** DSP applications and their execution environments are often subject to unpredictable changes, e.g., in terms of node availability, workload size, network congestion. If a DSP system can benefit from adaptation mechanisms, what is the role that the placement problem and the elasticity play in reconfiguring the application in response to changes? Are there any different needs in determining the initial deployment and the following runtime adaptations? Is it useful to have differentiated policies? What are the challenges of designing efficient elasticity mechanisms and policies?
- RQ4 **Online Placement:** In a more general setting, a distributed infrastructure can host multiple concurrent DSP applications, each arriving and departing over time with unforeseen requirements and characteristics. How can the system trade off between the QoS requirements of multiple DSP applications and the service provider objectives, e.g., profit maximization and/or optimal resource utilization?
- RQ5 **Cross-layer Optimization:** The emerging Software Defined Networking (SDN, [11]) promises an increment of flexibility and performance of networks, exploiting the programmability of forwarding logic within network devices. Leveraging on SDN functionalities, is it possible to exploit a cross-layer optimization, so to improve the execution of concurrent applications while reducing the load and congestion of the network?
- RQ6 **Extensions and Customizations:** The generality of the placement problem allows us to extend in different contexts the achieved research outcomes. Examples are the deployment of software containers in a distributed Cloud environment, and the resolution (or service selection) of service-oriented applications (both SOA and micro-service applications). How can the developed placement model for DSP applications be customized to represent features and requirements of a different context?

To support our investigations, we also develop a DSP framework, which can: (1) execute applications in a heterogeneous environment; (2) be equipped with centralized and decentralized placement policies; and (3) provide adaptation capabilities, to cope with the dynamism of the execution environment.

III. Related Work

The DSP placement problem has been widely investigated in the literature under different modeling assumptions and optimization goals. Due to its hardness, several heuristics have been proposed.

A. Placement Problem

In a geographically distributed environment, differently from a locally distributed one, system resources comprise computing nodes together with the logical connectivity among them. Among the first proposals, Zhu et al. [5] studied a placement problem which accounts for computational and communicational delays. However, their results are not suitable for nowadays DSP applications, because neglecting resource capacity they assume that a node can host at most a single operator. Huang et al. [12] rearrange the Multi-Constrained Optimal Path problem to model and solve the service composition problem. Nevertheless, their model allows only to express constraints on communication links but not on computing resources. Most of the other existing formulations specializes different aspects of the model, either aiming at developing heuristics (e.g., [13]), improving the expressiveness of constraints (e.g., [14]), or discovering strong theoretical properties that can simplify the problem itself (e.g., [15]).

The emerging SDN is attracting interest from different communities, because it enables to dynamically define networks on the basis of the application needs. As presented in [16],

Big Data applications (such as DSP) can simplify the tasks of data acquisition, transmission, and processing by exploiting a cross-layer design that leverages on the SDN features. Gu et al. [17] show that a joint optimization of application placement and routing logic performs much better than a two-step optimization (in their study, benefits are in terms of energy cost reduction).

B. Heuristic Approaches

A great variety of placement algorithms have been proposed in the literature and Lakshmanan et al. [6] provide a comprehensive overview of them. As the authors show, they differ each other on assumptions and optimization goals, leading to strategies not fully comparable one to each other. The different heuristic approaches optimize a diversity of utility functions, such as to minimize the application end-to-end latency (e.g., [8]), the inter-node traffic (e.g., [7], [18]), and the network usage (e.g., [9], [10]). In a first set of works, the proposed heuristics reduce the placement problem to the graph partition problem (e.g., [8]), where the application operators are grouped and assigned while reducing a given utility function. In another set of works, the problem is modeled as a bin-packing problem, which is then solved using a best-fit or first-fit heuristic. Relying on the best-fit meta-heuristic, Aniello et al. [7] and Xu et al. [18] propose centralized algorithms that assign operators with the goal of reducing the inter-node traffic exchanged on the network during the application execution. It is worth mentioning also the decentralized approaches. Pietzuch et al. [9] minimize the network usage in a completely distributed way by modeling the application as an equivalent system of springs. Rizou et al. [10] exploit the mathematical properties of the network usage function that allows a distributed and iterative computation of the optimal operators placement.

C. DSP Frameworks

As technologies and needs evolve in time, the DSP paradigm has experienced different generation of architectures [19], where the last one relies on Cloud-based resources. However, most DSP systems are still designed to run in a local cluster, where the often homogeneous nodes are interconnected with negligible network delays (e.g., [20], [21], [22]). These assumptions do not hold any more when the DSP system runs in geographically distributed and dynamic environments, where a great heterogeneity of devices are interconnected with not-negligible network latencies. Apache Storm [21] is a framework of the last generation, that provides an abstraction layer to execute event-based applications. Several research efforts have used Storm to either evaluate new operator placement algorithms in a real environment or to propose some architectural improvements (e.g., [7], [19], [23], [18]). Apache Spark [24] is a general-purpose framework for large-scale processing. It extends and improves the MapReduce programing model (batch processing) and, using the Spark Streaming module, can reduce the size of each batch and process streams of data (micro-batch processing). Another emerging framework is Apache Flink[1], that proposes a unified framework for batch and stream processing. Aside the specific functionalities, all of these frameworks use directed graphs to model DSP applications.

The most popular open-source DSP frameworks, such as the ones presented so far, do not fully support the adaptation of running applications. Nevertheless, with the advent of Cloud computing, which promotes the on-demand resource principle, the elasticity has gained increasing interest (e.g., [25], [26], [27], [28]). The challenges of designing elastic systems emerge especially when applications must handle an internal state, because the elasticity of stateful operators requires efficient state migration strategies, needed to preserve the applications integrity [26]. The placement problem also appears in several works that investigate *when* to scale and *where* to relocate the operators when a scaling decision is needed (e.g., [27], [28]).

IV. Research Approach

We are interested in determining the placement of DSP applications over an infrastructure, which comprises heterogeneous and (possibly geographically) distributed computing and networking resources. In other words, we want to explore the benefits of infrastructure-awareness while determining the application placement, because in the emerging environment network latencies and QoS attributes are non-negligible. Unlike the approaches proposed so far in the literature (Section III), our goal is to provide a general formulation, which can be easily extended to include new constraints and QoS attributes. We also need to address the vagueness of the "best placement" concept, that strictly depends on the application context. To this end, we need an analytical definition of the user preferences about the desirable application performances; this definition should account for different and possibly conflicting optimization goals, such as minimize latency while maximizing the throughput. Exploiting the generality of the placement problem in the distributed systems, we want to conclude our research project by investigating the possibility of reusing the developed solutions in different contexts.

The problem under investigation is complex, because it requires: (1) to understand the interactions between different entities and quality attributes; (2) to model these relationships and the optimal placement problem; (3) to investigate solution approaches that efficiently compute the placement; and (4) to address the dynamism of the execution environment, exploiting runtime adaptation strategies and online placement algorithms. However, the research questions, identified in Section II, sketch an outline that drives our research and that we follow by adopting an engineering-oriented approach.

Firstly, we analyze and understand the placement challenges relying on a thorough exploration of the literature and on empirical experiments on a distributed system. To this end, we develop a DSP framework that can integrate the prototype solutions that will be designed. Once proved that adaptation capabilities are beneficial for DSP systems, we can design

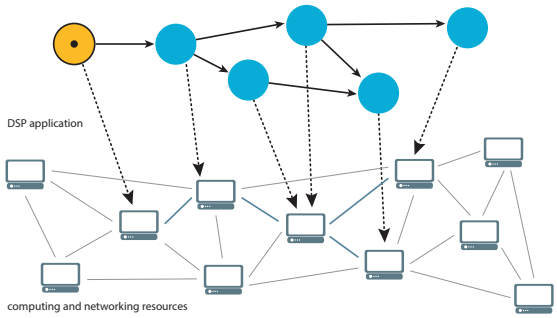
[1] https://flink.apache.org/

adaptation mechanisms and investigate the needs of runtime reconfigurations in terms of operators placement and replication.

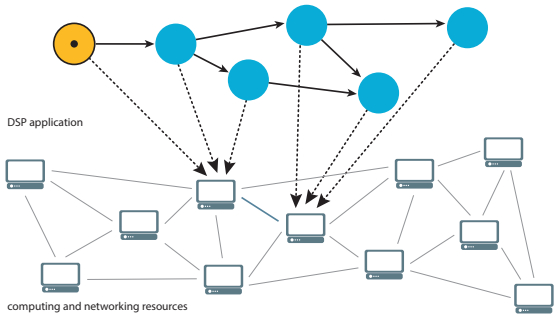
Secondly, we identify a proper representation of applications and resources, and we model the placement problem with tools provided by the optimization theory. The resulting placement model provides also a general framework against which existing centralized and decentralized placement solutions can be evaluated.

Thirdly, we devise an efficient resolution heuristic for the placement model, which is NP-hard because of its combinatorial nature. Differently from the existing solutions (Section III), the formalized optimal placement model should steer the design of the heuristics. As a consequence, we can provide quantitative information on the heuristic quality.

Finally, we enhance the placement model according to two directions: to optimize the online placement of applications; to customize the model for other application contexts. By definition, an online algorithm must take decisions facing a lack of complete knowledge of future resource requests [29]. The research challenge is to optimize resource utilization in the long run, while coping with uncertainty and incomplete knowledge. We are interested in applying the placement solution in different contexts, because, abstracting away the peculiarities of the different kinds of applications, there is the common need of deploying distributed applications on heterogeneous infrastructures. The general validity of the outcomes can confirm the quality of the research work.



(a) The utility function prefers a wide distribution of the operators



(b) The utility function reduces the number of used network links

Fig. 2: Example of the application placement, resulting by different utility functions.

V. Current Status

We now present the current status of the ongoing research and how the research questions outlined in Section II have been addressed so far.

A. A QoS-aware and Self-adaptive DSP Framework

To answer the first three questions, aside exploring the existing literature, we developed a prototype DSP framework, which can execute both centralized and decentralized placement and adaptation strategies on a distributed and heterogeneous infrastructure. In a preliminary phase, this framework has been useful to empirically evaluate realistic applications on large-scale infrastructures and identify the key features that a placement policy must consider. This approach is alternative to the a simulation-based one, where the observable phenomena strictly depend on the system representation previously defined within the simulator. To focus our efforts, we did not develop the framework from the scratch, but we extended an existent open-source framework, Apache Storm. The extension, presented in [30], has introduced some key components that are seamlessly integrated and provide distributed monitoring, scheduling and management capabilities.

Since DSP applications are long-running and subject to unpredictable load burst, soon after the first experiments we realized that an efficient DSP system needs adaptation capabilities. In fact, a static (one-time) definition of the application placement cannot keep the pace of the execution environment dynamism. On the other hand, a dynamic reconfiguration requires to migrate the operators state, which, as side effect, reduces the application availability and temporary increases its response time. In other words, differently from the initial deployment, adapting running applications has a "cost" that cannot be neglected. Our extension has introduced in Storm the MAPE control cycle [31], the reference model for autonomic systems, which allows to perform continuous adaptation of the application placement according to a decentralized placement policy.

Leveraging on this QoS-aware and self-adaptive DSP framework, we have investigated some placement policies. In [32] and [33], we evaluated two relevant distributed solutions proposed in the literature and we have shown the challenges of designing fully decentralized and uncoordinated placement strategies. The outcomes of the research show that QoS-aware placement policies are truly beneficial for DSP applications, because network latencies cannot be neglected in this execution environment [30]. The results empirically show that different policies, although they share the same optimization goal, can lead to different application performances [33]. Moreover, designing decentralized placement policies further increases the challenges, because the lack of an efficient coordination system can lead to some instability that negatively affects the application QoS (e.g., reducing the application availability) [32].

B. The Optimal Placement Model

The research outcomes emphasize the importance of QoS attributes while determining the application placement in distributed environments. Several QoS-aware placement policies have been proposed in the literature. Nevertheless, although they are intended for solving the same problem, they differ from each other on their own assumptions and optimization goals, leading to strategies not fully comparable one to each other.

Therefore, in [34], we proposed a general formulation of the optimal DSP placement (for short, ODP) as an Integer Linear Programming problem which explicitly takes into account the heterogeneity of computing and networking resources. Unlike the existing solutions (see Section III), our formulation explicitly models application operators, streams, computing nodes, and network connectivity. Each of these entities exposes its QoS attributes, whether they are requirements or capacities; specifically, in [34], we focus on latency, availability, and exchanged data rate. ODP provides a general modeling framework that can be easily extended to include new constraints and QoS attributes, such as memory, operator replication, pay-as-you-go pricing policy, and privacy. Exploiting the optimization theory, ODP can find the optimal placement solution (i.e., the best one) taking into account the application needs, the set of available computing and network resources, and the user preferences on the application placement (i.e., an utility function). Figure 2 exemplifies the effect of different utility functions on the optimal application placement. Even if solving ODP does not scale as the number of involved entities (i.e., applications, resources) grows, it gives an insight of the optimal placement, can steer the design of heuristic approaches, and provides a benchmark against which other centralized and decentralized placement algorithms can be compared. We also developed an ODP-based scheduler for Apache Storm and evaluated some well-known placement solutions proposed in the literature (i.e., [9], [10], [18]).

VI. Expected Contribution

We have answered almost half of the research questions, therefore in the following months we will continue along these directions: heuristic algorithms, dynamic placement reconfiguration, online placement, and extensions and customizations.

A. Placement Heuristics

Since the optimal placement is an NP-hard problem, heuristic resolution algorithms are needed. Similarly to the approach pursued in [35], we want to use the optimal placement model as reference point to develop heuristic algorithms. This enables a quantitative evaluation of the trade-off between solution optimality and time needed for its computation. An efficient solution provides a twofold benefit. Firstly, due to the reduced computational time, an heuristic can be executed periodically, so to continuously evaluate the application placement and to trigger adaptation decisions when a performance degradation is experienced. Secondly, it increases the scalability of a DSP system that can faster accommodate incoming applications. As presented in Section III, most of the heuristics proposed in the literature customizes some well-known approaches, such as greedy best-fit and first-fit bin packing, tabu search, genetic algorithms, and simulated-annealing. We are currently investigating how to customize these strategies to solve our placement problem.

B. Placement and Adaptation

Answers of some research questions have highlighted that DSP systems call for adaptation capabilities. In [33], we presented how the self-adaptation is performed according to the MAPE control cycle. Specifically, in a single iteration of MAPE, the DSP system *monitors* the environment and the applications, *analyzes* if reconfigurations are needed, and, in positive case, *plans* and *executes* the corresponding adaptation actions. In this general framework, the plan phase is in charge of reconfiguring the application in order to address the external changes; this step may require to solve the placement problem when either an application has to be reassigned or an operator has to be relocated. Differently from the initial deployment, reconfiguring a running application introduces some costs that cannot be neglected (e.g., reduction of the application availability, increase in response time). A differentiated placement policy, which distinguishes between initial placement and runtime adaptation, can thus help to improve the application performance, because it can better address the different needs of these placement decisions. To this end, we want to extend the optimal placement in order to include these "adaptation costs" and subsequently determine a new heuristic that can take runtime placement decisions while optimizing the adaptation overhead.

C. Online Placement

The ODP model determines the placement of a single application at a time. It can be easily extended to compute the placement of multiple applications, nevertheless assuming that all the applications require resources at the same time or that we have in advance a complete knowledge of the applications that the system will host and execute in the future. In a real scenario, these assumptions are not realistic and, if ODP accommodates an application at a time with a greedy definition of its optimal placement, in the long run, it will lead the system to a sub-optimal utilization of the available resources. Therefore, we want to investigate modeling techniques that can guarantee near-optimal performances even when the system has to take placement decisions facing a lack of complete knowledge of future resource requests. This is one of the most challenging research questions, but also the most interesting one from the service provider standpoint.

Different works address the online allocations of virtual machines (e.g., [36]), but, at the best of our knowledge, none of them investigate this problem for distributed DSP applications. We plan to explore online placement policies as the latest step of the research project, when we may take advantage of the knowledge gathered from the heuristics development (e.g., how to deal with soft and hard QoS requirements).

D. Extensions and Customizations

We have already integrated the DSP system with a SDN controller, and we are going to investigate the cross-layer optimization; to this end, we have explicitly represented the physical network connectivity within our optimal placement model. The combination of DSP applications and SDN networks opens a wide variety of research topics, such as multi-level optimization, reservation and overbooking strategies, and efficient multi- and cross-level adaptation strategies (e.g., whether to relocate an application operator or an entire computing node).

Our placement model can be readily customized to solve similar problems. In the context of SOA applications, we will adapt the optimal placement model and develop a general benchmark that can evaluate decentralized service selection policies, which map abstract services on a subset of available concrete services. Our formulation is particularly suitable for evaluating architecture where the orchestration between services is fully decentralized and the network connectivity between concrete services cannot be neglect.

VII. Acknowledgments

The author would like to thank his supervisor Valeria Cardellini, and Francesco Lo Presti and Vincenzo Grassi for the useful and valuable advices.

References

[1] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," in *Proc. of 1st Workshop on Mobile Cloud Computing*, ser. MCC '12. ACM, 2012.

[2] A. Brook, "Low-latency distributed applications in finance," *Communications of the ACM*, vol. 58, no. 7, 2015.

[3] M. R. Garey and D. S. Johnson, *Computers and Intractability*. San Francisco, Calif.: W. H. Freeman and Co., 1979.

[4] A. Benoit, H. Casanova, V. Rehn-Sonigo, and Y. Robert, "Resource allocation strategies for constructive in-network stream processing," in *Proc. of 2009 IEEE Int'l Symp. on Parallel & Distributed Processing*, ser. IPDPS '09. IEEE Computer Society, 2009.

[5] Q. Zhu and G. Agrawal, "Resource allocation for distributed streaming applications," in *Proc. of 37th Int'l Conf. on Parallel Processing*, ser. ICPP '08, 2008.

[6] G. T. Lakshmanan, Y. Li, and R. Strom, "Placement strategies for Internet-scale data stream systems," *IEEE Internet Computing*, vol. 12, no. 6, 2008.

[7] L. Aniello, R. Baldoni, and L. Querzoni, "Adaptive online scheduling in Storm," in *Proc. of 7th ACM Int'l Conf. on Distributed Event-based Systems*, ser. DEBS '13. ACM, 2013.

[8] A. Chatzistergiou and S. D. Viglas, "Fast heuristics for near-optimal task allocation in data stream processing over clusters," in *Proc. of 23rd ACM Int'l Conf. on Information and Knowledge Management*, ser. CIKM '14. ACM, 2014.

[9] P. Pietzuch, J. Ledlie, J. Shneidman, M. Roussopoulos *et al.*, "Network-aware operator placement for stream-processing systems," in *Proc. of IEEE 22nd Int'l Conf. on Data Engineering*, ser. ICDE '06. IEEE Computer Society, 2006.

[10] S. Rizou, F. Durr, and K. Rothermel, "Solving the multi-operator placement problem in large-scale operator networks," in *Proc. of 19th Int'l Conf. on Computer Communications and Networks*, ser. ICCCN '10, 2010.

[11] B. Nunes, M. Mendonca, X. N. Nguyen, K. Obraczka *et al.*, "A survey of software-defined networking: Past, present, and future of programmable networks," *IEEE Communications Surveys Tutorials*, vol. 16, no. 3, 2014.

[12] J. Huang, G. Liu, Q. Duan, and Y. Yan, "QoS-aware service composition for converged network-cloud service provisioning," in *Proc. of 2014 IEEE Int'l Conf. on Services Computing*, ser. SCC '14, 2014.

[13] Y. Huang, Z. Luan, R. He, and D. Qian, "Operator placement with QoS constraints for distributed stream processing," in *Proc. of 7th Int'l Conf. on Network and Service Management*, ser. CNSM '11, 2011.

[14] C. Thoma, A. Labrinidis, and A. Lee, "Automated operator placement in distributed data stream management systems subject to user constraints," in *Proc. of IEEE 30th Int'l Conf. on Data Engineering Workshops*, ser. ICDEW '14, 2014.

[15] R. Eidenbenz and T. Locher, "Task allocation for distributed stream processing," in *Proc. of IEEE INFOCOM '16*, 2016.

[16] L. Cui, F. R. Yu, and Q. Yan, "When big data meets software-defined networking: SDN for big data and big data for SDN," *IEEE Network*, vol. 30, 2016.

[17] L. Gu, D. Zeng, P. Li, and S. Guo, "Cost minimization for big data processing in geo-distributed data centers," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, 2014.

[18] J. Xu, Z. Chen, J. Tang, and S. Su, "T-Storm: traffic-aware online scheduling in Storm," in *Proc. of IEEE 34th Int'l Conf. on Distributed Computing Systems*, ser. ICDCS '14. IEEE Computer Society, 2014.

[19] T. Heinze, L. Aniello, L. Querzoni, and Z. Jerzak, "Cloud-based data stream processing," in *Proc. of 8th ACM Int'l Conf. on Distributed Event-Based Systems*, ser. DEBS '14. ACM, 2014.

[20] M. Cherniack, H. Balakrishnan, M. Balazinska, D. Carney, U. Cetintemel *et al.*, "Scalable distributed stream processing." in *Proc. of CIDR 2003*, vol. 3, 2003.

[21] A. Toshniwal, S. Taneja, A. Shukla, K. Ramasamy, J. M. Patel *et al.*, "Storm@Twitter," in *Proc. of 2014 ACM SIGMOD Int'l Conf. on Management of Data*, ser. SIGMOD '14. ACM, 2014.

[22] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proc. of USENIX HotCloud'10*, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1863103.1863113

[23] T. Li, J. Tang, and J. Xu, "A predictive scheduling framework for fast and distributed stream data processing," in *Proc. of 2015 IEEE Int'l Conf. on Big Data*, 2015, pp. 333–338.

[24] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma *et al.*, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *Proc. of USENIX NSDI '12*, 2012.

[25] Y. Wu and K. L. Tan, "ChronoStream: Elastic stateful stream computation in the cloud," in *Proc. of IEEE 31st Int'l Conf. on Data Engineering*, ser. ICDE '15. IEEE Computer Society, 2015.

[26] B. Gedik, S. Schneider, M. Hirzel, and K.-L. Wu, "Elastic scaling for data stream processing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, 2014.

[27] V. Gulisano, R. Jiménez-Peris, M. P. no Martínez, C. Soriente, and P. Valduriez, "StreamCloud: An elastic and scalable data streaming system," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, 2012.

[28] T. Heinze, V. Pappalardo, Z. Jerzak, and C. Fetzer, "Auto-scaling techniques for elastic data stream processing," in *Proc. of IEEE 30th Int'l Conf. on Data Engineering Workshops*, ser. ICDEW '14, 2014.

[29] A. Borodin and R. El-Yaniv, *Online computation and competitive analysis*. Cambridge University Press, 2005.

[30] V. Cardellini, V. Grassi, F. Lo Presti, and M. Nardelli, "Distributed QoS-aware scheduling in Storm," in *Proc. of 9th ACM Int'l Conf. on Distributed Event-Based Systems*, ser. DEBS '15. ACM, 2015.

[31] J. Kephart and D. Chess, "The vision of autonomic computing," *Computer*, vol. 36, no. 1, Jan. 2003.

[32] V. Cardellini, V. Grassi, F. L. Presti, and M. Nardelli, "On QoS-aware scheduling of data stream applications over fog computing infrastructures," in *Prof. of 2015 IEEE Symposium on Computers and Communication*, ser. ISCC '15. IEEE Computer Society, 2015.

[33] M. Nardelli, "A framework for data stream applications in a distributed cloud," in *Proc. of ZEUS 2016*, 2016.

[34] V. Cardellini, V. Grassi, F. Lo Presti, and M. Nardelli, "Optimal operator placement for distributed stream processing applications," in *Proc. of 10th ACM Int'l Conf. on Distributed Event-Based Systems*, ser. DEBS '16. ACM, 2016.

[35] I. Stanoi, G. Mihaila, T. Palpanas, and C. Lang, "WhiteWater: distributed processing of fast streams," *IEEE Trans. Softw. Eng.*, vol. 19, no. 9, 2007.

[36] F. Hao, M. Kodialam, T. V. Lakshman, and S. Mukherjee, "Online allocation of virtual machines in a distributed cloud," in *Prof. of 2014 IEEE Int'l Conf. on Computer Communications*, ser. INFOCOM '14. IEEE Computer Society, 2014.

QoS-aware Application Placement Over Distributed Cloud

Francesco Bianchi
Department of Civil Engineering and
Computer Science Engineering
University of Rome "Tor Vergata", Italy
Email: f.bianchi@ing.uniroma2.it

Francesco Lo Presti
Department of Civil Engineering and
Computer Science Engineering
University of Rome "Tor Vergata", Italy
Email: lopresti@info.uniroma2.it

***Abstract*—We are facing an increasing interest about virtualization technologies that enable new paradigms and technical approaches like Cloud Computing, Software Defined Networks. Infrastructure Providers (InPs), managing distributed physical resources, need to perform the mapping of the requested virtual resources on the physical ones. For this purpose, they need to optimize the allocation process in order to: (1) optimally use the various types of available resources shared among customers, and (2) be compliant with Service Level Agreement (SLA) to meet the QoS attributes. This mapping phase, called Virtual Network Embedding, is known to be an NP-hard problem. The aim of the PhD thesis project described in this paper is to propose algorithmic solutions that allow the InP to optimize the mapping of virtual computational nodes and network links in order to improve objective metrics (e.g., revenue) while satisfying multiple QoS attributes agreed in the SLA. To achieve this goal, the thesis will address these main topics: (1) efficient online heuristics for network-aware embedding of multiple network requests, (2) multi-constrained embedding according multiple QoS metrics (e.g., maximum expected delay). Early results show that our algorithms performs better than existent ones, encouraging us to continue with the future work.**

***Index Terms*—Network Virtualization, Cloud Computing, Virtual Network Embedding (VNE), Workflows, Topology-Aware, QoS, Node Ranking, Software Defined Networking (SDN)**

I. Introduction

The Internet is growing very fast in terms of number of hosts, clients, applications, the amount of created data, i.e. Big Data, and exchanged data traffic through the links. Despite this situation, the Internet's structure does not change at the same velocity and seems to reluctant to a radical modification towards new technologies and architectures, due also to the conflicting interests of the several stakeholders. Thus, we are experiencing an "ossification" [1]–[3]. A viable and widely adopted solution is network virtualization [4]–[6] which, thanks to its nature, makes the infrastructure more resilient and customizable, thus virtual networks (VNs) are the base for the future Internet [1], [5], [7]–[9], not just employed to evaluate new services and protocols [2], but being largely used in research testbeds, and to enable new computing models like Cloud Computing [10], [11], Software Defined Networking (SDN) [12]–[14], Data Stream Processing (DSP). Thanks to network virtualization it is possible to supply several VNs into one substrate/physical network enabling the sharing of networking and computing resources for the users in a transparent way. Usually a VN is a set of virtual nodes (with certain assigned CPU, RAM and storage attributes, among others, acting for example as virtual routers) connected through virtual links with certain bandwidth constraints, giving life to various types of widely editable topologies.

In the future Internet, based on the Infrastructure as a Service (IaaS) cloud service model, the classic ISP is replaced by the Service Provider (SP) and the Infrastructure Provider (InP) [15]. InPs which owning and managing the physical infrastructure, lease parts of physical resources to the users of the cloud, namely SPs [16], that make VNs providing end-to-end services to the final users [1], [3], [7], [15]. The basic principle is the sharing of the resources giving life to a multi-tenant business model and a flexible multi-layer infrastructure. In this scenario, the challenge for the InP is how to optimally operate the infrastructure to achieve several goals: maximize the revenue, minimize the costs, satisfy as many requests as possible (or the best using admission control policies), and make efficient use of resources (e.g., computational capacity, disk space, memory, bandwidth of network links and energy). Along with the above goals, InP must be compliant with Service Level Agreements (SLAs) towards customers that expect customized performances on their specific applications. Thus, simultaneously, the InP has to deal with QoS (e.g., the total application delay allowed, the minimum network bandwidth amount) during the mapping/scheduling of the virtual computing nodes and network links on the physical ones. In practice the InPs have to dynamically determine a subset of physical nodes and links with sufficient resources to host each Virtual Network Request (VNR) composed by a set of virtual nodes and links with a certain amount of resource attributes (e.g., CPU, bandwidth, disk storage, RAM, ...) in a way that should be QoS-aware. This mapping process also called Virtual Network Embedding (VNE) is known to be an NP-hard problem [17], [18]. Its solution is a win-win solution for both the InP and the clients benefiting of the advantages.

Because of the complexity of the general resource allocation problem (e.g., the specific VNE problem), several authors address it using heuristic approaches trying to find a trade-off between goodness of the results and runtime of the algorithms. A number of recent works devised some solutions that underline the ability to find a good compromise for the

above challenge. From the first results, they seem to lack in reaching good performance regarding objective metrics. In particular they lack in fully exploiting the physical resources in a balanced way.

The research track of the PhD thesis project reported in this paper, starts trying to find more efficient and effective solutions to the resource allocation problem (beginning with VNE) devising online heuristics addressing various objectives. For this purpose, we implemented a heuristic two-stages topology-aware VNE algorithm able to better detect the global resources of the network, so that it is possible to make a more uniform use of them. Early results show that it outperforms other algorithms in literature, with respect to the benchmark objective metrics.

As previously stated, virtualization techniques and SDNs, help us to devise and implement complex systems, unthinkable until a short time ago, supplied in a dynamic and autonomous way and processing dynamic loads of data. In particular, since we noted that the infrastructure is shared among the users, the InPs are facing another big challenge: meeting various types of SLAs. It is necessary to grant a quality and reliable service according to customized user expectations. The natural improvement of the research track of the thesis project, tries to address this second problem taking into account several QoS attributes (also simultaneously) during the embedding process. This multi-constrained problem, like the resource allocation one, is NP-hard, thus very difficult to deal with. A wider literature search is in progress, but it seems that the works that treat this problem do not deepen all the various aspects. Most of the works usually evaluate just the standard performance metrics without worrying about the final user experience metrics. We have improved our first algorithm so that is able to deal with the latency over the links. Early results, show that our improved algorithm is able to reduce the total end-to-end latency guaranteeing good performance in terms of resource optimization.

The thesis project, on a stepwise refinement basis, is going to cover other resource allocation problems (e.g., mapping of workflows) taking into account the QoS, towards the final objective that involves heterogeneous and distributed resources in the Cloud, making use of SDNs. To do that is necessary to address some challenges that arise: the scalability of the solution, the online and adaptive scheduling, and the problems deriving from the use of distributed networks.

The rest of the paper is organized according to the following structure. The next section introduces the research topics and challenges. The adopted approach and preliminary results are discussed in section III. The future work is described in section IV. In section V we present the related work. Finally, section VI concludes the work.

II. Research Topics and Challenges

The long-term research objective of this thesis project is to address the problem of the resource allocation in a distributed infrastructure, considering the QoS. This main problem has several facets, so can be characterized in several

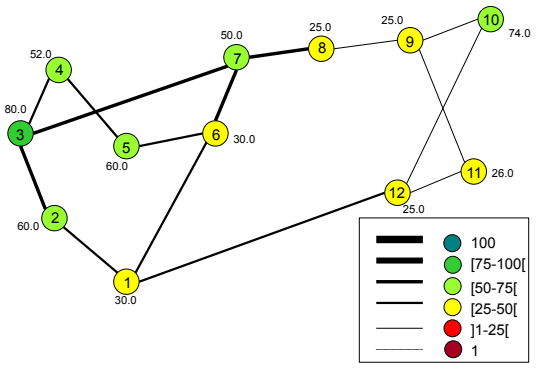


Fig. 1. NSFNET network with available nodes and links resources expressed with values in the range 0-100. CPU values are related with colours of the nodes and bandwidth values are related with thickness of the lines.

sub-objectives/topics that can be also achieved step-by-step. The first one is to help the InPs to achieve a deployment of the requested virtual resources onto physical ones, optimizing the use of available resources while satisfying the SLAs, taking into account multiple QoS attributes. We start applying this topic to the VNE problem. A second topic to be faced is to improve the previous model introducing multiple types of available physical resources that need to be optimized together. A further topic consists in the application/adaptation of the previous optimized and QoS-aware resource allocation to the DSP problem domain, involving the mapping of workflows. Since the InPs will have to manage distributed networks in future (composed of heterogeneous networks and resources), we will deepen also this specific objective. Eventually, the theme of SDNs is of particular interest to test the designed algorithms on realistic testbeds that provide the necessary flexibility along with the possibility to be automatically provisioned/deployed. It is a viable choice letting us to implement the mapping process as easily as in simulation environments.

Facing these naturally interconnected topics, we will have to address some challenges: the need of scalability, the need of run-time operation and adaptivity of the designed solutions, the resource monitoring, and the network control.

The rest of this section will deepen the above topics/objectives.

Resource Optimization and QoS-awareness: Starting with the resource allocation problem of the VNE, we want to help the InPs to answer the question “how to select the physical nodes in order to successfully satisfy the request, optimizing the use of the network thus allowing to accept more future requests maximizing the revenue?”. One of the critical points for an automatic real-time and online algorithm of scheduling, is to evaluate very precisely the real state of the available physical resources of the network. This is in particular true when the embedding process is fully operative and the substrate network becomes almost fully used. One of the main problems claimed by so many literature works, is that the resources become fragmented. In Fig. 1 is reported

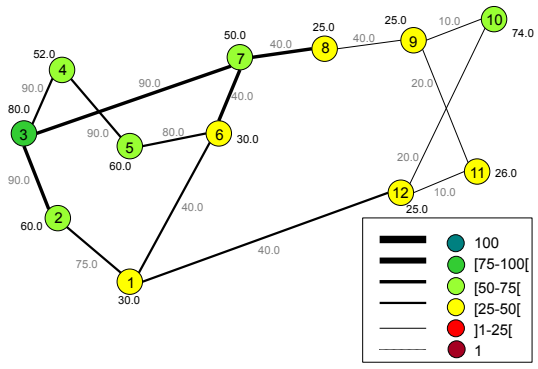


Fig. 2. NSFNET network with available nodes and links resources expressed with values in the range 0-100. CPU values are related with colours of the nodes and bandwidth values are related with thickness of the lines. Latency values are added near each link.

a representation of the real network NSFNET with computational and bandwidth resources. The example reports the state of the network in terms of resources at a certain time of utilization. It is possible to note that node 10 has good computational resources but it does not seem to be an optimal choice to map a virtual node there because the nodes and links around it have quite scarce resources. Probably, a mapping that involves it will result in a rejected request due to insufficient amount of resources. If we model the problem using for example Integer Linear Programming, it is possible to achieve an optimal solution but since the problem is NP-hard, we can adopt it just for small instances of the problem. This scenario is very challenging and means that optimal solutions do not scale according to the dimensions of the problem, so to realize an online scheduler with good speed and performance is a fortiori very challenging.

Furthermore, our aim is to help the InPs to carry out the resource allocation optimizing the resources guaranteeing some minimum performance requirements for virtual networks requested by customers. We need to consider QoS attributes along with various resource metrics. For example, some end-users run applications where the maximum total delay is critic for their results, others ask for a specific geographical location of the employed physical nodes, resulting in constraints on the maximum distance from the identified sites. Several scientific works try to address the last type of constraint (e.g., [19]). A very detailed literature review on this field is in progress, but there are not many works that model the latency/delay (e.g., [20]) or the constraint on the number of used links, to name a few. In Fig. 2 the network has been completed with fixed latency values on the links. In this scenario, it is even more difficult to choose the placement policy. There is an evident trade-off between saving resources and guaranteeing a maximum latency for the entire virtual network request. If latency is crucial for the user, it is worth to map the current virtual node on the right side of the network risking a rejected request. This multi-constrained problem is NP-hard like the resource allocation one, thus devising online scheduling solutions is even more challenging. Since the level of resources and the QoS attributes are dynamic, another challenge is to design a QoS-aware scheduler that automatically and in real-time adapts its policy of mapping according to the expected attributes thresholds of SLAs. Also post-allocation re-optimizations are included.

Multi-metrics Resource Optimization: We need to add multi-metrics to our models. Usually, the works that treat the resource allocation problem (e.g., VNE) adopt a simplified version of the model taking into account just two types of resources: the CPU and the bandwidth. Obviously this does not reflect reality, it's necessary to consider other resource metrics like: amount of memory (RAM), disk storage space, energy consumption.

Mapping of Workflows: Our aim is to design optimized solutions also for the more complex DSP problem domain. It involves the mapping of workflow applications instead of VNRs but it is similar to VNE. In this new scenario we can have fixed source and sink nodes. Also the multi-mapping of virtual nodes/operators on the same physical node is allowed. We will have to address the same challenges previously mentioned.

Distributed and Heterogeneous Cloud: With the advent of Cloud Computing and distributed Clouds, the InPs face the further difficulty of using heterogeneous (e.g., [21]) and distributed networks (e.g., [22]) as if they were a single one. In this new harder scenario, it is very challenging to do online resource allocation granting the QoS to a large amount of users sharing the network that for its nature has different hardware components and different performances. Another challenge to address is to develop improved adaptive scheduling systems because congestions on some parts of the big network and overloading on some nodes/links could affect negatively the overall performances. Also scalability of the solution is always challenging. Another challenge for the InPs is to grant the reliability of nodes, in a distributed environment are necessary effective solutions of resource monitoring and network control. Centralized solutions could become bottlenecks.

SDNs and Cross-layer Orchestration: We need to use SDNs to test our resource allocation solutions. For example, some stream and batch processing frameworks based on Cloud Computing are being deployed over SDNs. There is the need to address the fact that the application framework's scheduler, that allocates tasks to servers, differs from the scheduler of SDN, that allocates network routes [23]. It is necessary to find a mechanism that executes a cross-layer scheduling between the application layer and the network layer. In this way, the coordinated scheduling manages the mapping of the application tasks in combination with the selection of network routes.

III. Approach and Preliminary Results

The thesis project discussed in this paper, aims to address some of the research topics and challenges detailed in the

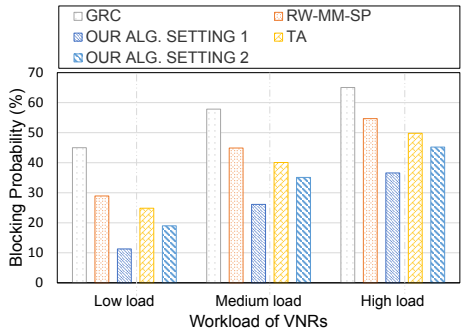


Fig. 3. Performance comparison, in terms of percentage of rejected VNRs, of our algorithm (in two different settings) with other existent algorithms. Under different requests loads.

previous section. At the moment, we have evaluated our solutions through simulation.

As first goal, the thesis addresses the first research topic, namely resource optimization, thus the design and evaluation of efficient and effective online algorithms to solve the resource allocation problem (starting with the specific VNE problem). We devised an algorithm that acting as centralized scheduler, computes in a very effective way the resource embedding potential of each node of the substrate network considering the level of resources in the area of each physical node. It is a heuristic online algorithm based on accumulated rewards of Markov Chains. It consists in a coordinated topology-aware two-stages algorithm.

In the first stage, the node mapping, we compute a very efficient resource capacity metrics for both VNR and substrate network nodes. Like done in [24], for the sake of comparison, we define the amount of available resources of each node n, $CB(n)$, as the product of its computational resources $C(n)$ and the sum of its bandwidth resources $B(e)$ divided by 2, as follows:

$$CB(n) = C(n) \cdot \sum_{e \in NeighE(n)} \frac{1}{2} B(e).$$

Then, normalization is performed:

$$Res(n) = \frac{CB(n)}{\sum_{m \in N} CB(m)}. \quad (1)$$

We formulate the resource metrics for a node n, $V(n)$, in a recursive manner, in the form of weighted summation of the local resources and adjacent nodes metrics, as follows:

$$V(n) = Res(n) + \gamma \sum_{m \in N(n)} \frac{Res(m)}{\sum_{h \in N(n)} Res(h)} V(m), \quad (2)$$

where $\gamma \in [0, 1)$ is the relative weight of the neighbours metrics. In (2), the contribution of each neighbour node to the summation is proportional to their level of resources with respect to the other neighbouring nodes (i.e., $\frac{Res(m)}{\sum_{h \in N(n)} Res(h)}$). Without going into details, if we consider the entire set of equations (one for each node) together, the metrics of node n, $V(n)$, corresponds to its expected discounted accumulated reward of a Markov Random Walk where the factor $\frac{Res(m)}{\sum_{h \in N(n)} Res(h)}$ is the transition probability from node n to node m, $Res(n)$ is the reward associated to the node n, and γ is the discount factor, namely a measure indicating the size of the neighbourhood to be considered in the metrics computation. Higher values of γ mean larger portions of graph taken into account in the metrics computation. The more the value of this metrics for a particular node is high, the more the amount of resources of that node and its neighbouring nodes is high. Thus, the value of this metrics indicates the embedding potential for each physical node, whilst it indicates the requested amount of resources for each virtual node. After computed these values for both the VNR and the substrate network, we rank in descending order the nodes according to these values and finally we carry out a greedy node mapping. So, for all the nodes, we try to map the current virtual node with highest metrics value on to the available physical one with highest metrics value.

In the second stage, we try to map every virtual link on the respective physical path using one or more physical links, computing the shortest path between the involved substrate nodes.

We conducted a wide set of simulations using the ANSNET topology [25] and also random networks. The experimental results show that the proposed heuristics performs significantly better than other algorithms in literature. For this purpose, in Fig. 3, we compare the performance of our algorithm and other existing two-stages topology-aware algorithms (GRC [26], RW-MM-SP [27], and TA [24]), in terms of percentage of rejected VNRs. We set our algorithm with different weights (i.e., parameter γ) to be compliant with the weights adopted by the authors of the other algorithms. Despite the fact that our algorithm works better with higher values of γ, the results show that, under various requests loads, it blocks less VNRs compared to other algorithms. This in turn results in better results in terms of gained revenue for the InP and in terms of utilization of substrate's resources (we omit further plots due to limited space). This algorithm is able to schedule a large number of frequent requests in real-time, it achieves a good balance between speed and goodness of the objective metrics results. Its main feature is the topology-awareness so that it not only coordinates the two stages, but also manages to estimate in a more efficient manner the global resource state of the entire physical network. This is important because the resource state of the network changes continuously in a dynamic way.

The further goal of this thesis project is to improve the above heuristics in such a way that multiple QoS attributes are considered, simultaneously, during the mapping process. We are adding the functionality that allows us to combine QoS metrics and resource metrics for a joint optimization, also using topics from the Graph Theory. This work addresses the challenges of the passage from just checking downstream the compliance or not with the QoS thresholds, to acting upstream

to the factors that affect the obtained QoS metrics values. As first step, we tested this improved algorithm on a VNE problem setting, adding the latency attribute as further optimization parameter. Early results show that our new solution is able to reduce the overall latency of each request, compared to the solution non-latency-aware, limiting the loss of performance in terms of resource profit and classic objective metrics. So a good trade-off can be found. The work is currently in progress.

IV. Future Work

In the section has been detailed the work done till now: some of the several topics of the thesis project have been tackled. The future research will continue along the following directions.

QoS-aware Placement: We need a multi-constrained QoS algorithm which is able to mediate the overall performance over more than one dimension of the problem. Of particular interest, is to obtain more realistic models, is to model the impact of a certain level of use of resource over a QoS attribute (e.g., the impact of the use of bandwidth over the latency value). In order to balance the contrast among different types of objectives (SLAs vs objective metrics and resource metrics), we need to design intelligent policies of admission control along with intelligent scheduling policies which aim is to adapt the optimality of the mapping according to the level of QoS requested (e.g., an application request without delays constraints can be run on high-delay links). Further QoS attributes will be introduced to the model, for example the availability of the links.

Knowing the importance to understand how our heuristic solutions are far from the optimal and suboptimal solutions, we are going to deepen approximation algorithms and the theory of Optimization/Operations Research, in order to carry out a good evaluation of the gap.

Multi-metric Resource Optimization: We intend to address this gap improving our previously mentioned algorithm so that all the realistic resource types will be considered.

Mapping of Workflows: A further step will be adapting our solutions for the VNE problem to the workflow mapping problem. In order to approach this even more difficult scenario, we are going to add admission control policies.

Distributed and Heterogeneous Cloud: This type of infrastructure is for its nature composed by heterogeneous resources asking a fortiori advanced algorithms based on QoS policies. Since the size of the network will be very large, a distributed version of the scheduling algorithm will be crucial to avoid bottlenecks and to grant good run time performances. We need to implement mechanisms to grant the reliability of nodes. The mixed nature of the distributed resources demands always an initial deployment but afterwards the development of an intelligent function of runtime adaptation is highly recommended: so we need to develop migration functions of mapped nodes and links like done in [18]. The introduction of locality constraints as new QoS attributes will be essential.

SDNs and Cross-layer Orchestration: We need to develop and test a cross-layer resource orchestration framework. With a such environment it is possible to develop applications very rapidly through the use of standard APIs. It is possible to create dynamic and on-demand services (e.g., NaaS). We need to expose physical network topology to applications in an efficient way. We will have to deal with data-path representation to find standard mechanisms.

V. Related Work

Due to its complexity, the resource allocation problem (in particular starting with the VNE) has been tackled with various heuristic solutions (e.g., [18], [19], [24], [27]–[33]).

In particular the first works on this field consider just a simplified version of the full problem or carry out the embedding without considering the particular topology features of the VNs and the substrate/physical network [28], [29]. A lot of approaches and categories of VNEs algorithm exist, for example there is the distinction between “one phase” [27], [32] and “two phases” [28], [31] algorithms. The first type performs the mapping of virtual nodes and links at the same time whilst the second type performs the two actions separately. Each one has its own advantages and disadvantages. The former aims to the accuracy and optimality of the results while the latter aims to be faster. In the last period, some works have proposed a kind of algorithms that try to combine the advantages of the above two categories [26] but they seem to lack in reaching good results in terms of objective metrics. In any case is important to underline that the time spent during the mapping phase to schedule the resources is critical in many use cases for example in DSP where we need to map many and frequent workflows without losing performance.

Some works (e.g., [18]) approach the problem dynamically using “migration” and “path splitting” mechanisms.

Houdi et al. [34] implemented a distributed algorithm using a multi-agent approach with exchange of messages between the physical nodes. There is a trade-off between centralized and distributed algorithms: although the first can get overloaded and can be single point of failure, they have a better overall knowledge of the resources achieving often better performances [15]. He et al. [35] proposed the “DaVinci” architecture, where periodically each physical link reallocates bandwidth amount among its virtual links. However, it does not supply a node mapping phase. Still not so many works address the VNE problem performing distributed solutions.

Several works try to model the problem in full detail using Linear Programming to obtain an optimal or approximated solution: Chowdhury et al. [19] use Mixed Integer Programming (MIP) and models the location constraints for the requests.

VI. Conclusion

The thesis described in this work aims to improve efficient online QoS-aware heuristics addressing the VNE problem, towards a more complex scenario where InPs have to manage heterogeneous resources/infrastructures. Granting the QoS and the reliability will be even more challenging. This objective

requires to deepen the main basic research challenges behind: (1) how to exploit near-optimally the physical network to maximize various resource metrics and objective metrics, (2) how to carry out the mapping with the same performances while meeting the SLAs and thus QoS attributes. So far, the thesis work has addressed the first one providing an online heuristic topology-aware algorithm which early results prove to be more efficient of other existing algorithms. The second one is partially addressed but good early results encourage to continue with the future work still to be tackled. The thesis project will be continued adding complexity to the models step by step towards the final test of QoS-aware application on a distributed cloud.

ACKNOWLEDGMENT

I would like to thank Prof. Valeria Cardellini for her very helpful supplementary supervision.

REFERENCES

[1] N. Feamster, L. Gao, and J. Rexford, "How to Lease the Internet in Your Spare Time," *SIGCOMM Comput. Commun. Rev.*, vol. 37, no. 1, pp. 61–64, Jan. 2007.

[2] T. Anderson, L. Peterson, S. Shenker, and J. Turner, "Overcoming the internet impasse through virtualization," *Computer*, vol. 38, no. 4, pp. 34–41, April 2005.

[3] J. Turner and D. Taylor, "Diversifying the internet," in *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, vol. 2, Dec 2005, pp. 6 pp.–760.

[4] A. Khan, A. Zugenmaier, D. Jurca, and W. Kellerer, "Network virtualization: a hypervisor for the internet?" *Communications Magazine, IEEE*, vol. 50, no. 1, pp. 136–143, January 2012.

[5] N. M. K. Chowdhury and R. Boutaba, "A survey of network virtualization," *Comput. Netw.*, vol. 54, no. 5, pp. 862–876, Apr. 2010.

[6] A. Nakao, "Network virtualization as foundation for enabling new network architectures and applications," *IEICE Transactions on Communications*, vol. E93.B, no. 3, pp. 454–457, 2010.

[7] N. Chowdhury and R. Boutaba, "Network virtualization: state of the art and research challenges," *Communications Magazine, IEEE*, vol. 47, no. 7, pp. 20–26, July 2009.

[8] K. Tutschku, T. Zinner, A. Nakao, and P. Tran-Gia, "Network virtualization: Implementation steps towards the future internet," *Electronic Communications of the EASST*, vol. 17, 2009.

[9] A. Berl, A. Fischer, and H. de Meer, "Using system virtualization to create virtualized networks," *Electronic Communications of the EASST*, vol. 17, 2009.

[10] F. Baroncelli, B. Martini, and P. Castoldi, "Network virtualization for cloud computing," *annals of telecommunications - annales des télécommunications*, vol. 65, no. 11-12, pp. 713–721, 2010.

[11] P. Endo, A. de Almeida Palhares, N. Pereira, G. Goncalves, D. Sadok, J. Kelner, B. Melander, and J.-E. Mangs, "Resource allocation for distributed cloud: concepts and research challenges," *Network, IEEE*, vol. 25, no. 4, pp. 42–46, July 2011.

[12] R. Jain and S. Paul, "Network virtualization and software defined networking for cloud computing: a survey," *Communications Magazine, IEEE*, vol. 51, no. 11, pp. 24–31, November 2013.

[13] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "Openflow: Enabling innovation in campus networks," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 69–74, Mar. 2008.

[14] S. Rahtz and H. Oberdiek. (2011) Open networking foundation. [Online]. Available: https://www.opennetworking.org/

[15] A. Fischer, J. Botero, M. Till Beck, H. de Meer, and X. Hesselbach, "Virtual network embedding: A survey," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 4, pp. 1888–1906, Fourth 2013.

[16] G. Schaffrath, C. Werle, P. Papadimitriou, A. Feldmann, R. Bless, A. Greenhalgh, A. Wundsam, M. Kind, O. Maennel, and L. Mathy, "Network virtualization architecture: Proposal and initial prototype," in *Proceedings of the 1st ACM Workshop on Virtualized Infrastructure Systems and Architectures*, ser. VISA '09. New York, NY, USA: ACM, 2009, pp. 63–72.

[17] D. G. Andersen, "Theoretical approaches to node assignment," Dec. 2002, unpublished Manuscript.

[18] M. Yu, Y. Yi, J. Rexford, and M. Chiang, "Rethinking virtual network embedding: Substrate support for path splitting and migration," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 17–29, Mar. 2008.

[19] N. Chowdhury, M. Rahman, and R. Boutaba, "Virtual network embedding with coordinated node and link mapping," in *INFOCOM 2009, IEEE*, April 2009, pp. 783–791.

[20] J. Inführ and G. R. Raidl, "Introducing the virtual network mapping problem with delay, routing and location constraints," in *Proceedings of the 5th International Conference on Network Optimization*, ser. INOC'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 105–117.

[21] J. Nogueira, M. Melo, J. Carapinha, and S. Sargento, "Virtual network mapping into heterogeneous substrate networks," in *Computers and Communications (ISCC), 2011 IEEE Symposium on*, June 2011, pp. 438–444.

[22] I. Houidi, W. Louati, and D. Zeghlache, "Exact multi-objective virtual network embedding in cloud environments," *The Computer Journal*, vol. 58, no. 3, pp. 403–415, 2015.

[23] H. Alkaff, I. Gupta, and L. M. Leslie, "Cross-layer scheduling in cloud systems," in *Cloud Engineering (IC2E), 2015 IEEE International Conference on*, March 2015, pp. 236–245.

[24] S. Zhang, Z. Qian, J. Wu, and S. Lu, "An opportunistic resource sharing and topology-aware mapping framework for virtual networks," in *INFOCOM, 2012 Proceedings IEEE*, March 2012, pp. 2408–2416.

[25] Z. Li and J. J. Garcia-Luna-Aceves, "Finding multi-constrained feasible paths by using depth-first search," *Wirel. Netw.*, vol. 13, no. 3, pp. 323–334, Jun. 2007.

[26] L. Gong, Y. Wen, Z. Zhu, and T. Lee, "Toward profit-seeking virtual network embedding algorithm via global resource capacity," in *INFOCOM, 2014 Proceedings IEEE*, April 2014, pp. 1–9.

[27] X. Cheng, S. Su, Z. Zhang, H. Wang, F. Yang, Y. Luo, and J. Wang, "Virtual network embedding through topology-aware node ranking," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 2, pp. 38–47, Apr. 2011.

[28] Y. Zhu and M. Ammar, "Algorithms for assigning substrate network resources to virtual network components," in *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, April 2006, pp. 1–12.

[29] J. Lu and J. Turner, "Efficient Mapping of Virtual Networks onto a Shared Substrate," Washington University in St. Louis, Tech. Rep. 2006-35, 2006. [Online]. Available: http://openscholarship.wustl.edu/cse_research/186

[30] J. Fan and M. Ammar, "Dynamic topology configuration in service overlay networks: A study of reconfiguration policies," in *INFOCOM 2006. 25th IEEE International Conference on Computer Communications. Proceedings*, April 2006, pp. 1–12.

[31] A. Razzaq and M. Rathore, "An approach towards resource efficient virtual network embedding," in *Evolving Internet (INTERNET), 2010 Second International Conference on*, Sept 2010, pp. 68–73.

[32] J. Lischka and H. Karl, "A virtual network mapping algorithm based on subgraph isomorphism detection," in *Proceedings of the 1st ACM Workshop on Virtualized Infrastructure Systems and Architectures*, ser. VISA '09. New York, NY, USA: ACM, 2009, pp. 81–88.

[33] N. Farooq Butt, M. Chowdhury, and R. Boutaba, "Topology-awareness and reoptimization mechanism for virtual network embedding," in *NETWORKING 2010*, ser. Lecture Notes in Computer Science, M. Crovella, L. Feeney, D. Rubenstein, and S. Raghavan, Eds. Springer Berlin Heidelberg, 2010, vol. 6091, pp. 27–39.

[34] I. Houidi, W. Louati, and D. Zeghlache, "A distributed virtual network mapping algorithm," in *Communications, 2008. ICC '08. IEEE International Conference on*, May 2008, pp. 5634–5640.

[35] J. He, R. Zhang-Shen, Y. Li, C.-Y. Lee, J. Rexford, and M. Chiang, "Davinci: Dynamically adaptive virtual networks for a customized internet," in *Proceedings of the 2008 ACM CoNEXT Conference*, ser. CoNEXT '08. New York, NY, USA: ACM, 2008, pp. 15:1–15:12.

Energy-aware control of server farms

Misikir Eyob Gebrehiwot
School of Electrical Engineering
Aalto University, Finland
Email: misikir.gebrehiwot@aalto.fi

Samuli Aalto
School of Electrical Engineering
Aalto University, Finland
Email: samuli.aalto@aalto.fi

Pasi Lassila
School of Electrical Engineering
Aalto University, Finland
Email: pasi.lassila@aalto.fi

***Abstract*—We study the energy-performance trade-off in server farms using multi-server models with parallel queues. Idle servers may be turned off to save energy. However, turning the server back on incurs a setup cost in the form of additional delays and energy consumption. In our model, a dispatcher assigns incoming jobs to the servers depending on the applied task (job) assignment policy. Once routed to a specific server, a job can either be queued or served immediately depending on the service discipline and state of the server. We consider a family of idle time control policies that keep track of the time spent in an idle state using a timer. Upon expiration of this timer the server is switched to a sleep state, and waits until a new job is assigned to it by the dispatcher. Earlier studies for single server systems, with Poisson arrivals, showed that the optimal control policy either leaves the server idling or switches it to sleep immediately when it becomes idle. However, in a multi-server setting, our study indicates that 10-30 % saving on the system cost can be achieved by using the right combination of idle time control and task assignment policy.**

I. Introduction

Energy efficiency of server farms in data centres has gained the attention of researchers in recent years [1], [2]. This can mainly be attributed to the fact that the servers spend $70-80\%$ of the time in an idle state [3], wasting about 70% of their peak power in the process [4]. Although an idle server can be turned off, or put to sleep, to save energy, turning it back on involves a substantially long setup delay. In most cases, the power consumption is also at its peak during this time.

Among other things, server farm performance and energy consumption values depend on how incoming requests are assigned to the servers, and what control actions are taken when a server in the server farm becomes idle. In queueing theoretic models, these are studied as *task assignment policies* and *idle time control policies*, respectively.

Energy-performance trade-off has been recently studied for single server systems, and optimal solutions under a variety of system settings and cost metrics have been derived, see [5]–[8]. However, due to the high number of possible task assignment and control policy combinations, the optimal control of a multi-server system has proven to be challenging.

The two models mainly used in the study of multi-server systems are the central and parallel queueing models. In a central queueing model, arriving requests are queued centrally and processed by the servers according to the implemented scheduling policy. In the case of parallel queueing models, each server has its own queue, and arriving requests are assigned to one of these queues by the dispatcher depending on the implemented task assignment policy.

In this paper, we consider a system where the idle time of servers can be controlled by a timer. When a timer expires, the server associated with that timer is switched to a sleep state to save energy. We give a summary of optimality results available in the literature for a single server system. In this case, when requests arrive according to a Poisson process, there is no actual role for the idling timer, because the optimal control either leaves the server idle or puts it to sleep immediately. In a multi-server parallel queue setting, the role of the idle timer in optimizing the energy-performance trade-off is not that straightforward. We study the impact of different task assignment and control policies on a homogeneous multi-server system with parallel queues, using a C^{++} based event driven simulator. We show that $10-30\%$ saving on system cost can be achieved by choosing the right combination of task assignment and idling time control policy. In this case, the optimal idling time value could have any finite value. Additionally, we show that a similar improvement can be achieved by leaving an optimally selected number of servers in an idle state, and switching the remaining servers to a sleep state when they become idle.

The rest of the paper is organized as follows. In the following section, the system model is defined formally. Section III covers the analysis of single server systems. A simulation based study of multi-server systems is given in Section IV. Concluding remarks and future directions are given in SectionV.

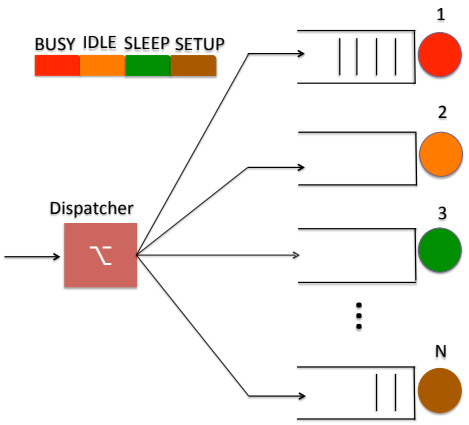


Fig. 1: System model

A. Related work

Gandhi et al. [5], [9] considered a homogeneous server farm with a central queue and exponential service times. In [10], they also conducted a trace based study and proposed the DELAYEDOFF control policy to be implemented with the *Most-Recently-Busy* (MRB) task assignment policy. DELAYEDOFF, as its name indicates, delays turning a server off by using an idling timer, and the MRB task assignment policy routes a job to the server that has been busy most recently (provided there are more than one idle servers). These policies will be discussed formally in later sections.

The total number of idle servers can also be controlled by allocating a threshold to the maximum number of servers that can stay idle simultaneously. These control policies are studied in [9], [11]–[14]. Adopting the notation in [9], we represent these policies by BUSY/IDLE(t), where t represents the threshold set for idle servers.

In the case of server farms with parallel queues, the existing literature has mainly focused on performance improvement, with little or no attention given to the energy aspect. In [15], the *Join-the-Shortest-Queue* (JSQ) is shown to minimize waiting time for a homogeneous multi-server system with exponential service times and Poisson arrivals. Energy-aware task assignment in a parallel server setting is studied in [16]–[18] by applying Markov Decision Process (MDP).

II. System model

We consider a system of N homogeneous parallel servers, see Figure 1. Service requests arrive to a dispatcher according to a Poisson process with rate λ, and are assigned to one of the servers based on the task assignment policy in use. Denoting the generic service times by S and their cumulative probability distribution by $F(s)$, the load on the system is given by $\rho = \lambda \mathrm{E}[S]$. We assume a stable system with $\rho < N$. In a single server case we consider the FIFO and PS scheduling policies, while our multi-server study focuses only on FIFO.

In addition to the *busy* and *idle* states, we assume that each server has at least one *sleep* state. Turning a server on from a sleep state involves a *setup* state during which the server is not available to process requests. We denote the associated setup delay by a random variable D. Further, we assume that

$$0 \leq P_{\text{sleep}} < P_{\text{idle}} < P_{\text{setup}} \leq P_{\text{busy}},$$

where P_x denotes the power consumption in state x.

Idling time control policies

In this setting, the following family of control policies will be considered:

NEVEROFF: Under this policy, all the servers are kept running even if they are idle, without regard to the energy wasted in the process.

INSTANTOFF: This policy switches a server to *sleep* immediately when it becomes idle. The server is turned on and switched to the *setup* state as soon as a new job is routed to it

DELAYEDOFF: These policies use timers to control the amount of time spent in an *idle* state. Each server is assigned its own timer. Thus, we have a generic list of timers as $\{I_1, I_2, ..., I_N\}$. For $1 \leq n \leq N$, if the idling timer I_n expires before a new request arrives, the nth server will be turned off. On the other hand, if a service request arrives to that specific server before the timer expires, the timer I_n is reset and the server starts working on the request immediately. The idling timers, I_n, are i.i.d random variables having general distributions.

BUSY/IDLE(t): For $0 \leq t \leq N$, these policies keep servers idling until the threshold, t, is reached. Any server that becomes idle afterwards is switched to a sleep state immediately. Note that NEVEROFF and INSTANTOFF are special cases of the BUSY/IDLE(t) family of control policies when $t = N$ and $t = 0$ respectively.

Task assignment control policies

We assume that the dispatcher has no a priori knowledge about the size of requests. However, the assumption is that it has instantaneous information about the servers, such as the state of each server and the number of requests waiting and/or being served in each server. With this assumption, we consider the *random (RND)*, *join the shortest queue (JSQ)* and *most recently busy (MRB)* task assignment policies.

RND assigns incoming requests randomly from the uniform distribution to the available servers, and hence does not need any additional information to make a forwarding decision. On the other hand, JSQ needs to keep track of the number of jobs being served in each server in order to assign a request to the server with the fewest jobs.

The MRB task assignment policy, like JSQ, needs to know the state of all the servers for its routing decision. Upon the arrival of a job, the state of all servers is checked and the job is forwarded to an idle server. If two or more servers are found to be idle, the most recently busy server is chosen. In other words, the job is routed to the server with the largest remaining idling time, allowing the other servers to run down their idling time (depending on the control policy implemented) and be turned off. However, there might not be an idle server on the arrival of a job. In this case, we will look into two variants of the MRB task assignment policy:

MRB-RND: This policy applies the MRB routing if one or more servers are in the idle state at the arrival time of a job. If there is no idle server, the arriving job will be routed randomly.

MRB-JSQ: In this case, MRB routing is applied first. If there is no idle server, an arriving job is assigned to the queue with the fewest jobs.

Metrics

The popular cost metrics associated with the study of energy-performance trade-off in server farms are the *Energy-Response time-Weighted-Sum* (ERWS) [6], [19],

$$w_1 \mathrm{E}[T] + w_2 \mathrm{E}[P], \tag{1}$$

and *Energy-Response time-Product* (ERP) [5], [7],

$$\mathrm{E}[T]\mathrm{E}[P], \tag{2}$$

where $w_1$and w_2 are weighting factors. Other cost metrics have also been applied in [6], [20].

III. Analysis of optimal single server control policies

There are several studies in the literature focusing on the energy-performance trade-off for a single server system with Poisson arrivals [5]–[8], [17], [18]. The energy-aware single server systems can be modelled as one of the parallel servers in Figure 1 without the dispatcher. Moreover, let Π denote the set of all idling time controls, i.e., Π = {NEVEROFF, INSTANTOFF, DELAYEDOFF}. Assuming finite valued idling timer I, the system dynamics would be as follows. The server would alternate between the *busy* and *idle* states, and eventually go to *sleep* when the idle timer expires. It is then put into *setup* at the arrival of the next request, and becomes *busy* again, once setup is finished.

Denoting the expiration of the idle timer as a regeneration point, there could be several visits to the *idle* state within a regeneration cycle. We denote the total sum of such idle periods within a cycle by I^{tot}. This can be expressed in terms of the number of individual idle times, and the length of each of these periods. Thus, the mean total idle time in a cycle is given by

$$\mathrm{E}[I^{\text{tot}}] = \frac{\mathrm{E}[\min\{I, A\}]}{P\{I < A\}},$$

where A represents the exponentially distributed interarrival times, see [7] for a more detailed discussion.

Energy-aware M/G/1-FIFO

For such a system, under the FIFO service discipline, the mean response time, $\mathrm{E}[T]$, and the mean power consumption, $\mathrm{E}[P]$, are given by

$$\mathrm{E}[T] = \mathrm{E}[T_{\text{M/G/1}}] + \frac{\frac{1}{\lambda}\mathrm{E}[D] + \frac{1}{2}\mathrm{E}[D^2]}{\mathrm{E}[I^{\text{tot}}] + \frac{1}{\lambda} + \mathrm{E}[D]}, \tag{3}$$

and

$$\mathrm{E}[P] = \mathrm{E}[P_{\text{M/G/1}}] + (1-\rho)\frac{\frac{1}{\lambda}(P_{\text{sleep}} - P_{\text{idle}}) + \mathrm{E}[D](P_{\text{setup}} - P_{\text{idle}})}{\mathrm{E}[I^{\text{tot}}] + \frac{1}{\lambda} + \mathrm{E}[D]}, \tag{4}$$

where $\mathrm{E}[T_{\text{M/G/1}}]$ and $\mathrm{E}[P_{\text{M/G/1}}]$ are the mean response time and mean power consumption of an ordinary M/G/1-FIFO queue, respectively, and are given by

$$\mathrm{E}[T_{\text{M/G/1}}] = \mathrm{E}[S] + \frac{\lambda \mathrm{E}[S^2]}{2(1-\rho)},$$
$$\mathrm{E}[P_{\text{M/G/1}}] = \rho P_{\text{busy}} + (1-\rho)P_{\text{idle}}.$$

Due to lack of space, we refer the reader to [7] for the detailed derivation.

Energy-aware M/G/1-PS

For an energy-aware M/G/1 - PS system, the mean conditional response time of a job of size t can be decomposed into

$$\mathrm{E}[T|S=t] = \mathrm{E}[Q] + U(t), \tag{5}$$

where $\mathrm{E}[Q]$ is the mean queueing time of the job, if it happens to arrive in the *sleep* or *setup* states, and it is given by

$$\mathrm{E}[Q] = (1-\rho)\frac{\frac{1}{\lambda}\mathrm{E}[D] + \frac{1}{2}\mathrm{E}[D^2]}{\mathrm{E}[I^{\text{tot}}] + \frac{1}{\lambda} + \mathrm{E}[D]}. \tag{6}$$

$U(t)$ is the mean conditional processing time of a job of size t, and it needs to be solved from the following integral equation,

$$U'(t) = 1 + 2\lambda \mathrm{E}[Q]\bar{F}(t) + \lambda \int_0^\infty U'(y)\bar{F}(t+y)\,dy + \lambda \int_0^t U'(y)\bar{F}(t-y)\,dy, \tag{7}$$

where $\bar{F}(t) = 1 - F(s)$. See [8] for the derivation.

Proposition 1: For an energy-aware M/G/1 system with setup delay, the optimal policy in Π under the ERWS (1) and ERP (2) cost metrics is either NEVEROFF or INSTANTOFF. This applies for both FIFO and PS scheduling.

See [8, Theorem 2] and [7, Propositions 1-4] for the proof.

Thus, for a single server system, with Poisson arrivals, the optimal control for a large class of cost metrics remains simple: that is, it is either NEVEROFF or INSTANTOFF. This simple result can be attributed to the memoryless property of Poisson arrivals. That is, the probability of having a new arrival after spending t units of time in an *idle* state is the same as the probability of having an arrival at the start of the *idle* state. Hence, the server might, as well, have been put to sleep immediately when it becomes idle to save energy, or could still be left idling to improve performance.

However, the optimal choice, among the NEVEROFF and INSTANTOFF policies, depends on the load, the setup delay of the *sleep* state and its power consumption. For a fixed setup delay value, lower load values would favour the INSTANTOFF policy, while NEVEROFF would be the optimal choice at higher load values.

In a multi-server setting, the choice of task assignment policy impacts the total cost of the system. Thus, the optimal control may not be summarized into two simple policies as is the case in single server models. In the following section, we study the energy-aware control of server farms by considering a collection of idle time control and task assignment policies.

IV. Energy-aware Multi-server control policies

We conduct a numerical study of the energy-performance trade-off in a server farm with 10 servers. Power consumption values of $P_{\text{busy}} = P_{\text{setup}} = 240$ W and $P_{\text{idle}} = 150$ W are used. We assume exponentially distributed service times, with mean $\mathrm{E}[S] = 1$ s. Both setup and idling time are assumed to be deterministic with idling time reset every time the server goes out of the idle state. Weighting factors of $w_1 = 1$ and

TABLE I: Combined mode of operation of the task assignment and idle time control policies

	NEVEROFF	INSTANTOFF	DELAYEDOFF	BUSY/IDLE(t)
RND	Incoming requests are routed randomly and servers are in either the *busy* or *idle* states.	Servers can be in any of the *busy*, *sleep* or *setup* states, and incoming request is routed randomly irrespective of the current state of servers.	Servers can be in any of the *busy*, *idle*, *sleep* or *setup* states, and incoming requests are routed randomly irrespective of the current state of servers.	There are at most t idle servers and at least $10-t$ in busy, sleep, or setup states. An incoming request is routed to an idle server, if any. Otherwise it is routed randomly.
MRB-RND	An arriving request is routed to the most recently busy server if there are multiple servers in the *idle* state, otherwise routing is done randomly. However, idle servers are not turned off.	No idle servers. Thus jobs are routed to any of the servers randomly. Note that this mode of operation is the same as the RND and INSTANTOFF combination.	If there are multiple servers in the *idle* state, an incoming job is routed to the most recently busy server, i.e. to the server with the largest remaining idling timer. Otherwise jobs are routed randomly. Servers are allowed to stay idle until the timer expires.	An incoming job is routed to an idle server (if any) that has been busy most recently. Otherwise jobs are routed randomly. Note that due to the BUSY/IDLE(t) policy there are no idling timers, hence the combination does not fully exploit the functionality of MRB.
MRB-JSQ	Behaves the same way as the combination of MRB-RND and NEVEROFF with RND replaced by JSQ	Behaves the same way as the combination of MRB-RND and INSTANTOFF with RND replaced by JSQ	Behaves the same way as the combination of MRB-RND and DELAYEDOFF with RND replaced by JSQ	Behaves the same way as the combination of MRB-RND and BUSY/IDLE(t) with RND replaced by JSQ

$w_2 = 0.1$ are used for the ERWS (1) cost metric. With these system parameters, the stochastic behaviour of the system given in Section II is studied using a C^{++} based discrete event simulator. For every case we carried out 10 independent simulations, each of them having 10^5 requests.

Our study focuses on the NEVEROFF, INSTANTOFF, DELAYEDOFF, and BUSY/IDLE(t) idle time control policies in combination with the RND, MRB-RND, and MRB-JSQ task assignment policies. Table I describes the mode of operation under the two family of policies.

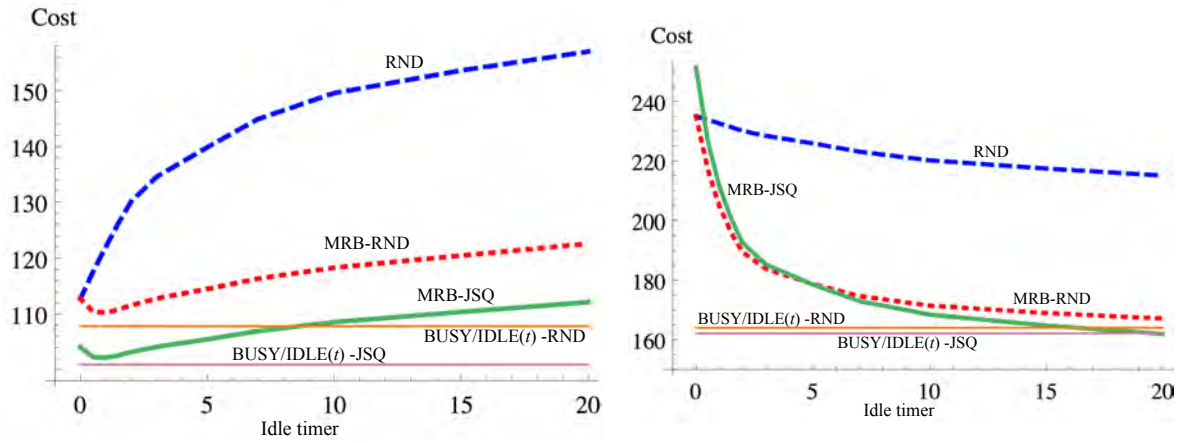
Figures 2-4 illustrate the ERWS cost of the system under a variety of load and setup delay values. The BUSY/IDLE(t) policy has t NEVEROFF and $10-t$ INSTANTOFF servers, and hence it has only two possible values of idling time. In the figures, it is plotted as a function of idling time only for the sake of comparison. Clearly, the value of the threshold t impacts the total cost of the system. Hence, a separate simulation was run to determine the optimal value of t that minimizes the system cost under the BUSY/IDLE(t) policy for each combination of load and setup time. Moreover, for notational convenience, we represented the combination of BUSY/IDLE(t) and RND task assignment policy by BUSY/IDLE(t)-RND, while that of BUSY/IDLE(t) and JSQ is BUSY/IDLE(t)-JSQ.

Figure 2 illustrates the ERWS cost metric of a lightly loaded system as a function of the idling timer. Setup delay values of $D = 0.5\mathrm{E}[S]$ (left panel) and $D = 6\mathrm{E}[S]$ (right panel) are considered. The leftmost part of the figure ($I = 0$ s) represents the INSTANTOFF policy. Any other value of the idling timer represents a DELAYEDOFF policy. As we approach the right most part of the figure ($I = 20$) s, we can safely assume that the control policy in action is close to NEVEROFF. In both the right and left panels, the ERWS cost of the RND task assignment policy is monotonous, indicating that the optimal control policy is either the NEVEROFF or INSTANTOFF. This, in fact, can also be deduced from Proposition 1 by applying the Poisson splitting property.

However, in the case of MRB-RND and MRB-JSQ task assignment policies, DELAYEDOFF provides the lowest ERWS cost. This is more visible in Figure 3 where the load value is 5. In the left panel, DELAYEDOFF has about 10% improvement in ERWS cost compared to NEVEROFF. On the other hand, the right panel shows that ERWS cost of DELAYEDOFF is about 20% lower than that of INSTANTOFF.

Figure 4 depicts the ERWS cost of a highly loaded server farm. Even at high load the MRB-JSQ task assignment policy has its lowest ERWS cost somewhere in between 0 and ∞, when the setup delay is low (left panel). For a larger setup delay value, the ERWS cost of all the task assignment policies appear to be a monotonously decreasing function of the idle timer. In almost all the cases studied, the combination of BUSY/IDLE(t) and JSQ has the minimum or near-minimum ERWS cost. However, keep in mind that the value of t had to be optimized with respect to the load to yield this result.

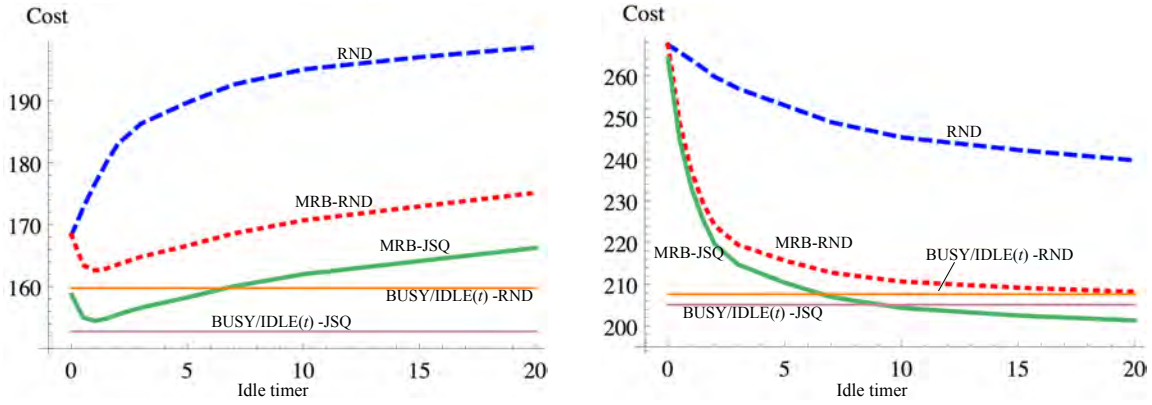
From all the figures, we can see that the single most important factor, in determining which combination of task assignment and control policy give the optimal result, is the setup delay of the system. When the setup delay is shorter compared to the mean request size, the rule of thumb is to either use DELAYEDOFF with MRB-JSQ or the BUSY/IDLE(t) with JSQ, by optimizing with respect to t or the idling timer



(a) Load = 3 and $D = 0.5\mathrm{E}[S]$

(b) Load = 3 and $D = 6\mathrm{E}[S]$

Fig. 2: ERWS cost of the system under the RND, MRB-RND, MRB-JSQ and BUSY/IDLE(t) policies with load = 3.



(a) Load = 5 and $D = 0.5\mathrm{E}[S]$

(b) Load = 5 and $D = 6\mathrm{E}[S]$

Fig. 3: Total cost of the system under the RND, MRB-RND, MRB-JSQ and BUSY/IDLE(t) policies with load = 5.

value depending on the load. Our results show that these combinations of task assignment and control policies can have up to 30 % ERWS saving compared to INSTANTOFF and up to 10 % saving compared to NEVEROFF.

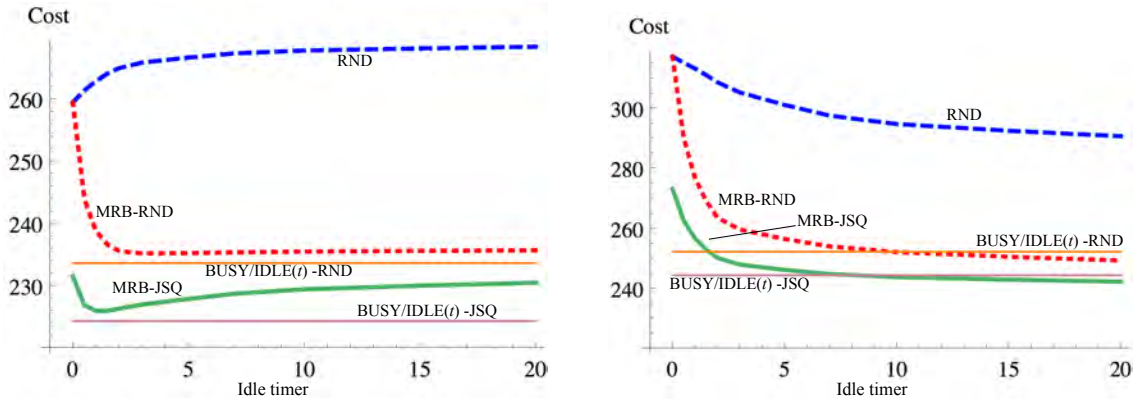
V. Conclusions

An attempt to reduce energy consumption often comes with a decline in performance. Thus, any applicable solution should strike the right balance between energy and performance. To study this, we used the weighted sum of mean response time and mean power (ERWS) metrics as a representation of the system cost. In the case of energy-aware single server systems, closed form expressions are available in the literature for performance and energy metrics. Studies show that either NEVEROFF or INSTANTOFF yield the optimal result.

In a multiserver system, the applied task assignment policy impacts the energy efficiency of the system in addition to the idle time control policy. A numerical analysis of the RND, MRB-RND and MRB-JSQ task assignment policies was performed in combination with the BUSY/IDLE, BUSY/IDLE(t), BUSY/OFF and DELAYEDOFF control policies. In addition to low, medium and high load values, setup time was varied relative to service time requirements and the resulting system was studied. The combination of the MRB-JSQ task assignment policy and the DELAYEDOFF control policy or that of JSQ and BUSY/IDLE(t) are found to reduce up to 30% of the system cost.

VI. Future work

In this simulation based study, we considered the energy-aware control of homogeneous multi-server systems under the assumption that requests arrive to the dispatcher according to a Poisson process. Request sizes are also assumed to be exponentially distributed, and the Energy-Response time-Weighted-Sum cost metric was utilized. Due to these restrictive



(a) Load = 8 and $D = 0.5\mathrm{E}[S]$

(b) Load = 8 and $D = 6\mathrm{E}[S]$

Fig. 4: ERWS cost of the system under the RND, MRB-RND, MRB-JSQ and BUSY/IDLE(t) policies with load = 8.

assumptions, the results can not be generalized to describe the energy-performance trade-off in server farms, whose arrival patterns and request sizes often can not be modelled by an exponential distribution. As part of our future work, we would like to consider more general request sizes and arrival patterns, with more general cost models that does not need a closed form expression for the mean response time and energy consumption.

ACKNOWLEDGEMENT

This research has been partially supported by the TOP-Energy project funded by the Academy of Finland (grant no. 268992).

REFERENCES

[1] J. Pierson, *Large-scale Distributed Systems and Energy Efficiency: A Holistic View*. Wiley, 2015.

[2] Y. Zhang and N. Ansari, "On architecture design, congestion notification, TCP incast and power consumption in data centers," *IEEE Communications Surveys Tutorials*, vol. 15, no. 1, pp. 39–64, 2013.

[3] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, pp. 50–58, 2010. [Online]. Available: http://doi.acm.org/10.1145/1721654.1721672

[4] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, Dec 2007.

[5] A. Gandhi, V. Gupta, M. Harchol-Balter, and M. A. Kozuch, "Optimality analysis of energy-performance trade-off for server farm management," *Perform. Eval.*, vol. 67, no. 11, pp. 1155–1171, Nov. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.peva.2010.08.009

[6] V. J. Maccio and D. G. Down, "On optimal policies for energy-aware servers," in *Proc. of IEEE 21st International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS 2013)*, Aug. 2013, pp. 31–39.

[7] M. Gebrehiwot, S. Aalto, and P. Lassila, "Optimal sleep-state control of energy-aware $M/G/1$ queues," in *Proc. of the 8th International Conference on Performance Evaluation Methodologies and Tools*, Bratislava, Slovakia, Dec. 2014.

[8] M. E. Gebrehiwot, S. Aalto, and P. Lassila, "Energy-performance trade-off for processor sharing queues with setup delay," *Operations Research Letters*, vol. 44, no. 1, pp. 101 – 106, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167637715001637

[9] A. Gandhi, M. Harchol-Balter, and I. Adan, "Server farms with setup costs," *Perform. Eval.*, vol. 67, no. 11, pp. 1123–1138, Nov. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.peva.2010.07.004

[10] A. Gandhi, M. Harchol-Balter, R. Raghunathan, and M. A. Kozuch, "Distributed, robust auto-scaling policies for power management in compute intensive server farms," in *Open Cirrus Summit (OCS), 2011 Sixth*, Oct 2011, pp. 1–5.

[11] C. Schwartz, R. Pries, and P. Tran-Gia, "A queuing analysis of an energy-saving mechanism in data centers," in *The International Conference on Information Network 2012*, Feb 2012, pp. 70–75.

[12] I. Mitrani, "Managing performance and power consumption in a server farm," *Annals of Operations Research*, vol. 202, no. 1, pp. 121–134, 2013.

[13] P. J. Kuehn and M. E. Mashaly, "Automatic energy efficiency management of data center resources by load-dependent server activation and sleep modes," *Ad Hoc Networks*, vol. 25, Part B, pp. 497 – 504, 2015.

[14] V. J. Maccio and D. G. Down, "Exact analysis of energy-aware queueing policies," 2016, technical report.

[15] W. Winston, "Optimality of the shortest line discipline," *Journal of Applied Probability*, vol. 14, no. 1, pp. 181–189, 1977. [Online]. Available: http://www.jstor.org/stable/3213271

[16] A. Penttinen, E. Hyytiä, and S. Aalto, "Energy-aware dispatching in parallel queues with on-off energy consumption," in *Proc. of IEEE 30th International Performance Computing and Communications Conference (IPCCC 2011)*, Nov. 2011.

[17] E. Hyytiä, R. Righter, and S. Aalto, "Task assignment in a heterogeneous server farm with switching delays and general energy-aware cost structure," *Perform. Eval.*, vol. 75-76, pp. 17 – 35, 2014.

[18] ——, "Energy-aware job assignment in server farms with setup delays under lcfs and ps," in *26th International Teletraffic Congress (ITC)*, Sept 2014, pp. 1–9.

[19] A. Wierman, L. L. H. Andrew, and A. Tang, "Power-aware speed scaling in processor sharing systems: Optimality and robustness," *Perform. Eval.*, vol. 69, no. 12, pp. 601–622, Dec. 2012. [Online]. Available: http://dx.doi.org/10.1016/j.peva.2012.07.002

[20] S. Kaxiras and M. Martonosi, *Computer Architecture Techniques for Power-Efficiency*, 1st ed. Morgan and Claypool Publishers, 2008.

SDN based service provisioning management in smart buildings

M. Tosic*, O. Ikovic * and D. Boskovic*
* Foundation VizLore Labs, Novi Sad, Serbia
milenko.tosic@vizlore.com, ognjen.ikovic@vizlore.com, dragan.boscovic@vizlore.com

Abstract - Foundation of every smart building is local area network connecting devices and systems for enabling provisioning of in-building and Internet based services. In the Internet of Things era, in-building local area networks are becoming more complex with support for plethora of access technologies, support for edge processing, interconnecting IoT gateways and supporting different mission critical services depending on low latency and high reliability all while supporting internal company business processes and high QoS for Internet based services like web browsing, content delivery and video streaming. It is evident that smart building local area networks require proper solution for context and service aware resource management. In this paper we introduce SDN-based approach and controller for context aware service provisioning management. We focus on mission critical services of building management system, VoIP for in-building voice communication and content delivery. We will show that our SDN controller enables context and service aware utilization of edge processing and storage capacities of the in-building IP infrastructure to ensure required QoS levels for these different classes of in-building and Internet based services.

I. INTRODUCTION

Internet of Things (IoT) enables sensing of the real world information through Internet services and applications. One of the main fields where the IoT is expected to have large positive impact are smart buildings and building management systems (BMS). At the basis of IoT enabled smart buildings are local area networks (LAN) for interconnecting occupants, sensors, appliances building infrastructure and services.

IoT systems and infrastructure in smart buildings are utilized for managing building resources and automating building processes in order to improve efficiency, reduce costs and improve comfort. Additionally, IoT systems and connected devices are rich source of contextual data which can be used for context aware service provisioning in the shared building LAN. In this paper we will focus on utilization of contextual insight obtained through smart building IoT system for Software Defined Networking (SDN) based management of service and resource provisioning throughout in-building LAN. First, we address networking trends set by IoT for smart buildings and requirements which arise from them and which defined features of our smart building SDN controller.

Trend 1 - modern mobile users. Modern mobile users are always connected and use Internet based services and applications for day to day decision making and activities. Their requirements and behavioral patterns guide resource and service adaptation. Requirement 1 - properly include contextual data coming from end users and their devices into SDN based decision making processes. For legacy network management procedures, client (mobile user) is approximated with his device (i.e. mobile phone or PC). In the era of wearables every network client is in fact an ad-hoc network in his own rights, comprising various connected devices communicating through different radio technologies. Therefore, the complexity of the access network is exponentially increasing and proper network resource and service provisioning management needs new set of tools and techniques which can properly scale and adapt to needs of modern mobile users.

Trend 2 - Smart Buildings are Big Data. The real value behind IoT is data rather than devices. Collecting contextual and sensory data from more building services and equipment provides a better insight into how exactly are buildings performing, but at the same time brings smart buildings closer to the domain of big data. Requirement 2 - distributed data analysis and decision making processes across the whole IoT ecosystem. BMSs need actionable insights to optimize building operations, increase energy efficiency, and improve occupants' comfort. Decision making processes can be divided into reactive and proactive. Reactive decision making provides actionable insight for real time management of resources and services. Proactive decision making performs strategic planning of smart building resources and infrastructure and derives policies for system modification. For reactive decision making it is important to detect triggers, status changes and alarms in monitored parameters. Therefore, required data analysis is simple, but needs to be achieved in real time and reliably. On the other hand, proactive decision making is based on in-depth analysis of collected contextual and sensory data focused on finding patterns, correlations and dependencies between system elements. Therefore, data analysis for proactive decision making requires a scalable platform with sufficient processing and storage resources for offline data crunching. In well designed IoT systems, reactive decision making utilizes edge processing resources while proactive decision making relies on cloud platforms.

Trend 3 - IP infrastructure as a building utility and the main enabler for provisioning of smart building services. Building owners and management companies are increasingly installing IP infrastructures to provide Internet to the residents but also to unify all kinds of different control subsystems that comprise BMSs. Current trends show that the cloud is descending to the network edge (IoT GWs, client personal devices and networking elements). This trend is called fog networking/computing [1], [2]. Properly managed smart building LAN can efficiently utilize this fog/edge computing potential for improving reactive decision making. Requirement 3 - unified smart building IP infrastructure and platform.

Networking technologies involved in local area networks and BMSs are often positioned as alternatives trying to replace one another (i.e. Bluetooth vs. Zigbee, WiFi vs. 3G, fiber optics vs. CAT6 etc.). Our position is that different technologies should complement each other (cost vs. performance, range vs. capacity, precision vs. scalability) in order to support pervasive service provisioning. All technologies have their place in IoT smart buildings. What they need is an unified IP framework able to properly utilize and combine different communication technologies for seamless and adaptive service provisioning.

Based on the identified trends and requirements in IoT era smart buildings, we have developed SDN controller for managing resources and service provisioning throughout smart building LAN. We acknowledge central role of in-building LAN for enabling and managing smart building services. Through context informed management of networking resources we are able to provide context aware and adaptive smart building service provisioning alongside other Internet of Everything (IoE) services all while ensuring highest levels of efficiency in resource utilization and maintaining required QoS levels.

II. Related Work

The latest achievements in automation have always been adapted for the needs of better management of building resources. University of California at Berkeley with their Software Defined Buildings family of projects [3], [4], [5] proposes different approaches for standardizing management of smart building resources and services. Their research efforts are focused on management of smart building efficiency. Our smart building SDN controller can complement their solution with management of networking resources and service provisioning throughout the shared in-building LAN.

SDN is created for management of data centers and core/distribution switches and routers. New trend is bringing the SDN logic to wireless access segment of a network (WiFi and cellular) [6]. SDN based network management in the IoT era cannot be achieved without comprehensive understanding of the environment with which the network is integrated. End users have never been as closely integrated into the networking services and infrastructure as in IoT systems like smart buildings and smart cities. Therefore, knowledge about how they utilize networking resources and smart building services should guide decision making algorithms for load balancing, handover and traffic engineering. Previous research focus on SDN techniques and their roles in sensory networks [7] and integration with context publish subscribe techniques for enabling efficient distribution of sensory and contextual data throughout IoT system of a smart building [8]. In [9] authors address main challenges for enabling SDN-based management of IoT services. What they neglect to address is the fact that smart buildings are based on LAN which is shared between smart building services and other Internet based services. Our SDN smart building controller jointly manages smart building services and other services which coexist on the shared LAN and compete for resources.

In our previous work [10] we have introduced concept of soft sensors and their role in enabling SDN-based management of networking resources. In this paper we continue with presentation of our work in context of smart buildings and resource and service provisioning challenges typical for these complex IoT systems. We will present three decision making solutions for efficient utilization of in-building LAN resources while addressing QoS requirements of different smart building services.

III. SDN controller for context aware service provisioning in smart buildings

SDN controller for smart buildings needs to be able to fulfill the following requirements: 1. Joint management of networking resources for ensuring uninterruptible support for mission critical smart building services while satisfying QoS requirements of other Internet based services (e.g. content delivery); 2. Enhancing smart building's edge processing potential through proper management of networking resources; 3. Utilizing rich sources of contextual data coming not only from networking devices, but from connected smart building systems and devices as well; 4. Making different wireless access technologies complement each other (short range for precise location, high capacity for pushing content, low delay for reactive decision making).

Smart building SDN controller presented in this paper (see Fig. 1) addresses all four requirements listed above. The controller is based on Open Networking Foundation (ONF) SDN Architecture 1.0 (currently being migrated to specification 1.1) [11]. As Control Data Plane Interface (CDPI - Southbound Interface) it incorporates the OpenFlow v1.3 for managing forwarding tables of network elements comprising smart building LAN. Controller elements are implemented using the Python 2.7 language and libraries for graph analysis, data analytics and API management. Controller's abstraction layer does not abstract just the LAN infrastructure, but also provides contextual overlays for network graphs representing smart building services and systems realized through the local area network. This layer also creates pool of edge processing and storage resources so that reactive decision making algorithms and accompanying data analysis processes are always efficiently executed. Contextual overlay for the SDN controller abstraction layer is created through soft sensors deployed throughout the smart building LAN and connected devices. Soft sensors [10] are software modules which can be deployed throughout IoT system. For enabling edge contextual and sensory data analysis in IoT systems, soft sensors are deployed on networking elements like access points, IoT gateways, end user mobile devices, BMS controllers and on other connected devices. Also, soft sensors can be deployed on cloud platforms for providing in depth data analysis. Soft sensors are building blocks for complex data analysis processes (implementation details depend on system/device on which the software sensor is deployed). Every soft sensor performs specific data analysis functionality (i.e. put temperature change in temporal and spatial context, detect if user enters/leaves region, derive elevator usage pattern). Results are used by the smart building SDN controller to construct different contextual overlays for LAN abstraction graph (different limitations on edges and different weights for nodes) required by decision making applications when making specific decision on how to manage available networking, processing and storage resources.

Smart building LAN usually consist of wireless access with overlapping coverage regions (for enabling mobility and increasing capacity) and multiple radio access technologies (e.g. WiFi, Bluetooth, Zigbee), distribution switches and firewall routers. Dedicated edge processing resources in form of IoT gateways are strategically placed throughout the building and connected through the LAN into a pool of edge processing resource and further with cloud platform. All these resources and LAN components are represented in the SDN controller abstraction layer.

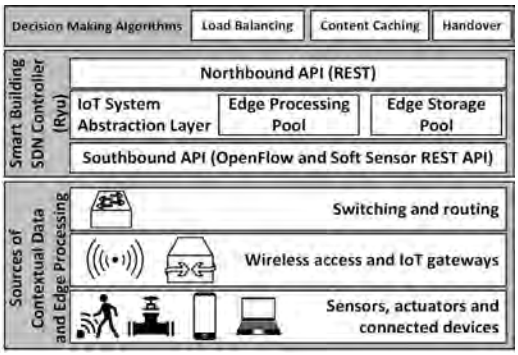


Figure 1. Smart building SDN controller

Smart building SDN controller's southbound interface comprises OpenFlow for managing forwarding tables in distribution switches and on WiFi APs thus enabling traffic engineering dictated by decision making algorithms. Also, the southbound interface includes a collection of REST APIs for communication with soft sensors and gathering their data analysis results which are used for updating abstraction layer graph and its contextual overlays. Controller's northbound interface is collection of APIs (REST) for networking decision making algorithms residing either on the controller or on IoT cloud platform. These decision making algorithms perform tasks related to context aware management of networking resources in line with QoS requirements of smart building and Internet based services. These tasks include (but are not limited to): 1. Load balancing based on available link capacity and congestions; 2. Load balancing based on available edge processing resources; 3. Device, user and service handover between access points based on load distribution and coverage regions; 4. Context aware utilization of edge storage resources for efficient content caching. As IoT platform we are utilizing our VizLore smart building management cloud platform deployed on the Google Cloud platform. The VizLore IoT cloud platform comprises Datastore for contextual and sensory data, data analysis and machine learning algorithms, configuration and reporting interfaces as well as handlers for APIs (REST towards soft sensors and our Python based SDN controller).

Currently we have developed and tested three decision making algorithms based on our smart building SDN controller logic. The first algorithm performs load balancing between IoT GWs strategically placed throughout a building with the main goal being efficient utilization of available edge processing resources (it addresses the Trend/Requirement 2 and 3 presented in the section 1). The second algorithm is context aware handover between wireless access points with overlapping coverage regions with the goal of handling mobility and improving QoS (it addresses the Trend/Requirement 1 and 3 presented in the section 1). The third algorithm is context aware content caching on edge storage resources for improved QoS and balancing load on distribution and core segments of the smart building LAN (it addresses the Trend/Requirement 1, 2 and 3 presented in the section 1).

IV. EVALUATION

A. *Use case 1 - context aware load balancing for efficient edge processing*

Smart buildings, as many other complex IoT systems, require distributed data analysis between cloud platform (for proactive decision making) and edge of the system (for reactive decision making and related data analysis). Reactive decision making (detection of alarms, triggers, status change detection and corrective actions and notifications) is mission critical and needs to be performed in (near)real time and reliably. Therefore, edge processing pool should be efficiently utilized through context aware load balancing.

Load balancing methods like the Round Robin, Weighted Round Robin (WRR) or more sophisticated traffic engineering approaches are well suited for SDN-based load balancing in data centers and for cloud platforms. Smart buildings are very dynamic systems since inhabitants are actively included into smart building processes. As people move throughout a building, load on edge processing points changes (i.e. meeting in one part of a building, arrival of group of visitors, party in an apartment etc.) which may overload certain points while leaving other underutilized. In smart buildings edge processing resources are strategically distributed so that required level of redundancy and scalability is ensured. IoT gateways process sensory and contextual data coming to them, perform reactive decision making based on data analysis and format data batches to be sent to the cloud for further analysis. In this way security and privacy requirements are addressed since GWs filter what and how leaves the smart building IoT system. When balancing load between gateways in a smart building it is important that every processing GW receives required set of sensory data to be able to make proper decision regarding particular smart building service (i.e. one IoT GW needs sensory data streams from all smoke detectors in a particular part of a building to be able to make decision whether or not to trigger alarms). This data set completeness is the main aspect of our SDN based load balancing approach. Bearing that in mind, the actual load balancing algorithm does not have to be very sophisticated (in our experiment we are using WRR algorithm [12]).

Our smart building SDN controller performs load balancing between IoT gateways with two main goals: 1. Even utilization of IoT GWs and their processing resources (typical approach in WRR); 2. While ensuring that load balancing does not compromise IoT GWs ability to make proper decision (ensure that all necessary sensory and contextual data streams are present at the GW performing decision making task - dataset completeness). In our WRR implementation weights assigned to particular IoT GWs correspond to their remaining processing capacity. More importantly, the load balancer recognizes sensory and contextual data flows and ensures that load is balanced not on a packet or single flow basis

but on basis of set of data flows (dataset completeness). The algorithm assigns identifiers to sensory and contextual data streams and groups them into classes based on smart building system that they belong to. The algorithm classifies data based on SDN controller's overall knowledge of the IoT system and LAN (which sensors/devices are source of particular data stream, static IP addresses, MPLS labels, transport protocols and sockets). The SDN controller implements publish/ subscribe context broker where sensors and connected devices advertise their availability and decision making processes/GWs subscribe to data sources necessary for completing data sets on which to perform data analysis.

For the purpose of validation we have built proof of concept testbed in the Mininet emulator [13]. WiFi APs and distribution switches are implemented as Mininet switches. IoT processing GWs, sensor nodes and mobile devices are emulated as Mininet hosts. SDN controller is Ryu (we use it for realization of other experiments as well), deployed on a Linux machine and connected with server running Mininet emulation through physical Ethernet interface bridged with the Open VSwitch instance. Python code for load balancing is deployed on the same machine as the SDN controller. The load balancing algorithm collects data about load on links connecting IoT GW hosts. Hosts acting as mobile devices have two different data flows from two client applications (UDP and TCP - this allows for easy stream classification). Mininet hosts acting as sensor nodes have only UDP client application. IoT GWs are Mininet hosts with server applications with which client applications establish connection and send dummy data streams of defined bitrate. Fig. 2 shows the testbed topology.

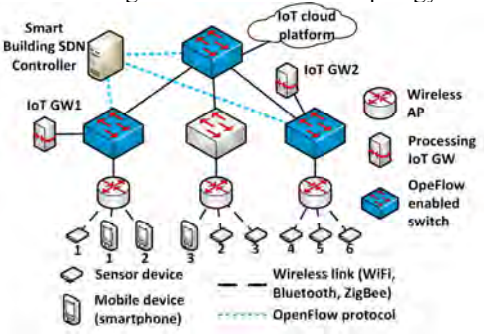


Figure 2. Use case 1 testbed configuration

Sensors 1, 2, 4 and 5 belong to smart building service class 1. Sensor 3 and UDP applications from mobile devices 1 and 2 belong to smart building service class 2. TCP client applications from mobile devices belong to smart building service class 3. Finally, UDP application from mobile device 3 and sensor 6 belong to smart building service class 4. This is pre programmed on the load balancer (in real life deployment this classification of data streams and data sources would be performed by the specific decision making algorithms deployed on the SDN controller as described in the 3rd paragraph in this section). Load from UDP client applications on mobile devices 1, 2 and 3 is increased in 0.05Mbit/sec increments from 0.05 Mbit/sec to 0.3Mbit/sec (these 6 steps represent 6 experiment steps analyzed in the next paragraph). Load from sensory devices 1-6 and TCP applications is kept at 0.05Mbit/sec. Each IoT GW can perform reactive decision making corresponding to all smart building service classes. At first, IoT GW1 performs reactive decision making for smart building service classes 1 and 3 while GW2 serves smart building service classes 2 and 4.

As benchmarks, we have approach without load balancing and Round Robin based load balancing. In Fig. 3 we can see that with increasing load from UDP host applications, IoT GWs are unevenly utilized when there is no load balancing (GW2 is nearly 3 times more loaded than GW1 at the end of experiment). On the other hand, standard load balancing approach (Round Robin) shows that IoT GWs are evenly loaded all the time, but smart building service class of data flows (IoT service dataset completeness) is not considered during load balancing thus making IoT GWs unable to conduct reactive decision making. Our algorithm achieves even IoT GW utilization while ensuring contextual dataset completeness for each IoT GW. In the first two steps of the experiment our algorithm does not change initial assignment of service classes to GWs. In step 3, in order to balance the load on GWs, the algorithm combines service classes 1 and 4 and sends them to the GW1 while classes 2 and 3 are sent to the GW2. At step 5 data flows belonging to service classes 1, 3 and 4 are sent to the GW1 while GW2 maintains only data flows from the service class 2. The proposed algorithm ensures dataset completeness for each IoT GW.

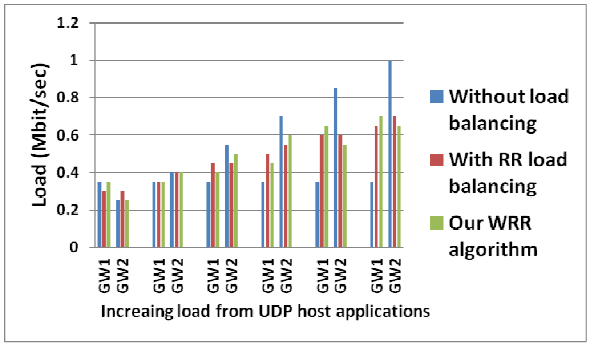


Figure 3. Load balancing for 6 steps of increasing UDP client app load

B. Use case 2 - context aware device/user/service handover in smart building WLAN

This use case is realized on a real smart building testbed (see Fig. 4) comprising 3 wireless access points with WiFi and Low Energy Bluetooth (BLE) interfaces deployed on two building floors, dual interface connected devices (WiFi + BLE interface) and single interface connected devices (only BLE). Dual interface devices are smartphones with custom built Android applications while single interface devices are Raspberry Pi boards emulating BLE sensors. The smart building SDN controller is our implementation of the Ryu controller and it is deployed on MikroTik RB800 running OpenWRT. Wireless access points also run OpenWRT with Open VSwitch instances for enabling OpenFlow based management of forwarding tables on virtual interfaces which are bridged with physical interfaces of APs. Southbound interface is OpenFlow towards wireless access points and AllSeen Alliance's AllJoyn framework [14] API towards connected devices. Single and dual interface devices have AllJoyn client applications which enable them to advertise their capabilities and discover other AllJoyn applications in the

local network. Wireless access points run AllJoyn router application which enables them to discover devices in their coverage regions and perform handover when instructed by the SDN controller decision making algorithm. Decision making algorithm is implemented in Python and performs context informed handover decisions based on contextual insights provided by soft sensors deployed on connected devices and on wireless access points. These soft sensors constantly feed SDN controller and decision making algorithms with updates regarding movement from one coverage region to another (based on BLE beacons identified by custom mobile applications and based on WiFi AP's capability to detect presence of mobile devices with BLE interface in advertising mode), load on interfaces, interference levels (SNR measurements on AP interfaces) and QoS requirement changes. The SDN controller provides the algorithm with abstraction layer in form of network graph which have edge for every possible handover option based on coverage region of wireless AP interfaces (known in advance based on spectrum measurements). These edges are assigned with forwarding priority based on current distribution of load on wireless access points, network problems (node outage, interference) and load on IoT GWs. Based on these priorities, the handover algorithm instructs wireless access points to modify their forwarding tables to accommodate the need for handover and instruct connected devices (through custom mobile application) to reconnect to another AP or another interface of the same AP.

Experiments are conducted in two stages. First, traffic generators are used to introduce increased load on wireless access points which triggered handover of devices that are in coverage region of two APs. Mainly dual interface devices are handed over from BLE interface of one wireless AP to WiFi interface of another AP in order to offload particular AP. In this experiment stage we show that through device/user/service handover between wireless access points with overlapping coverage regions we can achieve better control over load balancing and increased resilience to networking node outages and interference. The second experiment stage involved SIP/VoIP session handover for mobile users as they move through coverage regions of wireless access points. In typical smart building LAN deployment as mobile user exits coverage region of one access point and enters into coverage region of another AP, the session is interrupted for duration of time necessary for mobile device to reconnect to the new AP and reestablish SIP session (this typically takes 3-10 seconds). Our opportunistic handover approach utilizes both interfaces of wireless AP. SIP session is established and maintained through WiFi interface of APs. BLE interface of a mobile device is used to scan for beacons emitted by BLE interfaces of surrounding APs. Mobile application utilizes deployed soft sensor to continuously analyze handover opportunities and report them to the SDN controller who knows topology of the whole network (in graph like form with edges representing if handover between two APs is possible based on overlapping coverage regions), user's mobility pattern (if user leaves coverage region of one AP and enters into coverage region of another, the controller knows what is the next AP into which coverage region user can enter which is preprogrammed based on knowledge of network topology and building's floor layout - i.e. long corridors, exit points) and handover opportunities and capacities (based on the current measurements on AP interface queues). This knowledge is fed to the Python decision making script which prepares forwarding table on the AP (through the OpenFlow interface) towards which the mobile user is moving so that the L2 and L3 handover is ready as soon as device initiates physical handover (this speeds up the handover process to under 1 second). Android devices allow mobile applications to initiate reconnection to another WiFi AP when in range which enables decision making algorithm to control when device initiates physical handover. The proposed SDN controller and decision making algorithm are able to properly exploit opportunities for context aware handover.

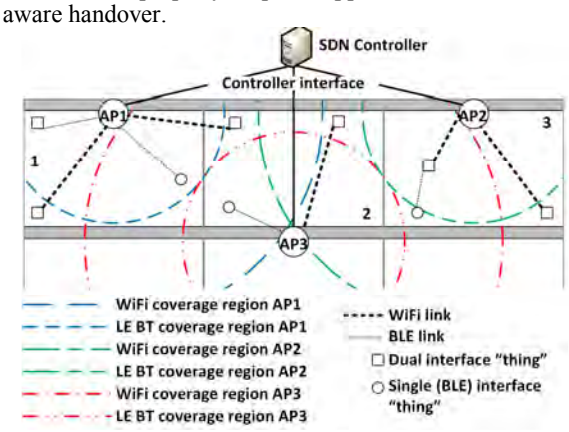


Figure 4. Use case 2 testbed configuration

C. *Use case 3 - context aware content caching in smart building edge*

Content delivery (especially video on demand and video streaming) are the main consumers of Internet and LAN capacity in buildings. If provided with sufficient distributed edge storage space, SDN controlled smart building can cache popular video content proactively which can result in drastic load reduction on leased Internet links and throughout the building LAN. This ensures that LAN resources are efficiently utilized between all smart building and Internet based services even during peak hours for Internet related activity of building inhabitants. In our previous work [15], [16] we have developed, implemented and tested a MILP (mixed integer linear program) model for deriving optimal content placement on routers in a wireless mesh network. Smart building wireless LAN with overlapping coverage regions can be seen as mesh network since users can be handed over between APs (as seen in the use case 2). The MILP model implemented in Python is used as a proactive decision making algorithm for our SDN controller.

The OpenFlow switches from the Mininet emulate smart building WLAN (see Fig. 5). Links between switches are defined so as to represent handover opportunities between APs. The MILP model utilizes soft sensors to obtain spatial and temporal distribution of content requests and profiles of connected devices. Based on history of content request distribution, available storage and end user device profiles, the model derives optimal content placement schedule (throughout the edge storage pool) in line with the optimization goal: maximizing total streaming capacity (serving maximal number of clients

before one of the links in the network becomes congested). We envision peer to peer based content delivery mechanism where content is always delivered from the local storage pool (if it is locally stored). The OpenFlow is used for modifying forwarding tables of the Mininet switches. If a Mininet switch (emulating AP) is selected for content placement, forwarding tables are modified so that requests for the content server are redirected to the selected switch. Results obtained through this emulation correspond to the main findings in our previous work [16] where the MILP model was implemented using MatLab. The results show that the number of users served with requested content from the edge storage pool enters saturation after content is stored on at least 65% of APs so there is no need to store content replica on every AP to completely offload content server. This indicates that, for a content which is very popular among smart building inhabitants, caching the content in the edge storage can significantly reduce burden on content servers and also reduce load on shared Internet backhaul links thus increasing overall content delivery capacity of the smart building LAN.

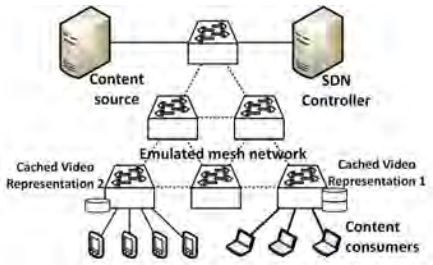


Figure 5. Emulated testbed for the use case 3

Furthermore, based on data obtained from soft sensors (mainly video reproduction capabilities of end user devices), the SDN controller is able to select which SVC (scalable video coding) representation of video content needs to be stored on which edge storage pool component in order to maximize QoE and storage resource utilization efficiency (see Fig. 5). The context aware handover algorithm presented in the use case 2 enables the SDN controller to adapt distribution of connected devices between WiFi APs with overlapping coverage regions. This provides another layer of opportunity for improving edge storage pool resource utilization (place high quality video representation on an AP where majority of connected devices have high resolution displays).

V. CONCLUSION

In this paper we have presented SDN controller as a key enabler for context aware resource and service provisioning management throughout smart buildings. We have elaborated trends and resulting requirements for proper management of smart building systems in IoT era. We have presented three decision making algorithms which utilize the proposed SDN controller logic for context aware load balancing for efficient edge processing, handover in wireless access network and content caching in smart building edge. The three proof of concept use cases provide insight into potential of the described smart building resource and service provisioning management solution based on SDN controller. In our future work we will address each of the presented decision making algorithms in more detail and with large scale experiments and real life deployments. We will focus on joint implementation of all proposed decision making algorithms and the SDN controller in real life smart building testbed that we operate (150 apartments on 15 floors) in order to further calibrate mathematical and data analysis models to be able to make proper decisions in real network traffic scenarios.

REFERENCES

[1] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog Computing and its role in the internet of things," Proceedings of the first edition of the MCC workshop on Mobile cloud computing, pp. 13-16, ACM, 2012.

[2] L. Vaquero, and L. Rodero-Merino, "Finding your way in the fog: towards a comprehensive definition of fog computing," ACM SIGCOMM Computer Communication Review, vol. 44, pp. 27-32, 2014.

[3] University of California at Berkeley - Software Defined Buildings. [Online]. Available: http://sdb.cs.berkeley.edu/sdb

[4] S. Dawson-Haggerty, A. Krioukov, J. Taneja, S. Karandikar, G. Fierro, N. Kitaev, and D. Culler, "BOSS: building operating system services," Proceedings of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI '13), April 2013.

[5] A. Krioukov, G. Fierro, N. Kitaev, and D. Culler, "Building application stack (BAS)," 4th ACM Workshop On Embedded Sensing Systems For Energy-Efficiency In Buildings, Toronto, Canada, Nov 2012.

[6] B. A. Nunes, M. Mendonca, X.N. Nguyen, and K. Obraczka, "A survey of software-defined networking: past, present, and future of programmable networks," IEEE Communications Surveys and Tutorials, Vol. 16, pp. 1617-1634, IEEE, 2014.

[7] T. Luo, H.P. Tan, and T.Q.S. Quek, "Sensor OpenFlow: enabling software-defined wireless sensor networks," IEEE Communication Letters, Vol. 16, pp. 1896-1899, IEEE, 2012.

[8] P. Kathiravelu, L. Sharifi, and L. Veiga, "Cassowary: middleware platform for context-aware smart buildings with software-defined sensor networks," Proceedings of the 2nd Workshop on Middleware for Context-Aware Applications in the IoT, pp. 1-6, ACM, 2015.

[9] Z. Qin, G. Denker, C. Giannelli, P. Bellavista, and N. Venkatasubramanian, "A software defined networking architecture for the internet-of-things," IEEE Network Operations and Management Symposium, pp. 1-9, 2014.

[10] M. Tosic, O. Ikovic, D. Boskovic, "Soft sensors in wireless networking as enablers for SDN based management of content delivery," 39th International Convention, MIPRO 2016, accepted for publication.

[11] Open Networking Foundation SDN Architecture 1.1 specification. ONF TR-521 February 2016.

[12] Weighted Round Robin (WRR [Online]. Available: https://en.wikipedia.org/wiki/Weighted_round_robin).

[13] Mininet (Mininet: An Instant Virtual Network on your Laptop. [Online]. Available: http://mininet.org/).

[14] AllSeen Alliance (AllJoyn Framework. [Online]. Available: https://allseenalliance.org/framework).

[15] D. Boscovic, F. Vakil, S. Dautovic, and M. Tosic, "Pervasive wireless CDN for greening video streaming to mobile devices," Proceedings of the 34th International Convention MiPRO, IEEE Xplore, pp. 629-636, 2011.

[16] M. Tošić, M. Ćirilović, O. Iković, D. Kesler, S. Dautović, and D. Boscovic, "Impact of different content placement and delivery strategies on streaming capacity of wireless mesh networks," Proceedings of the 11th International Conference, ADHOC-NOW, Springer, pp. 302-315. 2012 Belgrade, Serbia.

TELECOMMUNICATIONS & INFORMATION

INVITED PAPER

Time Series Analysis and Possible Applications

Mirjana Ivanović*, Vladimir Kurbalija*
* University of Novi Sad, Faculty of Sciences, Novi Sad, Serbia
e-mail: {mira, kurba}@dmi.uns.ac.rs

Abstract - Data presented in form of time series as its analysis and applications recently have become increasingly important in different areas and domains. In this paper brief overview of some recently important standard problems, activities and models necessary for time series analysis and applications are presented. Paper also discusses some specific practical applications.

I. INTRODUCTION

Recently data presented in form of time series as its analysis and applications have become increasingly important in different domains. A time series is a set of ordered observations on a quantitative characteristic of a phenomenon at equally spaced time points. One of the main goals of time series analysis is to forecast future values based on existing series. There are a lot of possible domains and various fields of research based on time series like: business, engineering, environometrics, economics, medicine, politics, social sciences, and a lot of others. Initial and significant book in the area *Time Series Analysis: Forecasting and Control*, has been published by Box and Jenkins (1970, 1976) bringing comprehensible methodologies and providing a modernized approach to time series analysis and forecasting. After that a numerous books and enormous number of research, theoretical and practical, papers have been published changing attitude and increasing significance of such kind of data.

One of the main activities in this domain is time series modeling which supports to carefully collect and rigorously study the past observations of a time series in order to develop an appropriate model. Such model serves to describe the inherent structure of the series and to produce adequate framework that then has been used to generate future values for the series, and make useful forecasts.

The use of computers and appropriate software tools is necessary in modern quantitative analysis. It is also extremely important in time series analysis where different complex algorithms and extensive computations are unavoidable. In many research areas and situations adoption of different methodologies and usage of adequate software tools are necessary to provide simplified and effective processes for complex subjects in time series analysis and forecasting.

The rest of the paper is organized as follows. In second section a brief overview of contemporary techniques for time series modeling and processing are presented. Section III is devoted to some of standard activities for better representation, pre-processing and similarity analysis of time series. Section IV briefly discusses some possible practical applications of time series in several domains. Section V introduces a framework for analysis and prediction in time series processing. Last section concludes the paper.

II. SOME IMPORTANT TIME SERIES MODELS

Time series analysis, modeling and forecasting have fundamental importance to various practical domains. To improve the efficiency and accuracy of time series modeling and forecasting in last two decades a lot of innovative, important models have been proposed. Some authors recognize [1] three important classes of time series models: the stochastic, neural networks and SVM based models.

Among the frequently used and evidently most popular stochastic time series models is the *ARIMA - Autoregressive Integrated Moving Average* [2, 3] model. The essential assumption in implementation of this model is: time series are linear and follows a particular known statistical distribution, like the normal distribution. ARIMA model is further explored for establishing other subclasses of models, like the *Autoregressive (AR)*, *Moving Average (MA)* and *Autoregressive Moving Average (ARMA)*. Box and Jenkins [4] had proposed very successful variation of ARIMA model for known phenomena of seasonal time series forecasting, the *SARIMA - Seasonal ARIMA*.

In spite the fact that ARIMA model is flexible, simple for optimal model building process it suffers of severe limitation. One of important limitations is pre-assumed linear form of the associated time series which becomes inadequate in different practical situations. Consequently various non-linear stochastic models have been developed in order to overcome this drawback. On the other hand these models are not so straight-forward and simple from implementation point of view as the ARIMA models.

Recently, general revitalization of artificial neural networks (ANNs) also has influenced its growing attentions in the fields of time series forecasting [2, 5]. ANNs especially have been successfully applied for forecasting and classification purposes in wade range of areas [5]. The excellent feature that influenced use of ANNs in forecasting problems is their inherent capability of non-linear modeling, without any presumption about the statistical distribution followed by the adequate observations [5].

ANNs possess two important features, first of all they are data-driven (as appropriate model is adaptively formed based on the given data) and on the other hand they are

self-adaptive [6]. During the past several years a considerable amount of research have been carried out towards the application of neural networks for time series modeling and forecasting.

There are various ANN forecasting methods and algorithms like the multi-layer perceptrons (MLPs), which are characterized by a single hidden layer *Feed Forward Network (FNN)* [6] but also widely used variation of FNN like *Time Lagged Neural Network (TLNN)*. A very interesting and new approach is oriented towards ANN model for seasonal time series forecasting *Seasonal Artificial Neural Network (SANN)*. It had been presented in more details in [7].

This model is unexpectedly simple and has been experimentally verified to be quite successful and efficient in forecasting seasonal time series.

A significant step forward happened in the area of time series forecasting with the development of concept of support vector machine (SVM) [8, 9]. The initial aim of SVM was to solve pattern classification problems but later they have been started to be applied in many other fields including time series prediction problems [9]. Apart from fact that it is useful for classification tasks it is also intended for a better generalization of the training data. For this reason the SVM methodology getting one of the frequently used techniques for time series forecasting problems. The objective of SVM is to use the *structural risk minimization (SRM)* principle to find a decision rule with good generalization capacity. Important feature of SVM is that here the training is equivalent to solving a linearly constrained quadratic optimization problem. Unlike the other traditional stochastic or neural network methods solution achieved by applying SVM method is always unique and globally optimal. Above all the most significant property of SVM is that the quality and complexity of the solution can be independently controlled, irrespective of the dimension of the input space [10]. Usually in SVM applications, the input points are mapped to a high dimensional feature space using special functions - *support vector kernels*. In last several years numerous SVM forecasting models have been developed for time series forecasting among the most frequently used are: *Least-square SVM (LS-SVM)* and *Dynamic Least-square SVM (LSSVM)*.

III. Analysis of Time Series

Time series embody one form of sequential data where the value of some numeric parameter in an observed process is recorded periodically in time. Each element of a time series describes the phenomenon under examination at a specific point in time. The parameters of interest can be various: the recorded values of some scientific sensor, the daily number of hits of some website, the value of shares on a stock market, the number of rainy days per year, etc. The observation rate can vary from one millisecond to several years, and in the general case the observation rate need not to even be uniform. However, the common characteristic of time series is that they usually consist of a large number of points (recorded values), which can make the handling of time series difficult.

Time series are used for storage, display and analysis of data across a wide range of different domains, including various areas of science, medicine, economics, ecology, telecommunications and meteorology [11–13]. In almost every scientific field, measurements are performed over time and the collected data can be organized in the form of time series with the aim of extracting meaningful knowledge.

Time-series analysis applies different methods, mainly from statistics, data mining and machine learning, in order to model and understand the process which generated time series or to make forecasts. Accordingly, several task types are important for time-series analysis, most notably indexing, classification, clustering, prediction, forecasting, segmentation, and anomaly detection [14–17]. In addition, there are three important concepts which need to be considered when dealing with time-series data, and which will be described in more detail: pre-processing transformation, time-series representation and similarity/ distance measure.

A. Pre-processing Transformations

Very commonly "raw" time series comes with different kinds of distortions. For most real-world applications these distortions are not meaningful, and they should be removed. This is the task of pre-processing phase. Generally, there are four groups of pre-processing transformations: **Offset translation, Amplitude scaling, Removing linear trend, Removing noise.**

Offset translation: Frequently, the values of various time series does not belong to the same interval, although their overall shape is very similar. In those situations, the values of time series should be translated into the same value-interval. The simplest solution for this problem is to subtract the mean value of each time series.

Amplitude scaling: In some situations, two or more time series have different amplitudes. If this property is undesirable, the amplitude scaling transformation should be applied. The most common approach is to subtract the mean value from time series and to divide this difference by standard deviation.

Removing linear trend: In some cases, time series contain some constant linear trend (growing or decreasing). In these situations, the most common approach is to fit a straight line through the time series, and then to subtract that line from the time series.

Removing noise: Very often, raw time series contains some kind of noise. There are many techniques for noise removal, but most of them are based on moving averages methods. The essence of these methods is to average each data point of time series with two or more of its neighbors.

Finally, it is worth mentioning that sometimes the distortions are the most interesting thing about the data, and in these cases pre-processing is not performed. The selection of pre-processing tasks depends on data and on task types which have to be performed.

B. Time-series Representations

As already mentioned, the dimensionality of time series, depending on domain, can be huge. Therefore, dimensionality of time series must often be reduced before applying certain data mining tasks upon it. Main purpose of time series representations is to decrease data dimensionality while keeping the important characteristics of the original time series which usually implies the overall behavior of time series and the shape. This process, of course, includes some information loss, so the main objective is to keep crucial information. By keeping important information, representation preserves time series features, behaving just like original time series in data mining algorithms. Among many proposed representation techniques, only the major ones will be described here.

Discrete Fourier Transform (DFT). DFT is actually used in many fields (such as digital signal processing, image processing, solving of partial differential equations, etc.), and has also found its application in time series data mining. Basic idea of DFT is that any signal can be represented by the super position of finite number of sine/cosine waves. After applying DFT, dimensionality of time series stays unchanged, so to reduce data dimensionality, we need to dismiss some Fourier coefficients. It is observed that only the first few coefficients appear to be dominant and therefore the rest can be omitted without great information loss [18]. In that way data series dimensionality can be efficiently decreased.

Discrete Wavelet Transform (DWT) [19]. Wavelets are functions that represent time series in terms of the sum and difference of prototype function, called mother wavelet. Unlike DFT, in which case each coefficient carries out only global information, DWT coefficients hold local information (each wavelet holds information of some time series segment). First coefficients are more global, while the each next group of coefficients refines global picture by adding more information to certain segments of time series. That is why they are very suitable for decreasing time series dimensionality. By taking first few DWT coefficients, we are actually taking global time series characteristics, and that is exactly what time series representation should hold.

Piecewise Aggregate Approximation (PAA) [20]. Underlying idea is very simple. Original time series is divided in N equal segments. For each segment mean value is calculated and stored. In that way dimensionality of original time series is decreased to N. Although very simple, this method proved to be reasonably good when used in data mining algorithms.

Piecewise Linear Approximation (PLA) [21]. Just like in PAA, time series is divided in N segments (here not necessarily of equal size). Each segment is then represented by a line. There are two ways of defining that line: it can be linear interpolation between start and end point of corresponding segment, or it can be linear regression that takes into account all points that are contained in the segment. Each segment is then presented by two numbers (*a* and *b* coefficients of line equation $y = ax + b$), so if dimensionality of final representation needs to be N, time series is divided into N/2 segments. As noted at the beginning, time series segments by definition do not have to be of equal size. Segments lengths are determined by different kind of algorithms whose main aim is to create representation of minimal reconstruction error. The problem with that approach is that distance measure upon this kind of representation does not support lower bounding, which is necessary in order to use representation in indexing algorithms.

Indexable Piecewise Linear Approximation (IPLA) [22]. IPLA solves the indexing problem of PLA in two aspects. According to first modification, IPLA does not allow segments of different lengths. Second modification introduces reset of x (time) component of each time series segment.

Symbolic Aggregate Approximation (SAX) [23]. This representation is built upon PAA. Idea of SAX is to convert data into a discrete format which is based on predefined alphabet. Before discretization, time series is transformed via PAA algorithm. In order to convert PAA coefficients to alphabet symbols, we must first define breakpoints that divide the distribution space into S equiprobable regions, where S is the size of alphabet. Each defined region is mapped into one alphabet symbol. Transformation of PAA representation is now trivial each coefficient is mapped and transformed into one alphabet symbol.

C. Similarity/Distance Measures

One of the most important aspects of time-series analysis is the choice of appropriate similarity/distance measure – the measure which tells to what extent two time series are similar. The choice of an appropriate time-series similarity measure is a critical point when dealing with many tasks in mining temporal data. While working with traditional databases we are interested in data that exactly match the given query, in the case of similarity-based retrieval of time series, we are looking for sequences that most resemble a given series. As similarity-based retrieval is explicitly or implicitly used in all main tasks of time series analysis (including classification, clustering, forecasting, etc.), it is important to carefully define the similarity measure between time series in order to reflect the underlying (dis)similarity of the specific data they represent. There is a large number of (dis)similarity measures for time series data proposed in the literature, but here we will present only the several major ones.

L_p-norm is one of the most popular classes of similarity measures that are used in data mining algorithms. Its definition is given in formula (1) where Q and C are time series of length n, q_i is the value of Q at time point i, and C_i is value of C at time point i. Variable p denotes the norm that is used. For example, p = 1 corresponds to the Manhattan distance, p = 2 is Euclidean distance, and $p = \infty$ is Chebyshev distance.

$$D(Q,C) = \sqrt[p]{\sum_{i=1}^{n}(q_i - c_i)^p} \qquad (1)$$

The advantage of Euclidean distance is that it is very easy to compute and to understand. There are, however, some disadvantages, too: the sequences must have the

same number of points (can be avoided by interpolation to equal length), it is sensitive to shifting and scaling along the y-axis (can be precluded by normalizing the series), and it is also sensitive to distortions and shifting along the time axis. The last problem is successfully solved with elastic measures; all measures that are presented in further text are elastic distance measures.

Dynamic Time Warping (DTW). Euclidean distance is based on linear aligning of related points of time series (Fig. 1 (a)): the i-th point of the first series is paired with the i-th point of the second one. The assessment of the distance can be improved by warping the time axis of one or both of the sequences (Fig. 1 (b)). One of the most popular distance measures based on non-linear aligning is the Dynamic Time Warping [24].

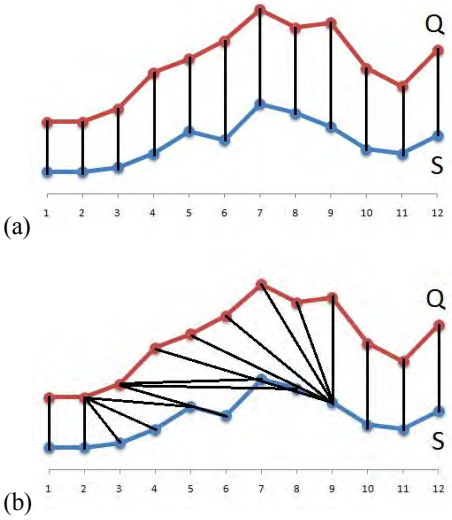


Figure 1. (a) Linear aligning of time series in case of the Euclidean distance, (b) Non-linear aligning of DTW

Longest Common Subsequence (LCS) [25]. This elastic similarity measure is one of the edit distances. The similarity of two time series is presented as the length of their longest common subsequence.

Edit Distance on Real Sequence (EDR) [26] is another edit distance that found its application in time series data mining. EDR value is actually minimal number of edit operations (insertions, substitutions and deletions) that need to be applied upon one time series in order to transform it to another time series. This similarity measure proved to be very robust to noise.

Edit distance with Real Penalty (ERP) [27]. Main problem with all mentioned elastic measures is that they do not satisfy the triangle inequality, so they cannot be used in indexing algorithms. ERP similarity measure combines L1-norm with DTW and EDR, dismissing segments of those measures that are causing them not to satisfy triangle inequality.

All of these elastic similarity measures are based on dynamic programming. It is well known that the computational complexity of dynamic programming algorithms is quadratic, which is often not suitable for larger real-world problems. However, the usage of global constraints such as Sakoe-Chiba band and Itakura parallelogram can significantly speed up the calculation of similarities. Furthermore, it is also reported [28] that the usage of global constraints leads to a qualitatively different measures which can have even improved accuracy of classification compared to unconstrained similarity measures. The accuracy of classification is commonly used as a qualitative assessment of a similarity measure.

IV. Some Possible Real Life Applications of Time Series Data

Significance and importance of time series analysis and practical use is getting more and more important in almost all areas of human life like science, engineering, technical, economical and sociological systems. So time series modeling is extremely important and this model is then used to generate future values for the series, i.e. to make appropriate forecasts. Time series forecasting thus can be termed as the act of predicting the future by understanding the behaviors detected in the past. Due to the indispensable importance of time series forecasting they are significant scientific trend in numerous practical fields such as business, economics, finance, science, medicine and engineering.

The choice of the best and most appropriate forecasting models depends on the forecasting horizon and usually there are three types of time series forecasts.

Point Forecast: a single number or a "best guess." It does not provide information on the level of uncertainty around the point estimate/forecast.

Interval Forecast: relative to a point forecast, this is a range of forecasted values which is expected to include the actual observed value with some probability.

Density Forecast: this type of forecast provides information on the overall probability distribution of the future values of the time series of interest.

However, the point forecast is the most commonly used type of forecast by businesses managers and policymakers.

A. Time Series in Finanial and Business Domains

Policymakers and business managers on a regular basis use forecasts of financial variables to help make important decisions about production, purchases, market conditions, and other choices about the best allocation of resources [29].

Time series analysis is an integral part of financial analysis with applications to the prediction of interest rates, foreign currency risk, stock market volatility, and a lot of other similar tasks and activities.

In comparison to other areas and domains, the modeling of economic time series is characterized by several problems [29]:

- The empirical sample sizes in economics (usually ranges between 25 - 100 observations) are generally small, especially compared to other

areas like physics or biology with more than several thousands of observations.

- Economic time series are highly dependent and they are correlated with other economic time series. In the economic science, problems are almost never concerned with univariate series. An example can be a function of income, and at the same time, consumption also affects income directly and through various other variables.
- Economic time series are also often dependent over time. Many series display high autocorrelation, as well as cross autocorrelation with other variables over time.
- Economic time series are generally non-stationary. Their means and variances change over time, implying that estimated parameters might follow unknown distributions as an alternative of standard tabulated distributions like the normal distribution. Non-stationarity arises from productivity growth and price inflation. In some situations, for example, inference in econometrics become quite complicated, and requires the development of novel statistical techniques for handling stochastic trends. Examples of such new techniques include the concepts of cointegration and common trends, and asymptotic theory for integrated variables.
- Economic time series cannot be assumed to be drawn from samples like in classical statistics. Since the sampling process can be controlled the variables which make up the sample can be seen as random variables. Hypothesis are then formulated and tested conditionally on the assumption that the random variables have a specific distribution. Economic time series are seldom random variables drawn from some underlying population in the classical statistical sense.
- Different variables like, money, prices, GDP, and dividends are given from history. To get a different sample it is necessary to re-run history, which is obviously impossible. The way statistic theory deals with this situation is to reverse the approach taken in classical statistic analysis, and build a model that describes the behavior of the observed data.
- Finally, from the view of economics, the subject of statistics deals mainly only based on the estimation and inference of covariances. On the other hand the econometrician must also give estimated parameters an economic interpretation. When it comes to time series, economic theory is an integrated part of the modeling process. Many econometric studies fail because researchers assume that their estimates can be given an economic interpretation without considering the statistical properties of the model, or the simple fact there is in general not a one to one correspondence with observed variables and the concepts defined in economic theory.

B. Application of Time Series in Medical Domains

Medical time series usually are combined with other methods like CBR-Case-based reasoning and data mining. These synergies may be of the form of pre-processing for feature mining from time series [30] or for retrieval of cases involving temporal features. Recent trends in applying CBR in medicine involve the representation of time series data for the purpose of reason about patient progress over time [31]. In medical domains important is to determine the transition of behavior over time rather than draw conclusions based on the absolute values in time series.

A characteristic example of combination of case-based decision-support system and time series data has been applied for diagnosing respiratory sinus arrhythmia (RSA) based on sensor readings [32]. The data mining facilities of this system allow for time-series classification and pattern identification. Following the case-based reasoning paradigm, classification of time series is performed by observing similar time series. In particular, the classifier examines sequences of patterns and their co-occurrence with class labels. The method is experimentally evaluated with a clinical case library, showing that it improves the classification accuracy.

Similar approach [30] the challenge of analyzing time series generated from lengthy and complex nature of signals is examined. In this approach a set of qualitative, interpretable features from original and usually real-valued time series data has been extracted. Features are generated by transforming the time series of real numbers into a symbolic series by temporal abstraction or symbolic approximation. Classification is then performed on such symbolic sequences and authors examine four alternative ways to index time series cases using discovered key sequences.

In [33] authors extend mentioned approaches by developing a hybrid case-based reasoning system for the application of stress diagnosis, which is based on time-series measurements in the form of signals and textual medical meta-data. To address the noise in the signals system uses a fuzzy similarity searching procedure. Fuzzy similarity aims at more accurate similarity estimates compared to traditional distance-based similarity measures for time series like Discrete Time Warping. More details on approaches in this direction can be found in [33].

C. Large-Scale Astronomical Time Series Databases

One of contemporary and modern areas where time series analysis can play important role are different areas of astronomy and astrophysics [34]. Time-domain astronomy (TDA) is rapidly changing faced by the exponential growth of the sample size, data complexity and data generation rates of new astronomical sky surveys. For example, the Large Synoptic Survey Telescope (LSST), which will begin operations in northern Chile in 2022, is planned to generate more than 150 Petabyte imaging dataset of the southern hemisphere sky. The LSST is intended to improve understanding of time-varying astrophysical objects, and also to reveal a plethora of yet unknown faint and fast-varying phenomena [34]. So this new field of data-driven astronomy

introduced new challenging areas: astroinformatics and astrostatistics. These new data-oriented paradigms encompass significant disciplines like statistics, data mining and machine learning, knowledge discovery, and computational intelligence. The role of time series analysis here is of great significance for using robust methods needed for the rapid detection and classification of known astrophysical objects as well as the unsupervised characterization of novel phenomena.

Photometry as a part of astronomy is aimed to the precise measurement of visible electromagnetic radiation from astronomical objects. Several techniques and methods are applied to transform the raw astronomical data into standard units of flux or intensity. The basic tool in the analysis of astronomical brightness variations is the light curve. A light curve is a plot of the magnitude of an object's electromagnetic radiation expressed as a function of time. This light curve analysis is challenging, because of the sheer size of the databases, and also due to the characteristics of the data itself. Astronomical time series are unevenly sampled due to constraints in the observation schedules, telescope allocations and other limitations. The sampling is randomized because observations for each object happen at different times every night under different weather conditions. Discontinuities in light curves can also be caused by different technical factors. Astronomical time series are also affected by several noise sources.

In general, errors in astronomical time series are non-Gaussian and heterocesdastic, i.e., the variance of the error is not constant, and changes along the magnitude axis. Other common problematic situations arising in TDA are the sample-selection bias and the lack of balance between classes. Generally the astrophysical phenomena of interest represents a small fraction of the observable sky, hence the vast majority of the data belongs to the "background class". This is especially noticeable when the objective is to find unknown phenomena, a task known as novelty detection. Sufficient coverage and exhaustive labeling are required in order to have a good representation of the sample, and to assure capturing the rare objects of interests and highly promising application of time series analysis.

V. FAP – Framework for Analysis and Prediction

All these concepts, when introduced, are usually separately implemented and presented in different publications. Every newly-introduced representation method or distance measure has claimed a particular superiority. However, this was usually based on comparison with only a few other representatives of the proposed concept. On the other hand, to the best of our knowledge there is no freely available software system for time-series analysis and mining which supports all important concepts, with the exception of the work proposed in [17]. Being motivated by these observations, we have designed a multipurpose, multifunctional system FAP – Framework for Analysis and Prediction [35]. FAP supports all important concepts and activities connecter to time series data like: different representations, applications of wide range of similarity measures and pre-processing tasks; with the possibility to easily change some existing or to add new concrete implementation of any other concept.

FAP is mainly designed for time-series researchers in order to test and compare existing and newly introduced time-series concepts, methods and models. However, it can also be useful to professionals from different domains not familiar with time-series analysis as an assistance tool for choosing appropriate methods for their own time series data sets. FAP is freely available and can be downloaded from http://perun.dmi.rs/fap/.

The Framework for Analysis and Prediction has been already successfully employed within various research domains including: investigation of the influence of global constraints on distance measures [28, 36], developing a distributed distance matrix generator based on agents [37], mining time series in the psychological domain [38, 39], time-series analysis in the medical domain [40], and in financial forecasting [41].

VI. Conlusion

Time series are characteristic representatives of complex data. Analysis and applications of time series have become progressively more important in a variety of fields of research and real life problems, including business, economics, engineering, medicine, social sciences, environometrics, politics, and others. Also time series analysis, modeling and forecasting has fundamental importance in these practical domains. In the paper we briefly presented several contemporary, innovative, and important models and discussed their possible applications.

References

[1] Adhikari R, Agrawal R. K. (2013) An Introductory Study on Time Series Modeling and Forecasting, arXiv:1302.6613v1.

[2] Zhang G.P. (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50: 159–175

[3] Cochrane J.H. (1997) Time Series for Macroeconomics and Finance. Graduate School of Business, University of Chicago.

[4] Box G.E.P., Jenkins G. (1970) Time Series Analysis, Forecasting and Control. Holden-Day, San Francisco, CA.

[5] Kamruzzaman J, Begg R, Sarker R. (2006) Artificial Neural Networks in Finance and Manufacturing, Idea Group Publishing, USA.

[6] Zhang G., Patuwo B.E., Hu M.Y. (1998) Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting 14:35-62.

[7] Hamzacebi C. (2008) Improving artificial neural networks' performance in seasonal time series forecasting. Information Sciences 178: 4550-4559.

[8] Farooq T., Guergachi and Krishnan S. (2007) Chaotic time series prediction using knowledge based Green's Kernel and least-squares support vector machines. Systems, Man and Cybernetics. ISIC. 7-10 Oct. 2007: 373-378.

[9] Raicharoen T., Lursinsap C., Sanguanbhoki P. (2003) Application of critical support vector machine to time series prediction. Circuits and Systems. ISCAS '03.Proceedings of the 2003 International Symposium on Volume 5, 25-28 May, pages: V-741-V-744.

[10] J.A.K. Suykens and Vandewalle J. (2000) Recurrent least squares support vector machines. IEEE Trans. Circuits Systems-I, 47(7):1109 – 1114.

[11] Hand DJ, Mannila H, Smyth P (2001) Principles of Data Mining. MIT Press, Cambridge, MA, USA

[12] Laxman S, Sastry PS (2006) A survey of temporal data mining. Sadhana 31:173–198

[13] Esling P, Agon C (2012) Time-series data mining. ACM Computing Surveys 45:12:1–12:34.

[14] Faloutsos C, Ranganathan M, Manolopoulos Y. (1994) Fast subsequence matching in time-series databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data; Minneapolis, Minnesota, USA: 419–29.

[15] Han J, Kamber M, Jei P. (2012) Data mining: Concepts and techniques. 3rd ed. Morgan Kaufmann Publishers.

[16] Keogh, EJ. (2006) A decade of progress in indexing and mining large time series databases. In: Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB); Seoul, Korea. p. 1268.

[17] Ding H, Trajcevski G, Scheuermann P, Wang X, Keogh EJ. (2008) Querying and mining of time series data: Experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment. 1:1542–52.

[18] R. Agrawal, C. Faloutsos, and A. Swami. (1993) "Efficient similarity search in sequence databases," in Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms, pp. 69–84.

[19] K.-p. Chan and A. Wai-chee Fu. (1999) "Efficient time series matching by wavelets," in Proceedings of the 15th International Conference on Data Engineering, pp. 126–133.

[20] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. (2001) Dimensionality reduction for fast similarity search in large time series databases. Knowledge and Information Systems, vol. 3, no. 3, pp. 263–286.

[21] E. Keogh and M. Pazzani, (1998) "An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback," in Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98), pp. 239–241.

[22] Q. Chen, L. Chen, X. Lian, Y. Liu, and J. Xu Yu, (2007) "Indexable pla for efficient similarity search," in Proceedings of the 33rd International Conference on Very Large Data Bases, pp. 435–446.

[23] J. Lin, E. Keogh, L. Wei, and S. Lonardi, (2007) "Experiencing sax: a novel symbolic representation of time series," Data Mining and Knowledge Discovery, vol. 15, no. 2, pp. 107–144.

[24] D. J. Berndt and J. Clifford, (1994) "Using dynamic time warping to find patterns in time series," in KDD Workshop, pp. 359–370.

[25] M. Vlachos, G. Kollios, and D. Gunopulos, (2002) "Discovering similar multidimensional trajectories," in Proceedings of 18th International Conference on Data Engineering, pp. 673–684.

[26] L. Chen, M. T. Ozsu, and V. Oria, (2005) "Robust and fast similarity search for moving object trajectories," in Proceedings of the 2005. ACM SIGMOD International Conference on Management of Data, pp. 491–502.

[27] L. Chen and R. Ng, (2004) "On the marriage of lp-norms and edit distance," in Proceedings of the 30th International Conference on Very Large Data bases, Volume 30, pp. 792–803.

[28] V. Kurbalija, M. Radovanovic, Z. Geler, and M. Ivanovic, (2014) "The influence of global constraints on similarity measures for time-series databases," Knowledge-Based Systems, vol. 56, pp. 49–67.

[29] Bo Sjö (2011), Lectures in Modern Economic Time Series Analysis, Editia a doua, Universitatea din Linköping, Suedia.

[30] Funk, P., Xiong, N. (2008) Extracting knowledge from sensor signals for case-based reasoning with longitudinal time series data. In: Perner, P. (ed.) Case-Based Reasoning in Signals and Images, pp. 247–284. Springer, Heidelberg

[31] Shahar Y., Miksch S., Johnson P. (1998) The Asgaard project: a taskspecific framework for the application and critiquing of time-oriented clinical guidelines. Artif Intell Med:14:29—51.

[32] Funk, P., Xiong, N. (2006) Case-Based Reasoning and Knowledge Discovery in Medical Applications with Time Series. Computational Intelligence: 238–253

[33] Begum S., Ahmed M. U., Funk P., Xiong N., Folke M.. (2011) Case-Based Reasoning Systems in the Health Sciences: A Survey of Recent Trends and Developments. IEEE Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews

[34] Huijse P., Estevez P.A., Protopapas P., Prıncipe J.C., Zegers P. (2015) Computational Intelligence Challenges and Applications on Large-Scale Astronomical Time Series Databases. arXiv:1509.07823v1 [astro-ph.IM] 25 Sep.

[35] V. Kurbalija, M. Radovanovic, Z. Geler, and M. Ivanovic, (2010) "A framework for time-series analysis," in Proceedings of the 14th International Conference, AIMSA, pp. 42–51.

[36] Z. Geler, V. Kurbalija, M. Radovanovic, and M. Ivanovic, (2015) "Comparison of different weighting schemes for the knn classifier on time-series data," Knowledge and Information Systems, pp. 1–48.

[37] Mitrovic D, Ivanović M, Geler Z (2014) Agent-Based Distributed Computing for Dynamic Networks. Information Technology And Control 43:88–97.

[38] Kurbalija V, Ivanović M, von Bernstorff C, et al (2014) Matching Observed with Empirical Reality– What you see is what you get? Fundamenta Informaticae 129:133–147.

[39] Kurbalija V, von Bernstorff C, Burkhard H-D, et al (2012) Time-series mining in a psychological domain. In: Proceedings of the Fifth Balkan Conference in Informatics on - BCI '12. ACM Press, New York, New York, USA, pp 58–63.

[40] Kurbalija V, Radovanović M, Ivanović M, et al (2014) Time-series analysis in the medical domain: A study of Tacrolimus administration and influence on kidney graft function. Computers in Biology and Medicine 50:19–31

[41] Kurbalija V, Ivanović M, Budimac Z, (2009) Case-based curve behaviour prediction. Software: Practice and Experience 39:81–103.

PAPERS

WIRELESS COMMUNICATIONS AND TECHNOLOGIES

Wireless Resonant Power Transfer – An overview

Žarko Martinović, Martin Dadić, Roman Malarić, Željko Martinović
Faculty of Electrical Engineering and Computing / Department of Electrical Engineering Fundamentals and Measurements
Unska 3, 10 000 Zagreb, Croatia
e-mail: zmartinovic@systec.danieli.com,
martin.dadic@fer.hr, roman.malaric@fer.hr, zeljko.martinovic@fer.hr

Abstract – Wireless power transfer represents the concept of energy transfer through resonant magnetic coupling based on Maxwell electromagnetic equations. The objective of this paper is to give an overview of the wireless power transfer. The introduction will give the description of the technology, describe the system components and working model. The focus will be kept on electromagnetic phenomena inside the resonant circuit. Furthermore, the paper will present a comparison between standard (conducted) and wireless resonant Witricity technologies.

I. INTRODUCTION

Witricity idea was born from Tesla's vision. His idea was to enable distribution of energy through Earth's ionosphere [1]. Tesla didn't scientifically prove his innovative idea, but he made significant experiments. Tesla's visions and conducted experiments resulted with hundreds of patents which have become his legacy. His researchs was based on nonradiative [2][3] and radiative wireless power transfer [4]. Several problems appeared during his research experiments such as low efficiency of power transfer [5][6], dangerous nature of the experiments and expensive equipment (financial) problems.

Wireless power (energy) transfer standards [7] can be divided:

- Standard, conducted, inductive, contactless method (electromagnetic coupling),
- EM (electromagnetic) radiation
- Resonant method (magnetic coupling)

1. 1. Inductive method

The inductive method is based on magnetic field induction principle. Electrical transformer represent an example of this working method. As it is known, the transformer has primary and secondary coil. Primary coil is connected to supply unit, while secondary coil by mutual inductance induce voltage [8]. This method is used worldwide in application such as electric toothbrushes, charging pad for cell phone or laptops, medical implants and etc. [9]

1. 2. Radiation method

Electromagnetic radiation is the radiative wireless energy transfer. This method is based on antenna use. System demands are transmitting and receiving antenna. Propagation channel without obstacles can be modeled with by Friis equation [10]

$$P_{RX} = P_{TX} \cdot G_{TX} \cdot G_{RX} \cdot \left(\frac{\lambda}{4\pi r} \right)^n \qquad (1)$$

with folowing parameteres:

n - Channel propargation coefficient

P_{RX} - Received power

P_{TX} - Transmit power

G_{TX} - Gain of transmit antenna

G_{RX} - Gain of receive antenna

λ - Wavelength

r - Seperation distance

Loses can be generated: signal transmission reflection (an incompatibility of antenna); obstacles reflection and diffraction between transmitting and receiving antenna; antenna polarization losses; antenna directivity losses.

Radiative transfer we can divide into omnidirectional and unidirectional radiation [11].

We should distinguish wireless power transfer and wireless information transmission. Electromagnetic waves transmit magnetic and electric energy through medium with different velocity. Power strength is represented by Poyting vector.

1. 3. Resonant method

More than one hundred years after Tesla vision, during the 2006. reasearch team lead by Marin Soljacic presented the idea of energy transmission by wireless resonance magnetic coupling [12]. Significant improvement in wireless power transfer was made a year later, during the

2007, by transfering 40% of energy on distance of 2m using wireless transfer by resonant coupling [13].

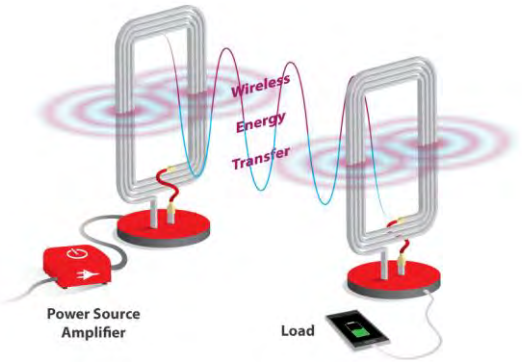


Figure 1. Equivalent circuit for Witricity

Resonant (nonradiative) transfer of energy is based on near – field magnetic coupling of conductive loops. Energy transfer was done on short distance, as we mentioned previously. The distance depends on the diameter of the loops [3]. Resonant power transfer can be achieved by inductive or capacitive approach. The papers that applies inductive method are [14][15][16][17][18]. Capacitive method has a lot of dissadvantages such as coupling area, frequency range connected with coupling area and power densitiy [19][20][21][22].

From aspects of the range power, wireless power transfer can be divided according to Table 1[23].

Table 1. WPT strenght and range

Range	P	f
Low	< 5W	110-205 kHz
Medium	5-120W	6.78-13.56 MHz
High	3-7 kW	20-150 kHz

Field of low power range is most active in wireless power transfer. During 2010. hundred and fifty companies arround the world announced idea of establishing Wireless Power Consortium (WPC). Purpose of the WPC is to set international standards for wireless power transfer in charging. Shortly after, WPC started with they program/technology „Qi“. Program has intention to solve problems in wireless power transfer in frequency range of 20kHz to maximum 5MHz. Another purpose of consortium and program is engourage reasearches to generate new more effective solution related with wireless power charging [24]. The high power range field is oriented mostly on vehicle and hybrid power charging. Research in this field had huge progress recently, but still there are a lot of the problems with frequency range. Except this, it is necessary for charging vehicle to have large air gap, high efficiency and a large amount of the power [25].

Another technology development in progress is bi - directional. Bi – directional technology operates as a current source in order to facilitate bi – directional power flow with inherent short – circuit protection. From both sides (transmitter/receiver) controlled rectifiers are set. [26][27].

II. WIRELESS (RESONANT) POWER TRANSFER

A. Resonance

Resonance is a pshysical phenomeneon which ocur in nature [28]. Two main parameters which can describe resonance are:

- resonant frequency ω_o

- intrinsic loss rate τ

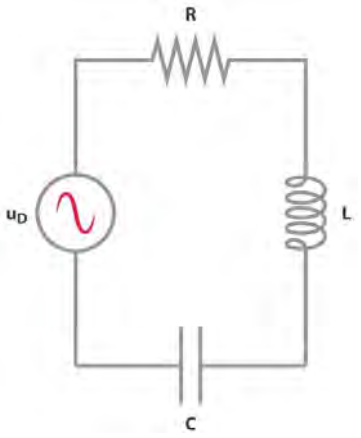


Figure 2. Equivalent circuit for series RLC with sine – wave voltage source

Second Kirchoff"s law applied to Figure 1. yields:

$$u_L + u_R + u_C = u_D \sin(\omega t) \tag{2}$$

$$i = \frac{dq_C}{dt} \tag{3}$$

$$L\frac{d^2 q_C}{dt^2} + R\frac{dq_C}{dt} + \frac{q_C}{C} = v_D \sin(\omega t) \tag{4}$$

If we divide relation (4) with L:

$$\frac{d^2 q_C}{dt^2} + \frac{R}{L}\frac{dq_C}{dt} + \frac{q_C}{LC} = \frac{1}{L} v_D \sin(\omega t) \tag{5}$$

$$\frac{d^2 q_C}{dt^2} + \frac{2}{\tau}\frac{dq_C}{dt} + \omega_o^2 q_C = \frac{1}{L} v_D \sin(\omega t) \tag{6}$$

We can derive value of the resonant frequency:

$$\omega_o = \frac{1}{\sqrt{L \cdot C}} \tag{7}$$

and intrinsic loss rate:

$$\tau = \frac{2L}{R}\,. \tag{8}$$

Energy stored in the system can be described with quality factor Q. Quality factor is the ratio of stored energy of the resonant circuit and the energy consumed in load.

$$Q = \frac{\omega_o}{2 \cdot \Gamma} = \sqrt{\frac{L}{C}} \frac{1}{R} = \frac{\omega_o \cdot L}{R} \tag{9}$$

From the power supply side, we can distinguish free and forced oscillation. Free oscillation can be gained by use of a short circuit on the source (v_D). In the initial moment, the current does not flow and the capacitor is charged with the initial charge q_{C0}. The energy oscillates at the resonant frequency (ω_o) between inductor (energy stored in the magnetic field) and the capacitor (energy stored in the electric field). Also, energy is dissipated in the resistor. When resistance is in the circuit very small, free oscillations are in form:

$$q_c = q_{C0} e^{-t/\tau} \cos(\omega t + \phi) \tag{10}$$

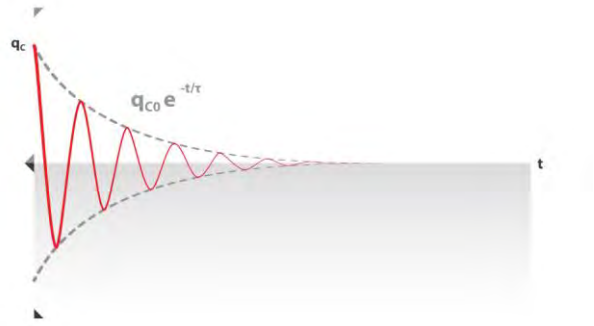


Figure 3. Free oscillation

The second case is with sine – wave source instead of short circuit. This case describes the forced oscillations. The solution is in the form:

$$q_C = A \sin(\omega t + \phi) \tag{11}$$

In Figure 4. we can notice the value of amplitude at the resonant frequency ω. If the frequency of the source is increased from zero, the amplitude of the oscillation increase until reaching a peak. At the maximum of frequency, amplitude of oscillation is higher than the amplitude of the source.

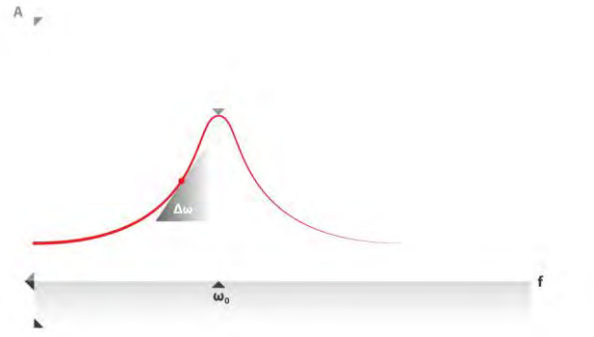


Figure 4. Amplitude of forced oscillation

B. *Coupled resonators*

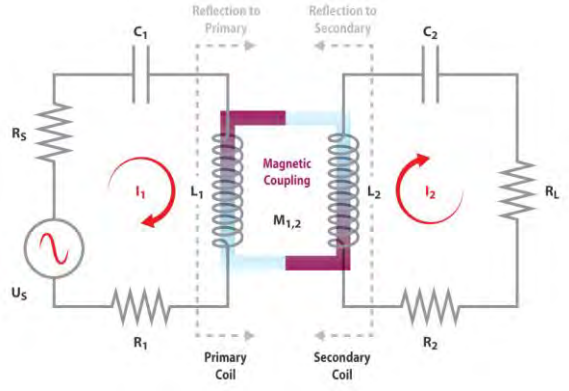


Figure 5. Equivalent circuit for coupled resonator system

Fig. 5 depicts a system of two coupled resonators with following parameters:

U_s - Voltage source
ω - Frequency
R_s - Internal resistance of the source
L_1 - Source inductance
L_2 - Device inductance
$M_{1,2}$ - Mutual inductance
C_1 - Source capacitors
C_2 - Device capacitors
R_L - Load resistance
I_1 - Source current
I_2 - Load current

Applying the electromagnetic theory, we can determine effective inductance and capacitance as:

$$L_{1,2} = \frac{\mu}{4\pi |I_0|^2} \iint dr dx' \frac{J(r) \cdot J(r')}{|r - r'|} \tag{11}$$

$$\frac{1}{C_{1,2}} = \frac{1}{4\pi e_0 |q_0|^2} \iint dr dr' \frac{\rho(r) \cdot \rho(r')}{|r - r'|} \tag{12}$$

With following parameters:

$J(r)$ - Spatial current
$\rho(r)$ - Charge density
I_0 - Inductance current
q_0 - Capacitance current

They are two possibilities how to reach maximum power transfer, both sources dependent. First of them use a constant AC current source. The second is with the constant AC voltage source.

1) Constant AC Current Source

The maximum wireless power transfer is maximized when the load current is at maximum value of the current, which is calculated from equivalent circuit:

$$I_2 = \frac{-\omega^2 M_{1,2} C_2}{(j\omega C_2) R_L + 1 - \omega^2 L_2 C_2} \cdot I_1 \qquad (13)$$

If we assume R_L as infinitely small value, the resonant frequency will be same as the resonant frequency of a parallel resonant circuit:

$$\omega_o (I_2 = \infty) = \frac{1}{2\sqrt{L_2 C_2}} \qquad (14)$$

2) Constant AC Voltage Source

Maximum wireless power transfer can be achieved by maximum load current I_2:

$$I_2 = \frac{-j\omega^3 M_{1,2} C_1 C_2}{((j\omega C_1) R_1 + 1 - \omega^2 L_1 C_1)((j\omega C_2) R_L + 1 - \omega^2 L_2 C_2) - \omega^4 M_{1,2}{}^2 C_1 C_2} \cdot U_s \qquad (15)$$

The resonant frequency is the same as the resonant frequency in a series resonant circuit:

$$\omega_o (I_2 = \infty) = \frac{\sqrt{L_1 C_1 + L_2 C_2 - \sqrt{L_1^2 C_1^2 + L_2^2 C_2^2 - 2(1-k^2) L_1 C_1 L_2 C_2}}}{2\pi\sqrt{2(1-k^2) L_1 C_1 L_2 C_2}} \qquad (16)$$

Ohmic and radiation resistance [12] of the transmitting coil can be, using the magnetic and electric dipole representation, determined as:

$$R_o = \sqrt{\frac{\mu_o \omega}{2\sigma}} \frac{l}{4\pi a} \qquad (17)$$

$$R_r = \sqrt{\frac{\mu_0}{\varepsilon_0}} \left[\frac{\pi}{12} n^2 \left(\frac{\omega r}{c} \right)^4 + \frac{2}{3\pi^3} \left(\frac{\omega h}{c} \right)^2 \right] \qquad (18)$$

With following parameters:

σ - Material conductivity
c - Speed of light
n - Number of turns
r - Radius of coils
a - Area of coils
h - Planck's constant

First part of Eq. 18 is parameter of magnetic dipole. It is necessary to achieve $r << 2\pi c / \omega$. Second part is representation of the electric coil dipole.

C. Wireless(resonant) power transfer

In this chapter we observe what's going on with efficiency and with power transfer in resonant state. Serial resonance wil be obtain in bi – directional mode of coils. In order to achieve necessary parameters we obtain models for primary and secondary coil according to the article [29]. Primary reflection obtain reflection of impedance and on secondary side reflection of impedance and source.

At Fig. 6. we can notice primary coil and their secondary reflection in primary circuit. At Fig. 7. we can see primary to the secondary reflection.

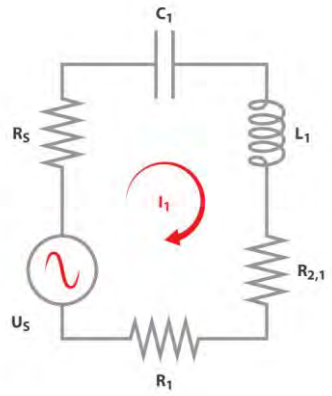


Figure 6. Simplified impedance matching by making secondary impedance reflection to be equivalent impedance in the primary coil

Resistance of secondary reflection in the primary coil:

$$R_{2,1} = \frac{\omega^2 M_{1,2}{}^2}{Z_2} \qquad (19)$$

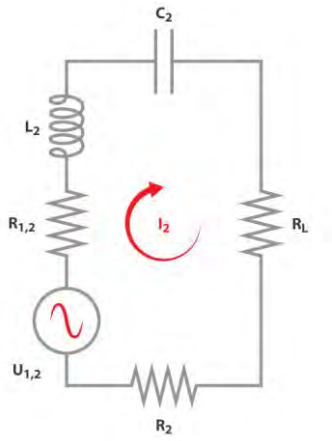


Figure 7. Simplified impedance and source matching by making primary reflection to be equivalent impedance in the primary coil

Resistance and source reflection to the secondary coil:

$$R_{1,2} = \frac{\omega^2 M_{1,2}{}^2}{Z_1} \qquad (20)$$

$$U_{1,2} = \frac{j\omega M_{1,2}}{Z_1} U_s \tag{21}$$

Modification of the equation (19) gives:

$$R_{2,1} = \frac{\omega^2 M_{1,2}{}^2}{R_2 + R_L} \tag{22}$$

Efficiency transfer can be determined as:

$$\eta = \frac{\omega^2 M_{1,2}{}^2 R_L}{(R_2 + R_L)\left[\omega^2 M_{1,2}{}^2 + (R_1 + R_s)(R_2 + R_L)\right]} \tag{23}$$

Mutual inductance of the system from Fig.5 is:

$$M = k\sqrt{L_1 L_2} \tag{24}$$

Efficiency of the system of Fig. 5 depends about resonator's parameter characteristic resonator and the energy coupling rate (k) between them. The energy coupling rate represents the fraction of the flux that's generated in primary coils and coupled in secondary coils. It can be maximized by selecting the radius of the coils.

Efficiency of system transfer can be increased by coil distance decreasing.

According to Tesla, we should analyze the connection between total available and delivered (load) power across frequency [3].
Maximum power transfer we can determine when the mutual inductances are fixed.

$$P_L = \frac{\omega^2 M_{1,2}{}^2 U_s{}^2 R_L}{\left[\omega^2 M_{1,2}{}^2 + (R_1 + R_s)(R_2 + R_L)\right]^2} \tag{25}$$

Highest efficiency we gain when the load is not fixed. Value of the load with maximum efficiency transfer and with maximum power transfer represents relations (25) and (26). In conclusion, adjustable value of load is not suitable for resonant circuit.

$$R_{L_{\eta\max}} = \sqrt{\frac{\omega^2 M_{1,2}{}^2 R_2}{R_1 + R_s} + R_2{}^2} \tag{25}$$

$$R_{L_{P_{L\max}}} = \frac{\omega^2 M_{1,2}{}^2 R_2}{R_1 + R_s} + R_2 \tag{26}$$

In this case mutual inductance with maximum power transfer we determined as:

$$M_{P_{L\max}} = \frac{\sqrt{(R_2 + R_L)(R_1 + R_s)}}{\omega} \tag{27}$$

Resonance coupling is simillar to communication between two communication systems. If they would like to communicate, both systems should be on the same level of communications. In this case, resonant system will exchange energy if they have the same level – same resonant frequency of the transmitter and receiver side.

Wireless resonant power transfer system consists from:

- power supply
- source resonators
- source loop
- load resonators
- load loop

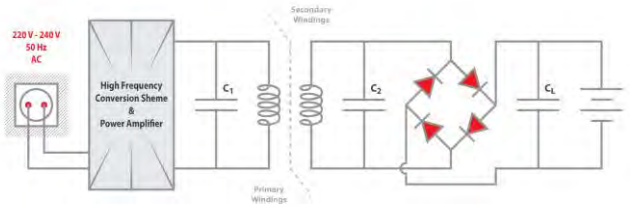


Figure 8. Basic design of wireless power transfer

From Fig. 8 we can see a basic system of wireless power transfer. By conducting resonance in transmitter circuit – we achieve power toward receiving circuit without electromagnetic waves.

As we mentioned previously, resonant coupling has a high energy efficiency transfer. High efficiency depends on resonant coils and their position. Position of them should be coaxially. Coils in both circuit needs to be on the same axis. To achieve high energy effciency transmitter and receiving coils shouldn't rotate in all directions[12].

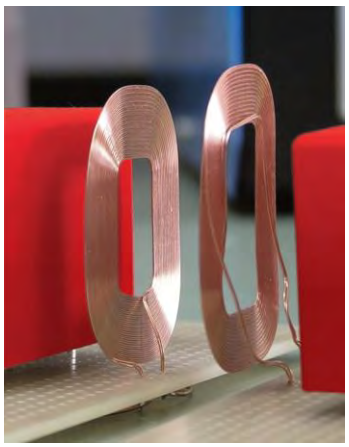


Figure 9. Resonant wireless power transfer in coaxially mode

Since 2006. when the MIT team developed Witricity power transfer based on resonant coupling, many new research appeared with the same aim. Increase power transfer efficiency with extension of transmission range have became a scientific world wide challange. Except

efficiency, other challenges are compact size and interference with other electromagnetic devices.
In recent years, researchers around the world conducted experiments changing different parameters such as: air gap, numbers and radius of the coils, number of turns and compact size (area). In Table 1. we can see the most relevant outputs of those researches in Witricity technology [30][31][32][33][34].

Table 1. Overview of wireless power transfer

Example	Turns [-]	Air gap [mm]	Frequency [MHz]	Efficiency [%]
1	6	50	5.5	73
2	10	30	43	48.8
3	14	30	30.4	51.4
4	18	30	23.5	48.8

From aspects of the coils, Kurs et al. presented idea of multiple receiving coils. By conducting of this technology, power can be transfered on multiple receiving devices simultaneously. Multiple receiving systems consists of three parts - receiving, transmitting and intermediate. Energy transfer is enabled via intermediate coil. Related to coils rotation, efficiency range is between 29% and 61.2% [35]. During 2008. Intel Company presented wireless power transfer of energy using flat coils [36] instead helix coils [7]. MIT team also presented the idea of several receiver coils [37]. Their idea became system with huge transmitter coil and energy transferring toward multiple smaller receiver coils. Analysis conclusion was that this method has better efficiency than original prototype [13]. Other experiments were made with four coil energy power transfer [38]. Four coil method is used in sensors network for medical implant [39]. According to the distance of transmission, efficiency was between 72% and 83%.
Research field in progress, but still worth of attention is field with adjustable frequency resonance body for biosensors [40][41][42]. According to the relation (4) this technology is oriented toward different materials which are used for resistors production. Different resistor materials can give us different stability and frequency respond as well different losses e.g.. If we reduce resistance losses, efficiency can be increased. Adjustable technology uses silver as material for receiving coils in order to reduce thermal resistance losses. Researches tried to use a different refractive material index to improve wireless power transfer [43][44]. One example is linear AC resistance used to reduce thermal losses [45].

Analytical study [46] related to interference between objects in resonant coupling showed the object position influence in the circuit. Object position can be in the middle between transmitter and receiver, or near to transmitter/receiver antenna side. According to the study, if the object is near to the transmitter side, the impedance of the source can change. Consequently, the power of the source can be changed. If the object is in the middle, interfering object influence on the receiver side. With object magnetic field, current in receiver circuit can be increased. Third case, if the interfering object is near to receiver side, magnetic field on receiver side is lowered. Because of the close proximitiy between interfering object and receiver coils this lead to impedance change as well power change.
Interaction of Witricity system with human body became hot scientific field in recent years. According to the study presented in [47] we can see that the human body is modified as interaction element - rectangular dielectric box. The box was made from dielectric material that are similar to the human tissue. Similartiy is set up related with dielectric constant (ε) and conductivity (σ) of tissue. If the box is near to the receiver side, resonant frequency fall down. Consequently, increases SAR (Specific Absorption Rate) of human tissue. The study presents also the value of distance where this factor should be smaller. According to the presented study if distance is greater than 0.5m – tissue absorption will be at a minimum. This mean SAR will be less than 2W/kg for the whole body. Power transfer efficiency related to human body interaction is reduced because of lower tissue conductivity. Tissue characteristics are the main reason for electric wave component attenuation. Magnetic field component decreases with square distance. As we can see electromagnetic field has a limitation with human body interaction.

III. CONCLUISION

One hundred year after Tesla's vision, Witricity became reality. In last decade witricity development was extraordinary expanded. Many research fields were opened in as output of Witrictiy development in recent years. Witricity technology became a reality with practical use in laptops, smarthphones, electric vehicles, medical implants and etc. Mid – range wireless power transfer will reduce problems with cabling, especially during charging of devices such as phones, tablets and etc. Witricity has intention to solve problems with electric vehicle charging. Practical use of witricity technology is still in progress. Purpose of Witricity technology is to enlarge productivity of new devices more convenient without batteries, cabling, more reliable and with decreased device costs. In order to achieve those aims, reasearchers around the world, put efforts to increase efficiency and distance of wireless transfer.

Literature

[1] Nikola Tesla, „ The transmission of electrical energy without wires“, Electrical World and Engineer, March 1905.

[2] H. Winfield Secor, „Tesla apparatus and experiments – how to build both large and small Tesla and Oudin coils an hot to carry on spectacular experiments with tem,“ Practicle Electrics, Nov. 1921.

[3] R. Lomas, The Man Who Invented the Twentieth Centruy: Nikola Tesla – Forgotten Genius of Electricity. New Yor, USA, QCS eBooks, 1999, p. 146, ISBN: 0 7472 6265 9.

[4] Nikola Tesla, „Systems of transmission of electrical energy,“ U. S. Patent 645 576, Mar. 20, 1900.

[5] E. Waffenschmidt and T. Staring, „Limitation of inductive power transfer for consumer applications,“ in Proc. Eur. Conf. Power Electron, Appl., 2009, pp, 1-10.

[6] J. O. Mur – Miranda, G. Fantim, Y. Feng, K. Omanakuttan, R. Ongie, A. Setjoadi, and N. Sharpe, „Wireless power transfer using

weakly coupled magnetostatic resonators,“ in Proc. IEEE Energy Conv. Congr. Expo. Conf., 2010, pp. 4179-4186.

[7] S. Y. Hu, „Planar Wireless Charging Technology for Portable Electronic Products and Qi“, in Proc. IEEE, 2013, v. 101, p. 1290-1301.

[8] C. V. Jones, „Unfield Theory of Electrical Machines, London, U.K.; Butterworths, 1967.

[9] G. Wang et al., „Power Supply Topologies for Biphasic Stimulation in Inductively Powered Implantas,“ Proc. IEEE Int. Symp. Circuits and Syst. (ISCAS), Kobe, Japan, May 23-26, 2005, pp. 2743-46.

[10] D.C. Hogg, „Fun with the FRIIS free-space transmission formula“, Antennas and Propagation Magazine, IEEE , vol. 35, pp. 33-35

[11] A. Sample et al., „Design of an RFID – based Battery – Free Programable Sensing Platform,“ IEEE Trans. Instrumentation and Mesearument, vol. 57, no. 11, Nov. 2008, pp. 2608-15.

[12] Andrè Kurs, Aristeidis Karalis, Robert Moffatt, J. D. Joannopoulos, Peter Fisher, Marin Soljačić, „Wireless Power Transfer via Strongly Coupled Magnetic Resonances,“ Science, vol. 317, pp. 83-86, July 2007.

[13] A. Karalis, J. D. Joannopoulos, Marin Soljacic, „Efficient wireless non radiative mid – range energy transfer“, Annals of Physics, (2008), vol.3 no. 23, pp 34-48.

[14] Y. Jang and M. Jovanovic, “A contactless electrical energy transmission system for portable – telephone battery chargers,” IEE Trans. Ind. Electron., vol. 50, no.4, pp. 520-527, Jun. 2003.

[15] C. – G. Kim, D. – H. Seo, J. – S. You, J. – H. Park, and B.H. Cho, „Design of a contactless battery charger for cellular phone,“ IEEE Trans. Ind. Electron.,vol. 48, no. 6, pp. 1238-1247, Dec. 2001.

[16] T. Bieler, M. Perrottet, V. Nguyen, and Y. Perriard, „ Contact power and infromation transmission,“ IEEE Trans. Ind. Appl., vol. 38, no. 5, pp. 1266-1272, Sep. – Oct. 2002.

[17] K. Oguri, „Power supply coupler for battery charger,“ U.S. Patent 6 356 049, 2000.

[18] H. Brockmann and H. Turtianen, „Charger with inductive power transmission for batteries in a mobile electrical device,“ U. S. Patent 6 118 249, 1999.

[19] G. Roberts, A. Owens, P. Lane, M. Humphries, R. Child, F. Bauder and J. Izquierido, „ A contactless transfer device for power and data,“ in Proc. IEEE Aerosp. Appl. Conf., 1996, vol. 2, pp. 333-345.

[20] L. Chao, A. P. Hu, and D. Xin, „A contactless power transfer system with capacitevely coupled matrx pad,“ in Proc. IEEE Energy Conv. Congr. Expo., 2011, pp. 3488-3494.

[21] M. Kline, I. Izyumin, B. Boser, and S. Sanders, „Capacitive power transfer for contactless charging,“ in Proc. IEEE Appl. Power Electron Conf. Expo., 2011, pp. 1398-1404.

[22] H. Fnato, Y. Chiku, and K. Harakawa, „Wireless power distribution with capacitive coupling excited by switched mode active negative capacitor,“ in Proc. Int. Conf. Electr. Mach. Syst., 2010, pp. 117-122.

[23] Shuichi Obayashi, Hitoshi Tsukahara, ”EMC Issuess on Wireless Power Transfer, Electromagnetic Conference, Tokyo, 2014.

[24] Wireless Power Consortium “Qi system description: Wireless power transfer,“ Volume 1:Low Power, Part 1: Interface Definition, Version 1.1, Apr. 2012.

[25] T. C. Beh, M. Kato, T. Imura, S. Oh, Y. Hori, „ Automated Impedance Matching System for Robust Wireless Power Transfer via Magnetic Resonance Coupling“, IEEE Industrial Electronics, 2013., pp. 3689-3698,

[26] Nayuki, et. al., Electric Power Engineering Research Laboratory Rep., No. H10007, March 2011.

[27] Madawala, U. K. et all., IEEE Trans. IE, Vol. 58, No.10, pp 1789-1796, Oct. 2011.

[28] Vujnović Momir, „Oscilaori“, Školska knjiga, 1987.

[29] X. Wei, Z. Wang, H. Dai, „ A Critical Review of Wireless Power Transfer via Strongly Coupled Magnetic Resonances“, Journal Energies, no.7, pp. 4316-4341, 2014.

[30] H. Zhou, S. Yang, „Resonant Frequency Calculation of Witricity Using Equivalent Circuit Model Combined with Finite Element Method,“ International Conference on Electromagnetic Field Problems and Applications (ICEF), 2012 Sixth, vol., no., pp1, 4, 2012.

[31] H. Zhou, J. Wang, W. N. Fu, M. Sun, „ A comparative Study Between Novel Witricity and Traditional Inductive Magnetic Coupling in Wireless Charging,“ IEEE Transactons on Magnetics, vol. 47, no. 5, pp. 1522, 1525, 2011.

[32] J. Wang, S. L. Ho, W. N. Fu, M. Sun, „Finite element analysis and corresponding experiments of resonant enery trasmission for wireless transmission devices using witricity,“ 14th Biennial IEEE Conference on Electromagnetic Field Computation (CEFC), on, vol., no., pp. 1,1, 2010.

[33] J. Wang, S. L. Ho, W. N. Fu, M. Sun, „Analytical Desing Study of a Novel Witricity Charger With Lateral and Angular Misaliagnments for Efficient Wireless Energy Transmission,“ IEEE Transactions on Magnetics, vol. 47, no. 10, pp. 2616, 2619, 2011.

[34] M. H. Salleh, N. Seman, R. Dewan, „ Reduced – Size Witricity Charger Design and Its Parametric Study,“ IEEE International RF and Microwave Conference (RFM), December 9-11, 2013.

[35] R. E. Hamam, A. Karalis, J.Joannopoulos, „Efficient weakly – radiative wireless energy transfer: An EIT – Like approach“, Annuals of Physics, (2009), vol. 324, no. 8, pp. 1783-1795.

[36] L. Xie, Y. Thomas Hou, W. Lou, „ Wireless Power Transfer and Applications to Sensor Networks“,

[37] A. Kurs, R Moffatt, and M. Soljacic, „ Simultaneous Mid – Range Power Transfer to Multiple Devices,“ Appl. Phys. Lett., vol. 96, Jan 2010, pp. 044102-1-44102-3.

[38] A. K. Ran Rakhyani, S. Mirabbasi, C. Mu, „Design and Optimization of Resonance – Based Efficient Wireles Power Delivery Systems for Biomedical Implants“, Biomedical Circuits and Systems, IEEE Transactions, (2011), vol. 5. no.1, pp. 48-63.

[39] A. Kumar, S. Mirabbasi, M. Chiao, „Resonance – based wireles power delivery of implantable devices“, IEEE Biomedical Circuits and Systems Conference, (2009).

[40] Y.–H. Kim, S.-Y. Kang, M. –L. Lee, „Optimization of Wireless Power Transmision through Resonant Coupling“, Power, (2009), pp. 426-431.

[41] Y.–H. Kim, S.-Y. Kang, A. Cheon, „ Wireless Power Transmision to multi devices through resonant coupling“, International Conference on Electrical Machines and Systems (IECMS), vol.1, (2010), pp. 9-11.

[42] F. Zhang, S. A. Hackworth, X. Liu, „ Wireless energy transfer platform for medical sensors and implantable devices“, Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (2009), pp. 1045-1048.

[43] B. Wang, T. Nishino, „Wireless power transmission efficiency enhancement with metamaterials“, Wireless Information Technology, (2010), vol. 43, no. 7, pp. 708-714.

[44] B. Wang, K. H. Teo, T. Nishino, W. Yerazunis, J. Barnwell, J. Zhang, „ Experiments on Wireless Power Transfer with Metamaterials“, Applied Physics Letters, vol. 98, no.25, (2011).

[45] S.-H. Lee, R. D. Lorenz, „ Development and validations of model for 95% effciency, 220W wireless power transfer over a 30cm air – gap“, Energy Conversion Congress and Exposition, (2010), pp. 885-892.

[46] E. Bou, D. Vidal, R. Sedwick, E. Alarcon, „Magnetic Characterization of Interfering Objects in Resonant Inductive Coupling Wireless Power Transfer“, PIERS Proceedings, August 12 -15, 2013.

[47] Q. Yuan, Q. Chen, K. Sawaya, „ Effect of Human Body on Near – Field Resonant Coupling Wireless Power Transmission System“, Electromagnetic Conference (EMC), Kyoto, 2009.

Investigation of a Small Handheld PCB Nesting Two Antennas NFC 13.56 MHz and to RF 868 MHz

L. A. Iliev*, I. S. Stoyanov*, T. B. Iliev**, E. P. Ivanova** and Gr. Y Mihaylov**
*University of Ruse, Department of Electrical and Electronic Engineering, Ruse, Bulgaria
**University of Ruse, Department of Telecommunication, Ruse, Bulgaria
e-mail addresses: lailiev@uni-ruse.bg, stoyanov@uni-ruse.bg, tiliev@uni-ruse.bg, epivanova@uni-ruse.bg, gmihaylov@uni-ruse.bg

Abstract - The development of the electronics during the recent years aims at producing smaller devices with a very high integration, very energy efficient and cost-effective. There is also a tendency that every device should have such communication modules, so that it can connect to the cloud network which is evolving every day. Designing several different antennas into one single PCB, which is small in size, might be a problem and a challenge. Every electromagnetic antenna has its own requirements for structural design, placement on the PCB, physical surrounding components and objects, ground planes, environment dependency and many other factors. The article presents the process and results of a design of a small PCB for remote control, nesting two antennas: NFC 13.56 MHz and to RF 868 MHz. The investigation includes tuning the antennas with spectrum analyzer and impedance analyzer. Aspects of EMC are also examined.

Keywords: antenna, HF simulation, NFC

I. INTRODUCTION

The development of the electronics during the recent years aims at producing smaller devices with a very high integration, very energy efficient and cost-effective. There is also a tendency that every device should have such communication modules, so that it can connect to the cloud network which is evolving every day. For these reasons, device also must include different kinds of wireless or wired protocols. In case wireless communications are used, there should be an antenna or several antennas so that a single device be able to nest several communication interfaces like RF (MHz to GHz range) or near field communication (NFC). Designing several different antennas into one single PCB, small in size, might be a problem and a challenge. Every electromagnetic antenna has its own requirements for structural design, placement on the PCB, physical surrounding components and objects, ground planes, environment dependency and many other factors.

Adding up the other requirements such as minimum overall size of the device, minimum energy consumption, durability, cost-effectiveness the design process is getting very difficult.

The following article is about investigation of the antenna parameters of a small handheld device, a remote controller, which combines two wireless communication interfaces - RF 868 MHz and NFC 13.56 MHz. The device includes a keyboard, battery, TI's CC1200 chip for RF868 interface, TI's RF430CL330H chip for NFC, TI's MSP430FR5969 as a microcontroller, Atmel's ATECC508A for efficient cryptography algorithms for the transmitted data. The device two PCB antennas: one for 868 MHz and another for NFC. The size of the PCB is 40 mm x 55 mm and the overall size of the remote control is 45 mm x 70 mm.

II. ANTENNA FOR 868 MHZ COMMUNICATION

Wireless communication frequency of 868 MHz was chosen, because it is permitted in most of the countries worldwide, it has a larger range than 2.4 GHz, not so many devices are operating in this frequency range and the antenna size is smaller than for 433 MHz range [1].

There are several requirements for the antenna design for this remote control. It has to be as small as possible in size, very efficient, integrated into the PCB and not interfering with other components on the PCB.

Several different types of antennas were investigated, in order to choose the most suitable one for the design.

Whip (Monopole) antenna is a low cost, efficient solution, but it has a bigger size and it is complicated to nest it inside the enclosure and no outside antenna is permitted. Making the whip shorter than 1/4 wavelength lowers the efficiency and overall gain.

Spiral antenna has a lot of advantages concerning size/efficiency, but it is very dependant of the surrounding objects, in this case the remote control is hand held and the human hand will interfere with the antenna, resulting in reduced efficiency and altered antenna radiation pattern.

Chip antenna provides a very small size and good quality, but it is expensive compared to others and hard to tune. The loop antenna is a permissible solution, but is it very demanding for the ground plane. Altering the ground plane results in decline of the parameters [2, 3].

Helix antenna has a small size, because it consists of windings tuned to resonate. Usually it is built external to the PCB, making it not very suitable for manufacturing process and repeatability.

PCB helix antenna showed the best parameters for this design, because it has a small size, 12 mm x 19 mm is the calculated size for 868/915 MHz operation. The Total Radiated Power is measured to be 1.83 dBm and the Efficiency around 65 %. It has a relative homogene XY directivity, which is important for the remote control direction. [4].

III. DESIGN CONSIDERATIONS.

The impedance of this type of antenna is far from 50 Ω and in order to operate properly it has to be matched by additional circuit. It also requires a balun circuit for balancing, which is impedance matched. The matching circuit consists of capacitors and an inductor, chosen to represent 50 Ω impedance for the antenna. Fig. 2 shows this circuit.

Components C202 and L201 are the tuning elements for the 50 Ω impedance matching. Those values will vary depending on the PCB material, thickness, manufacturing process, ground plane size, external enclosure, temperature, humidity and objects situated near the antenna.

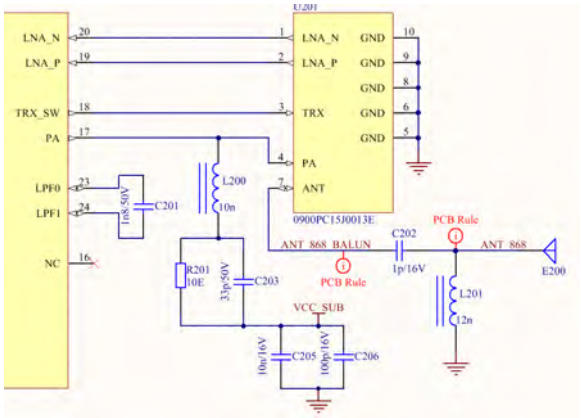


Figure 2. Tuning circuit for 868 MHZ helix PCB antenna.

As it is impossible to match the impedance exactly to 50 Ω in all conditions, it should be kept in the range from 45 to 55 Ω. Calculation and simulation results showed that the optimal values for the components are 1pF for the capacitor and 12nH for the inductor. There values might be changed after detailed tuning following the enclosure manufacturing. This tuning is performed using a spectrum analyzer and a coax cable connecting the antenna to the input of the analyzer. It is necessary to perform a calibration first with Open, Short and 50 Ω Load conditions in order to calculate the right results. After the calibration, the antenna is connected to the analyzer and S-Parameters are measured in dB. Also the complex impedance values can be derived from the S-Parameters measurements so that 50 Ω impedance can be tuned. Simulation results can be obtained using CST Studio Suite. A model exported from the PCB design software is imported into CST PCB Studio and all the parameters of the PCB are adjusted in correspondence with the real values. After that, the PCB is transferred to CST Microwave Studio as a 3D model containing all PCB layers and components. Fig. 3 shows a 3D model of the PCB imported into CST Microwave studio

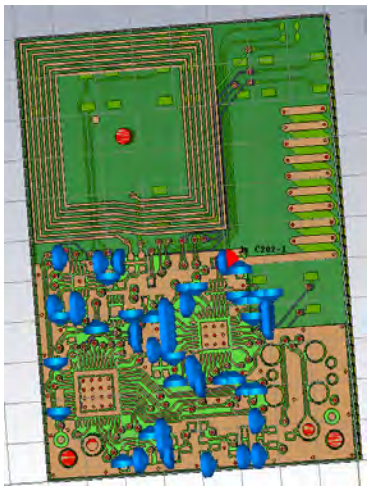


Figure 3. Remote control PCB imported in CST Microwave Studio.

In this case, S-Parameters are measured at the input of the antenna in reference to the ground plane. The software is able to calculate the impedance and the VSWR values taking into account the physical and electrical properties of the components and their placement.

The calculated impedance is within the range of 47-52 Ω for frequency range from 800 MHz to 900 MHz, which is acceptable. The next step is to make real measurements.

The RF868MHz was tested in real life conditions and it showed good results. The communication range was about 50 meters indoors and more than 300 m outdoors, which is quite enough for the intended operation. It does not interfere with the other components on the PCB and does not cause degradation on the device operation.

NFC antenna

Today NFC communication protocols are widely used in many aspects of everyday life. Price tags in the supermarkets, tickets in some transportation services, payment methods, authorization methods, etc. In the recent years the use of NFC is spreading drastically and it will keep on the same way in the future. There are more and more Smart phones produced with integrated NFC transponders and also other mass usable devices start to include NFC.

As NFC abbreviation says, this protocol is a Near Field Communication protocol and it is intended to operate in distance less than 10 cm. This is very suitable for activities which require high level of security, because RF transmitted data can be detected from a great distance, but NFC propagates only to about 10 cm from the tag and the reader. It cannot be detected from long distances.

NFC operates at different frequencies, but for the current design, 13.56 MHz is chosen, because it is more common today. Most of communication is performed using magnetic fields, rather than electromagnetic. There

are also 868/900 MHz NFC operation, but it is not considered in this article. NFC can be used in devices which operate without batteries, as the energy transferred to the antenna is sufficient to power the device and to exchange data.

The NFC antenna is simply a coil windings into a loop with a defined inductance. The inductance of the antenna depends on the following parameters: size, number of turns, space between turns, thickness of the wire. There are some software tools which provide a graphical design guides for NFC antennas. Fig. 5 shows the designed NFC antenna on the PCB [5].

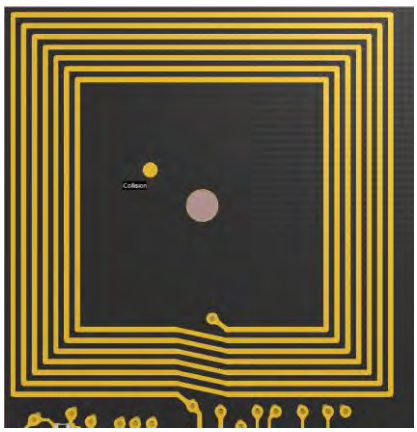


Figure 5. NFC antenna designed.

The designed NFC antenna is routed directly on the PCB and it has 7 windings, 0.35mm width, 0.35mm gap between them and the overall size of the antenna is 25mm X 25mm. The simulated impedance is 2.3 µH, which is acceptable.

The problem with the design is that any tracks or ground wires in the vicinity of the antenna decline its characteristics and must be avoided. Unfortunately it was not possible in the current design, so the tracks are moves aside as possible. This will however change the impedance at some point and respectively change the resonant frequency.

Simulation results from CST Studio showed that the NFC antenna designed has an inductance in the range from 2.1 to 2.5 µH which is close to the initial design values.

The antenna was physically tested with LG Nexus phone NFC reader and it successfully exchanged data from 2.5 cm range, which is considered a good performance. Fig. 6 shows the circuit for NFC transponder with the antenna.

If matching of the antenna capacitors is needed, it can be done by arraying the capacitor value until resonance is reached. For this reason 2 capacitors are placed (C300 and C301) parallel to each other in order the antenna to be tunable depending on the working conditions. Resistors R301 and R302 are placed so that the antenna Q-Factor can be adjusted. Q-Factor between 30 and 40 is required for optimal operation efficiency and stability. When adjustment is not needed, 0 Ω resistor values can be placed [6]. The RF430CL330 NFC transponder from Texas Instruments has the ability to detect NFC reader presence and initiate a wake-up signal for the MCU. When electromagnetic energy is induced into the antenna by the reader device, a transistor is triggered by the RF430CL330 IC and the MCU is awakened. This is very convenient feature especially from energy efficiency point of view. The device hibernates most of the time and only when a NFC reader presents near, the MCU is awaken and data is transmitted. Modes of PCB hibernation usually consume power less than 50pW. Wake-up states are short which saves battery life.

NFC antennas are affected by metal objects situated near them, so the enclosure must not contain metal planes covering the antenna. The data transmit rate is not required to be higher than 2-3 Kbit/s so design considerations are not so complex.

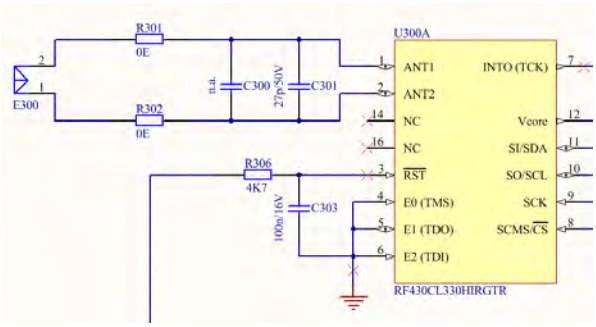


Figure 6. NFC circuit.

On overview of the designed PCB can be seen on Fig. 7. The main MCU is chosen to be MSP430FR5969 from TI, which is very energy efficient and provides enough flexibility, CC1200 IC from Ti as RF 858MHz transceiver. The power supply is designed with TPS61098 IC boost converter with low quiescent current and high efficiency. The main concept of the remote control device designed is to ensure high level of security and convenience for the user. This is the reason why NFC antenna is used as one of the communication methods, because it is more difficult to intercept than RF signal propagated up to hundreds of meters.



Figure 7. Remote control PCB overview.

Next steps

Next step will include EMC pre-compliance measurements when the device enclosure is manufactured. For this purpose Hameg HM5014-2 spectrum analyzer will be used with Log-periodic antenna 30-1000 MHz. When the pre-compliance is successful, then full compliance measurements will be made in EMC laboratory using anechoic chamber.

IV. CONLUSION

The current design proved by simulation and real measurements that this combination of two antennas (RF 868 MHz and NFC 13.56 MHz) is possible and gives satisfying results. The simulation was made using CST Studio Suite and imported PCB layout designed on a CAD system. By this method, many parameters were taken into account, as physical dimensions of the PCB and the components, placement and electrical parameters. Further optimization is needed in order to minimize the effect of the human hand in the current case of a remote control.

REFERENCES

[1] Compact Integrated Antennas, Freescale Semiconductor, Inc., Document Number: AN2731, Rev. 3, 09/2015.

[2] Antenna Design and RF Layout Guidelines, AN91445, Cypress Semiconductor Corp.

[3] Wireless communications design handbook, Volume 3, Reinaldo Perez, 1998.

[4] Miniature Helical PCB Antenna for 868 MHz or 915/920 MHz, Texas Instruments, Design Note DN038.

[5] NFC design navigator (web based) - http://www.semicon.panasonic.co.jp/en/tool/nfcdesignnavigator/.

[6] AN11535 - Measurement and tuning of a NFC and Reader IC antenna with a MiniVNA, NXP, Rev. 1.1 — 3 November 2014.

The Coverage Belt for Low Earth Orbiting Satellites

Shkelzen Cakaj[1, 2]

[1] Faculty of Electrical and Computing Engineering, Prishtina University, Kosovo

[2] Telecom of Kosovo, Prishtina, Kosovo

E-mail: shkelzen.cakaj@fulbrightmail.org

Abstract - **Low Earth Orbit (LEO) satellites are used for public networking and for scientific purposes. Communication via satellite begins when the satellite is positioned in its orbital position. Ground stations can communicate with LEO satellites only when the satellite is in their visibility region, since LEO satellites move too fast relative to a fixed point on the Earth. The satellite's coverage area on the Earth depends on orbit altitude and elevation. The satellite coverage area is defined as a region of the Earth where the satellite is seen at a minimum predefined elevation angle. Under zero elevation it is the largest coverage area. Because of LEO satellites movement related to a fixed observer on Earth surface, also the coverage area changes its position continually creating a coverage belt. The wideness of coverage belt strongly depends on orbit altitude. To conclude about the coverage belt variation for low orbiting satellites, the simulation for attitudes from 600km to 1200km is presented through this paper.**

Keywords - LEO; satellite; coverage belt.

I. INTRODUCTION

Generally, satellite' circular orbits are categorized as Geosynchronous Earth Orbits (GEO), Medium Earth Orbits (MEO) and Low Earth Orbits (LEO). The main difference among them is in the altitude above the Earth surface, thus consequently in different coverage areas [1], [2].

The satellites traversing in orbits of attitudes up to around 1400 km (limited by Van Allen belt [3]) are considered as LEO satellites. LEO satellites move at around 7.5 km/s velocity relative to a fixed point on the Earth (ground station). The characteristics of LEOs are: the *shortest distance* from the Earth compared with other orbits and consequently *less time delay*. Satellites in these orbits have an orbital period of around (90-110) minutes. The communication duration between the satellite and the ground station takes (5-15) minutes, (6-8) times over the day [1], [2]. Satellite coverage architecture shows considerable performance enhancement especially for Low Earth Orbit (LEO) satellite systems [4].

LEO satellites provide opportunities for investigations for which alternative techniques are either difficult or impossible to be applied. Thus, it may be expected that such scientific missions will be further developed in the near future especially in the fields where similar experiments by purely Earth-based means are impracticable. The Sun synchronization is a specific feature of LEO satellites utilized for remote sensing and Earth observation. [5].

From the communication perspective, the goal of the future satellite communication systems is to provide high quality broadband services with global coverage [6]. LEO satellite networks provide significant wide coverage area, unique broadcast capability and ability to meet different QoS requirements [7]. The satellite constellation is a convenient network solution for real time global coverage. The constellation is a system of low Earth orbit (LEO) identical satellites, launched in several orbital planes with the orbits having the same altitude. The satellites move in a synchronized manner in trajectories relative to the Earth. The application of low Earth orbit satellites organized in a *constellation* is an alternative to wireless telephone networks.

When designing a satellite network some decisions such as the selection of the orbit parameters, coverage model, the network connectivity and routing model must be made. LEO satellite networks as Iridium, Teledesic, Globalstar, have architecture differences affecting the capabilities and services. The main problem for the global coverage is the handover process from one coverage area to another one. This is called satellite hand over process [8]. Different deterministic models for coverage time evaluation of low Earth orbiting satellites are developed. Models involve statistical coverage time assessments [9]. The analyses are particularly useful for probabilistic investigation of intersatellite handovers in LEO satellite networks [9]. The probability of service interruption and hand over mechanism becomes important for the overall system performance [10]. Atmospheric impairments and interference aspects must necessarily be considered within link budget as system performance factors. [11], [12].

For complete coverage of the Earth's surface some overlapping between the adjacent satellites is necessary, to keep the continuity of real time services. The global coverage can be considered as an interoperable complementary networking process of multiple satellites organized in constellation, each of them contributing with its individual coverage. This is achieved because LEO satellites move with respect to a fixed observer on Earth surface, and along with satellite movement also the coverage area changes its position continually creating a coverage belt.

This paper discusses the single LEO satellite coverage aspects as an overture to the global coverage. Some characteristics of LEO satellites are given followed by coverage geometry. Finally the results of coverage belt simulation under different attitudes are presented.

II. LEO SATELLITES COVERAGE

The position of the satellite within its orbit considered from the ground station point of view is defined by *Azimuth* (*Az*) and *Elevation* (ε_0) angles. The azimuth is the angle of the direction of the satellite, measured in the horizon plane from geographical north in clockwise direction. The range of azimuth is 0° to 360°. The elevation is the angle between a satellite and the observer's (ground station's) horizon plane. The range of elevation is 0° to 90°.

The satellite's coverage area on the Earth depends on orbital parameters. The coverage area of a single satellite is a circular area (Fig. 1) on the Earth surface from where the satellite can be seen under an elevation angle. The largest coverage area is achieved under elevation of 0º, but in order to avoid obstacles caused by natural barriers at too low elevation, usually for the link budget calculations it is determined the minimal elevation angle which ranges on (2-10)º. For simulation purposes of the coverage it is considered the elevations up to 10º in steps of 2º.

Ground station (GS – defined as LUTKOS) can communicate with LEO (Low Earth Orbiting) satellite only when the ground station is under coverage area (satellite footprint). In Fig. 1, only communication base stations (BC1 and BC4) can communicate with satellite being within its coverage area and two others (BC2and BC3) have no communication as they are out of the coverage area. The duration of the visibility and consequently the communication duration vary for each LEO satellite pass over the ground station. Along with a satellite, the coverage area (footprint) moves also, leaving the ground station out of the footprint and consequently loosing the communication.

Usually, the satellite's coverage area is expressed (in percentage) as a fraction of the Earth's area. LEO satellites are known for the smallest coverage area, of few percents (3% - 8%) of Earth's surface. In Fig. 2, for synchronized orbits at attitudes of 600km, for different inclination (few orbits) at elevation of 10 º applying satellite orbit analysis software it is presented the case of LEO coverage area. Small circles represent LEO coverage area on the Earth's surface.



Fig. 1. The LEO coverage area.

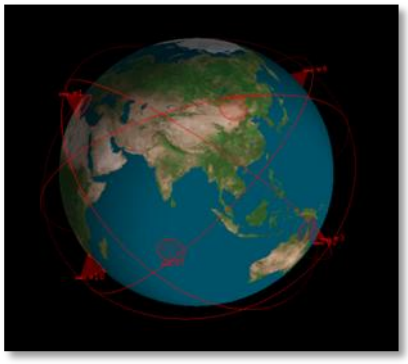


Fig. 2. LEO coverage area. (Source: NOAA, 2009).

The orbital plane is in principle fixed and defined by orbital parameters. Thus, as satellite orbits around the Earth, consequently the coverage area also moves virtually vertically around the Earth, as presented in Fig. 3.a. On the other hand, because of Earth's rotation around its N-S axis for angle β, as depicted in Fig. 3.b, the ground station changes the position relatively to orbital plane, so the pointing (azimuth and elevation) from the ground station to the satellite is not identical for the both satellite passes (see a) and b) in Fig. 3) [13]. Thus, as satellite orbits the Earth and at the same time the Earth rotates around its axis, consequently the coverage area also virtually moves horizontally. The Earth's area swept by LEO satellite's coverage area is known as coverage belt as given in Fig. 4. The wideness of the appropriate coverage belt depends on satellite altitude. Satellites under the same attitude under different inclination make different belts, enabling global coverage.

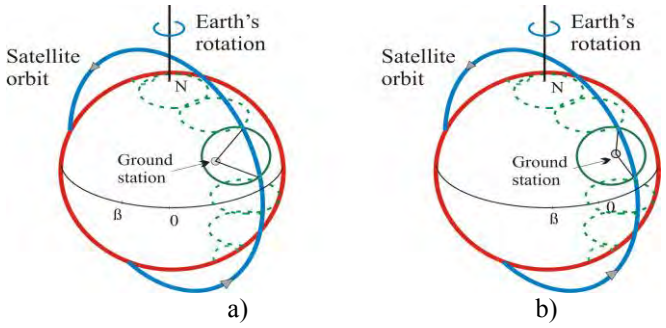


Fig. 3. Satellite pass for an Earth rotation angle of β per orbit a) first pass and b) second pass.

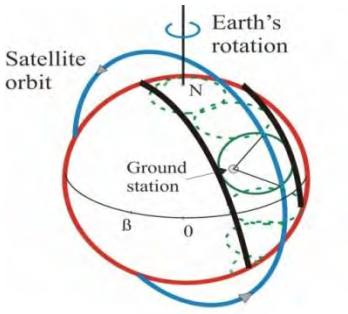


Fig. 4. Coverage belt.

III. LEO COVERAGE GEOMETRY

The basic geometry between a satellite and ground station is depicted in Fig. 5. The points indicate the satellite (SAT), ground station (P), and then the third is the Earth's center. The line passing at point P represents horizon plane. The subsatellite point is indicated by T. Two sides of this triangle are usually known (the distance from the ground station to the Earth's center, $R_E = 6378x10^3$ m and the distance from the satellite to Earth's center-orbital radius). There are four variables in this triangle: ε_0 - is elevation angle, α_0 - is nadir angle, β_0 - is central angle and d is slant range. As soon as two quantities are known, the others can be found with the following equations [14]:

$$\varepsilon_0 + \alpha_0 + \beta_0 = 90 \quad (1)$$

$$d\cos\varepsilon_0 = r\sin\beta_0 \quad (2)$$

$$d\sin\alpha_0 = R_e\sin\beta_0 \quad (3)$$

The most needed parameter is the slant range d (distance from the ground station to the satellite). This parameter will be used during the link budget calculation, and it is expressed through elevation angle ε_0. Applying cosines law for triangle at Fig. 5 yields out:

$$r^2 = R_e^2 + d^2 - 2R_e d\cos(90+\varepsilon_0) \quad (4)$$

Solving (4) by d, substituting, $r = H + R_e$ at (5) and applying (1), (2), (3) finally we will get the slant range as function of elevation angle ε_0 [14].

$$d(\varepsilon_0) = R_e\left[\sqrt{\left(\frac{H+R_e}{R_e}\right)^2 - \cos^2\varepsilon_0} - \sin\varepsilon_0\right] \quad (5)$$

H is the satellite attitude above the Earth's surface. Transforming Fig. 5 from the coverage point of view it looks like in Fig. 6.

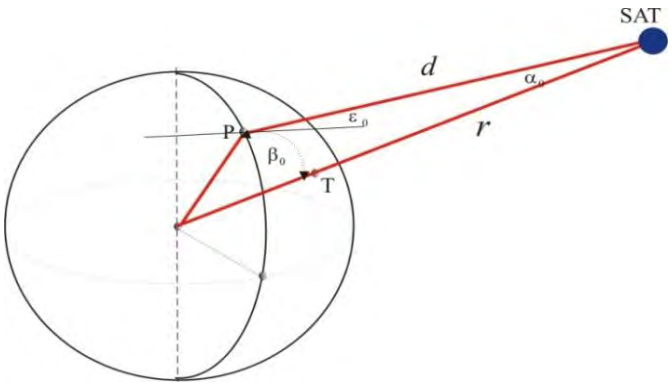


Fig. 5. Ground station geometry.

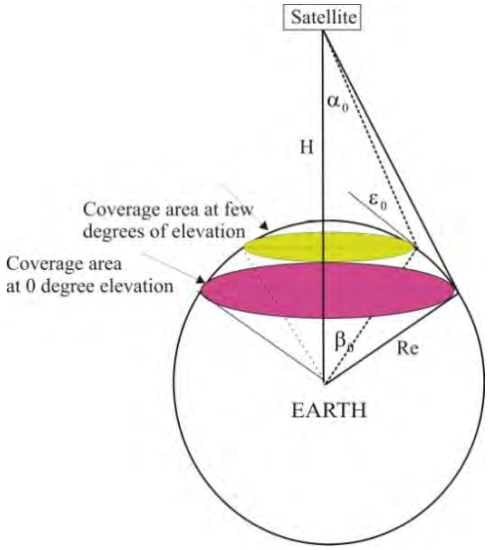


Fig. 6. Coverage geometry.

In Fig. 6, there are two triangles. The larger one represents the case of full coverage under elevation of 0º. The distance d under elevation of 0º represents the radius of the largest circle as coverage area. The wideness of the coverage belt is twice of the largest radius, expressed as follows and presented in Fig. 7.

$$d_{(\varepsilon_0=0)} = d_{\max} = R_e\left[\sqrt{\left(\frac{H+R_e}{R_e}\right)^2} - 1\right] \quad (6)$$

$$D_{BELT} = 2d_{max} \quad (7)$$

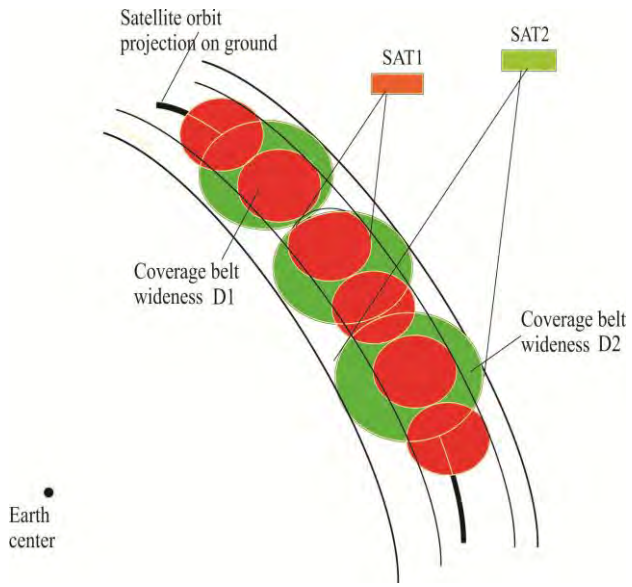


Fig.7. Coverage belt of different satellite altitudes.

IV. LEO Coverage Belt Results

Based on (6) and (7) it is obvious that the satellite coverage belt strongly depends on elevation angle. The largest one is under elevation of 0°. To conclude about the coverage belt variation for low orbiting satellites, the simulation for attitudes from 600km up to 1200km is further discussed.

For attitudes of H=600km, 800km, 1000km and 1200km which are typical low orbit attitudes it is simulated and calculated the coverage belt wideness for elevation of (0-10)° by steps of 2°, and results are presented in Table I and Fig. 8. Table I and Fig. 8 confirm the decrease of coverage belt wideness as elevation angle increases for the already defined attitude H, and the increase of the coverage belt wideness as attitude H increases keeping the fixed elevation.

TABLE I: Coverage belt wideness.

Orbital Attitude [km]	**H 600 [km]**	**H 800 [km]**	**H 1000 [km]**	**H 1200 [km]**
Elevation(ε_0)	***D*** [km]	***D*** [km]	***D*** [km]	***D*** [km]
0°	5633.0	6579.0	7416.0	8177.8
2°	5215.2	6157.2	6991.4	7751.2
4°	4824.4	5760.2	6590.6	7347.2
6°	4463.0	5386.8	6210.0	6959.8
8°	4141.6	5048.8	5859.2	6601.2
10°	3857.4	4745.0	5541.8	6273.6

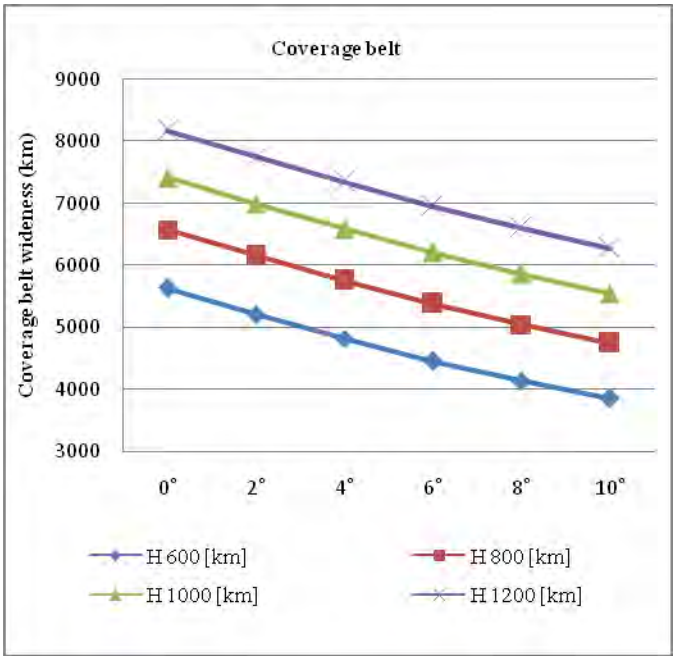


Fig.8. Coverage area variation for different attitudes at low elevation.

Conclusion

The satellite's coverage belt on the Earth depends strongly on orbit attitude and elevation angle. LEO satellites organized in constellations, each one of them with its own coverage belt, act as a convenient solution for real time global coverage.

The widest coverage belt is achieved under elevation of 0º, but in order to avoid obstacles caused by natural barriers at too low elevation, usually for the link budget calculations it is determined the minimal elevation angle which ranges on (2-10)º. For the higher elevation consequently the narrower coverage belt is generated.

Through simulation for typical LEO attitudes on range of (600-1200) km it is confirmed that the coverage belt wideness covered by satellites at respective attitudes ranges from 5633 km to 8177km.

References

[1] G. Maral and M. Bousquet, "Satellite communication systems", John Willey & Sons, Ltd, Chichester, England, 2002.

[2] M. Richharia, "Satellite communication systems", McGraw Hill, New York, 1999.

[3] M. Walt, "Introduction to Geomagnetically Trapped Radiation", Cambridge University Press, 2005.

[4] C.Qian, S. Zhang, W. Zhou, "Traffic-based dynamic beam coverage adjustment in satellite mobile communication" 6th International Conference on Wireless Communications and Signal Processing (WCSP), Hefei, China, 2014, pp. 1-6.

[5] S. Cakaj, M. Fischer, A. L. Schotlz, "Sun synchronization of Low Earth Orbits (LEO) through inclination angle", 28th IASTED International Conference on Modelling, Identification and Control, MIC 2009, Feb. 16 –18, 2009, Innsbruck, Austria, pp. 155-161.

[6] A. Botta, A. Pescape, "New generation satellite broadband Internet service: should ADSL and 3G worry", TMA 2013, co-lacted with IEEE INFOCOM 2013, April 2013, Turin, Italy.

[7] S. Olariu and P. Todorova, "QoS on LEO satellites" *IEEE Potentials,* Volume 23, Issue 3, pp. 11-17, 2004.

[8] I. F. Alyildiz, H. Uzunalioglu, M. D. Bender,Turkiye, " Handover management in ow Earth Orbit (LEO) satellite networks" Mobile Networks and Applications 4 (1999) 301 -310.

[9] P. Papapetrou, S. Karapantazis, F.N. Pavlidou, " Handover Policies in LEO Systems with Satellite Diversity", International Conference on Advanced Satellite Mobile Systems (ASMS 2003), 10-11 July, 2003, Frascati, Italy.

[10] Y. Seyedi, S. M. Safavi, "On the Analysis of Random Coverage Time in Mobile LEO Satellite Communications", *Communications Letters*, IEEE, Volume 16, Issue 5, May 2012.

[11] S. Cakaj, W. Keim, K. Malaric, "Intermodulation by Uplink Signal at Low Earth Orbiting Satellite Ground Station", 18th International Conference on Applied Electromagnetics and Communications, ICECom, IEEE, 12-14 October 2005, Dubrovnik, Croatia, pp. 193 - 196.

[12] S. Cakaj, "Rain Attenuation Impact on Performance of Satellite Ground Stations for Low Earth Orbiting (LEO) Satellites in Europe", *International Journal of Communications, Networks and System Sciences (IJCNS),* Volume 2, Issue 6, 2009, USA, pp. 480-485.

[13] S. Cakaj, "Practical Horizon Plane and Communication Duration for Low Earth Orbiting (LEO) Satellite Ground Stations", *WSEAS Journal: Transactions on Communications,* Volume 8, Issue 4, April 2009, pp. 373-383.

[14] G.D. Gordon, W.L. Morgan, „Principles of communication satellites", John Wiley & sons, Inc. 1993.

The investigation of the effect of the carrier frequency offset(CFO) in SC-FDMA System

Necmi TA¯PINAR*, Merve BALKK**
*Department of Electrical and Electronic Engineering, Erciyes University, Kayseri, Turkey
**Institute of Natural and Applied Science, Erciyes University, Kayseri, Turkey
taspinar@erciyes.edu.tr, mervebalki@gmail.com

Abstract- In this paper, the effect of the carrier frequency offset (CFO) on system performance was investigated by computer simulations in single carrier Frequency Division Multiple Access (SC-FDMA) system. By computer simulations, figures of signal-to-noise ratio (SNR) versus symbol error rate (SER) of the system with and without CFO are obtained. Besides, the figure of SNR versus SER is obtained for different CFO values. From these results, it is seen that when the value of CFO increases, the performance of the system decreases. Moreover, the figures of SNR versus SER is obtained for different FFT sizes and different modulation levels for the same CFO value. It is seen that when the size of FFT increases, the system yields better performance. When the level of modulation increases, the performance of the system decreases. In addition, the estimation of CFO is performed by computer simulations. In the simulations, the estimation algorithm based on cyclic prefix is used. The figures of mean square error (MSE) versus SNR related with estimation and different modulation levels are obtained. It is seen that MSE decreases when SNR increases. By using CFO values obtained with estimations, correction process is also performed to obtain performance values of the system.

I. INTRODUCTION

Single Carrier Frequency Division Multiple Access (SC-FDMA) is a promising technique for high data rate uplink communication and has been adopted by third Generation Partnership Project (3GPP) for next generation cellular system, called Long-Term Evolution (LTE).SC-FDMA is a modified form of Orthogonal Frequency Division Multiple Access (OFDMA) with similar throughput performance and complexity. Mostly OFDMA is used in high speed data transmission due to its strength against frequency selective channels [1]. But OFDMA system has high peak-to-average power ratio (PAPR). Therefore SC-FDMA is used as an alternative to OFDMA [2].Thus SC-FDMA systems are more convenient than OFDMA in that it has lower PAPR [3]. But SC-FDMA system is very sensitive to carrier frequency offset (CFO). CFO destroys the orthogonality between subcarriers. Thus the performance of the system deteriorates due to effect of CFO [4].Therefore frequency offset estimation and correction at the receiver side is very important for the system performance.

There are some studies on the estimation of CFO in SC-FDMA system in the literature [4-10].In [4], it was compared difference between SC-FDMA and OFDMA. CFO was corrected by adopting multistage interference cancellation (IC) techniques with the estimated CFO values. Thus, SC-FDMA more convenient than OFDMA in that system performance. In [5], it was presented joint low-complexity regularized zero-forcing (JLRZF) equalizer for multiple-input multiple-output (MIMO) SC-FDMA systems to reduce the complexity and to avoid from the noise enhancement of conventional ZF equalizer. In [6], it was investigated the effect of CFO on the single carrier interleaved frequency division multiple access (SC-IFDMA) system and was proposed low complexity time domain CFO compensation (TD-CC) scheme. In [7], it was proposed a combined minimum mean square frequency domain equalization (MMSE-FDE) and interference cancellation scheme to eliminate CFO. The performance of SC-FDMA transmission over frequency-selective fading channels was investigated in [8]. In [9], it was suggested a semi-blind CFO estimation scheme for SC-FDMA system with interleaved allocation using parallel factor analysis. In [10], it was proposed a method to estimate the frequency offset in SC-FDMA systems. In this method, it was assigned a part of each SC-FDMA symbol to pilot signals and developed an estimation method based on maximum a posteriori (MAP) criterion.

In this paper, CFO values are estimated by the estimation method based on cyclic prefix (CP) by computer simulations for SC-FDMA system as different from the studies in the literature. Negative influence of CFO was observed on the system performance and correction process was needed. Thus correction process that was used in OFDM system is arranged according to SC-FDMA system and it was applied to SC-FDMA system. In addition, the effect of the carrier frequency offset on system performance was investigated by computer simulations in SC-FDMA system for different FFT sizes, different modulation levels and different CFO values by considering the changes of symbol error rate and minimum square error versus signal-to-noise ratio. This paper is arranged as follows: The structure of SC-FDMA system is presented in Section II. In Section III the frequency offset method is explained. Simulation results are presented in Section IV. Section V includes conclusions.

II. SYSTEM MODEL

Fig. 1 illustrates a basic single carrier Frequency Division Multiple Access (SC-FDMA) system. In the transmitter part, one block of data is modulated firstly. The data block turn into M complex modulation symbols after from modulation process. The M point discrete Fourier transform (DFT) produce M frequency domain symbols. These M frequency domain symbols are assigned to M adjacent subcarriers out of N available subcarriers at the subcarrier mapping block. The inverse

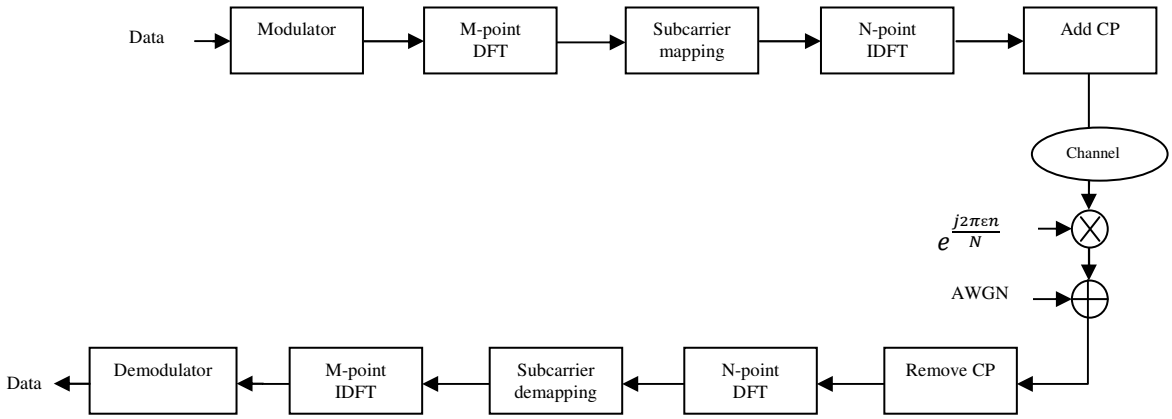


Fig.1. Structure of SC-FDMA System.

discrete Fourier transform (IDFT) generates time domain symbols of sizes *N* and cyclic prefix block inserts a set of symbols. SC-FDMA system has *K* active users and *N* subcarriers. Each of user has *Q=N/K* subcarriers. The signal is transmitted over the channel. At the receiver, firstly cyclic prefix block is removed. Then *N*-point DFT is performed to produce *N* frequency domain symbols. After that subcarrier demapping process is made. Then *M* point inverse discrete Fourier transform is performed to generate time domain *M* modulation symbols. These symbols are demodulated to obtain data block.

In the transmitter, the data transmitted for kth user is

$$x^{(k)} = [x_0^{(k)}, x_1^{(k)}, \dots, x_{Q-1}^{(k)}]^{\mathrm{T}} \quad (1)$$

where *x* is data and $X^{(k)}$ is output of modulator. We can write output of *N* point IDFT block for kth user as

$$s^{(k)} = D_M M I_N X^{(k)} \quad (2)$$

where D_M, *M*, I_N represent *M* point DFT size, subcarrier mapping process, *N* point IDFT, respectively. Cyclic prefix (CP) with length of L_{CP} is added to the transmitted sequence. Added CP eliminates the effect of inter symbol interference (ISI) when these symbols are passing from channel. The L_{CP} should be longer than maximum channel delay spread. The channel is time fixed for each user. So Channel Impulse Response (CIR) for kth user represents by $h^{(k)}$. After the transmitted signal passing from the channel, CFO and additive white Gaussian noise (AWGN) are added to the signal. At the receiver side, the summed signal that comes from different users after removing CP, could be written as

$$r(n) = \sum_{k=1}^{K} e^{\frac{j2\pi\varepsilon_k n}{N}} (s^{(k)} \otimes h^{(k)}) + w \quad (3)$$

where ε_k represents frequency offset for the kth user,$\otimes$ denotes convolution and *w* is AWGN for the kth user. Then correction process of the carrier frequency offset is performed. For this purpose, the following equation can be used

$$Y^{(k)} = \frac{1}{N}\sum_{n=0}^{N-1} r(n).\, e^{-\frac{j2\pi\varepsilon_k n}{N}} \quad (4)$$

Where $Y^{(k)}$ is corrected signal for kth user. Correction is performed after process of CFO estimation. Then it is performed subcarrier demapping, IFFT and demodulation [11].

III. CFO ESTIMATION METHOD

In the CFO estimation method, cyclic prefix is used for the CFO estimation. In this technique, effect of channel is ignored. So the kth SC-FDMA symbol with CFO can be written as

$$y^{(k)}(n) = s^{(k)}(n) e^{\frac{j2\pi\varepsilon n}{N}} \quad (5)$$

Considering CP process, the formula can be written as:

$$y^{(k)}(n+N) = s^{(k)}(n+N) e^{\frac{j2\pi\varepsilon(n+N)}{N}} \quad (6)$$

$$y^{(k)}(n+N) = s^{(k)}(n+N) e^{\frac{j2\pi\varepsilon n}{N} + j2\pi\varepsilon} \quad (7)$$

We can observe difference between CP and without CP when comparing"(5)" and "(7)". This difference is phase difference and equals to $2^{-}K$ In general, the amount of CFO is found from phase difference using "(8)" as follows:

$$\acute{\varepsilon} = \frac{1}{2\pi} arg\left\{\sum_{n=\mathrm{Lcp}}^{-1} y^{(k)^*}(n).\, y^{(k)}(n+N)\right\},\ \mathrm{n}= -1, -2, \dots, -\mathrm{L_{CP}} \quad (8)$$

where L_{CP} is the length of cyclic prefix, έ is estimated CFO value and * denotes conjugate of symbols. Argument operator that was used in "(8)"means $tan^{-1}(\,)$ operator.CP based estimation method is useful method for CFO estimation. This method has low mean square error (MSE).Finally we can make correction process from estimated CFO values by using "(4)" [12].

IV. SIMULATION RESULTS

The effect of the carrier frequency offset on system performance was investigated by computer simulations in SC-FDMA system for different FFT sizes, different modulation levels and different CFO values by considering the changes of sembol error rate and minimum square error versus signal-to-noise ratio.

Fig. 2 shows symbol error rate (SER) performance of SC-FDMA system for different CFO values. In the simulations, FFT size is used as 512, modulation level is taken as QPSK, subcarrier mapping method is chosen as localized mapping, and CFO values are taken as 0, 0.1, 0.2, 0.3, 0.4 and 0.5. It is seen from Fig. 2 that SER values decrease while SNR values increase. It is also seen from Fig. 2 that the system performance deteriorates while CFO value increases.

Fig.3 shows SER performance comparison of SC-FDMA system under the CFO value of 0.25 and without CFO. In the simulations, FFT size is used as 512, modulation level is taken as QPSK, subcarrier mapping method is chosen as localized mapping.

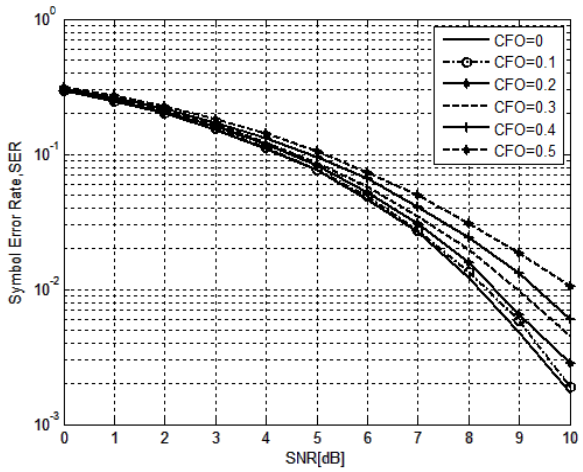


Fig. 2. SER performance of SC-FDMA system for different CFO values .

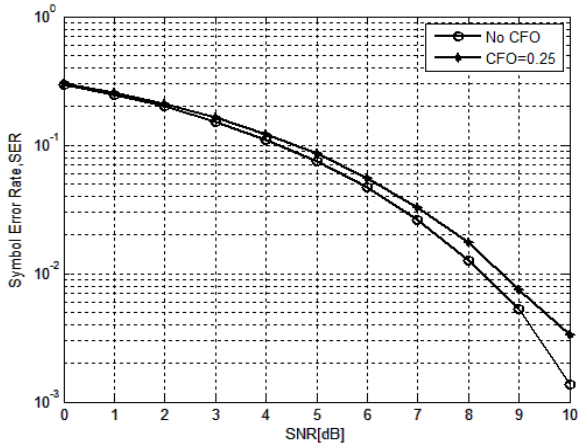


Fig.3. SER performance comparison of SC-FDMA system for including CFO and not including CFO.

It is seen from Fig. 3 that CFO affects the system negatively. It is also seen from Fig. 3 that SER values decrease while SNR values increase.

SER performance of SC-FDMA system with different FFT sizes is given in Fig. 4. In the simulations, modulation level is taken as QPSK, subcarrier mapping method is chosen as localized mapping, CFO value is taken as 0.25 and FFT sizes are taken as 512, 1024 and 2048. It is seen from Fig. 4 that SER values decrease while SNR values increase. It is also seen from Fig. 4 that the system performance decreases while FFT size increases.

Fig. 5 and Fig. 6 are about CFO estimation. These figures are obtained by using CP based estimation method. Fig. 5 shows estimated CFO values versus SNR values with CP based estimation method. In the simulations, FFT size is used as 512, modulation level is taken as QPSK, subcarrier mapping method is chosen as localized mapping, and CFO value is taken as 0.25. It is seen from Fig. 5 that the difference between the actual CFO value and estimated CFO. Estimated CFO value appears to actual value while SNR value increases.

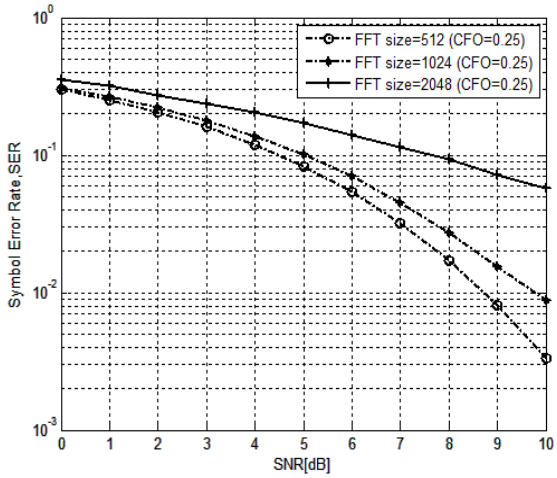


Fig.4. SER performance of SC-FDMA system with different FFT sizes.

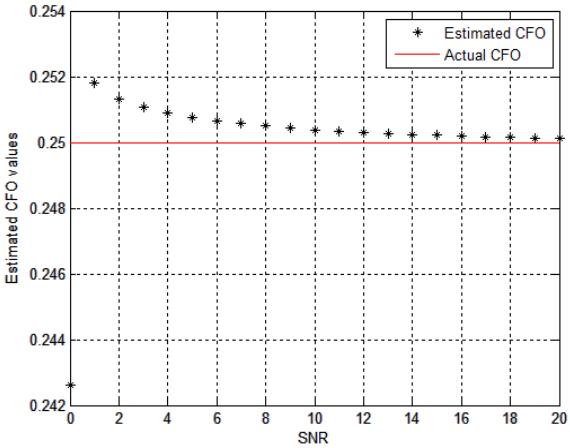


Fig. 5. Estimated CFO values with CP based estimation method

After CFO estimation, mean square error (MSE) is calculated by computer simulations and the results are given in Fig. 6. In the simulations, FFT size is used as 512, modulation level is taken as BPSK and QPSK, subcarrier mapping method is chosen as localized mapping, and CFO value is taken as 0.25. It is seen from Fig. 6 that MSE value decreases while SNR value increases. Besides BPSK modulation yields better performance than QPSK modulation.

SER performances of SC-FDMA system under CFO and corrected by using "(4)" are given in Fig. 7. In the simulations, FFT size is used as 512, modulation level is taken as QPSK, subcarrier mapping method is chosen as localized mapping, and CFO value is taken as 0.25. It is seen from Fig. 7 that corrected CFO values are near to the real CFO values. It is also seen from Fig. 7 that SER values decrease while SNR values increase.

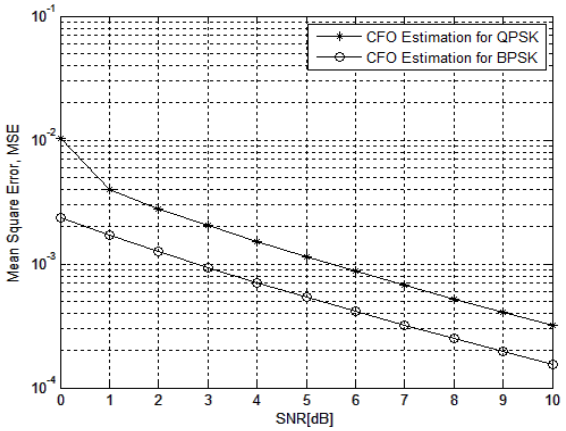


Fig. 6.MSE performance of CP based estimation method for BPSK and QPSK modulations.

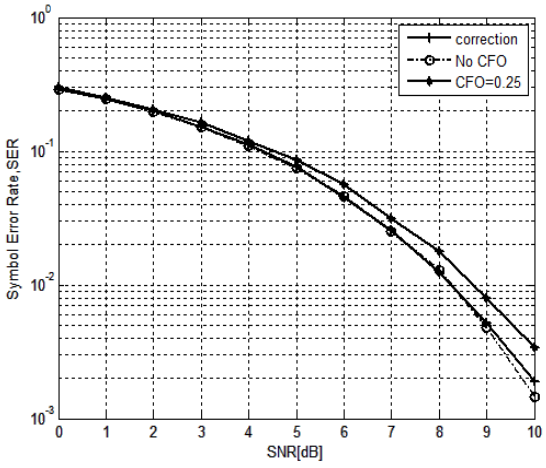


Fig.7. SER performance of SC-FDMA system under CFO and corrected by CP method.

V. CONLUSION

In this paper, the effect of the carrier frequency offset (CFO) on system performance was investigated by computer simulations in single carrier Frequency Division Multiple Access (SC-FDMA) system to see the effects of FFT size, modulation levels and CFO values. It is seen from computer simulations that the increase in FFT size, modulation level and CFO value decreases the system performance. CFO is determined while using CP based estimation method and it has low error. Finally, CFO is corrected and the system is fixed to the state without CFO.

ACKNOWLEDGMENT

This work was supported by Erciyes University Scientific Research Projects Coordination Unit (Project No:FYL-2014-5573).

REFERENCES

[1] H. G. Myung, J. Lim and D. J. Goodman, "Single carrier FDMA for uplink wireless transmission," IEEE Vehicular Technology Magazine, vol. 1, pp. 30-38, Sept. 2006.

[2] D. Falconer, R. Dinis, C.T. Lam and M. Sabbaghian, "A multiple access scheme for uplink broadband wireless systems," in Proc. of IEEE Global Commun. Conf. (GLOBECOM 2004), pp. 3808-3812, Dallas, Tx, Dec. 2004.

[3] H. G. Myung, J. Lim and D. J. Goodman, "Peak-to-average power ratio of single carrier FDMA signals with pulse shaping," in Proc. of 17th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'06),pp. 1-5, Helsinki, Sept. 2006.

[4] K. Raghunath and A. Chockalingam, "SC-FDMA versus OFDMA: sensitivity to large carrier frequency and timing offsets on uplink," in Proc. of IEEE Global Commun. Conf. (GLOBECOM 2009), pp. 1-6, Dec. 2009.

[5] F.S. Al-Kamali, M.I. Dessouky, B.M. Sallam, F. Shawki and F.E. Abd El-Samie, "Uplink single-carrier frequency division multiple access system with joint equalization and carrier frequency offsets compensation," IEEE Trans. on Wireless Communications, vol. 11, pp. 869-873, March 2012.

[6] Y Zhu and K. Letaief, "CFO estimation and compensation in single carrier interleaved FDMA systems," in Proc. IEEE Global Commun. Conf.(GLOBECOM 2009), ", pp. 1-5, Dec. 2009.

[7] G. Chen, Y Zhu, and K. Letaief, "Combined MMSE-FDE and interference cancellation for uplink SC-FDMA with carrier frequency offsets," in Proc. IEEE International Conference on Communications", pp. 1-5, May 2010.

[8] H. Wang, X. You, B. Jiang, and X. Gao, "Performance analysis of frequency domain equalization in SC-FDMA systems," in Proc. of IEEE International Conference on Communications (ICC'2008),pp. 4342-4347, May 2008.

[9] N. Sidiropoulos, G. Giannakis, and R. Bro, "Blind PARAFAC receivers for DS-CDMA systems," IEEE Trans. on Signal Processing, vol. 48, no. 3, pp. 810-823, Mar. 2000.

[10] K.Shamaei and M.Sabbaghian,"Frequency offset estimation in SC-FDMA systems,"in Proc. of IEEE Global Telecommunication Conference, pp. 216-220, Nov. 2012.

[11] K.Selvaraj, M.Alice Margaret and P.Ganesh Kumar, "Low Complexity CFO Compensation Scheme for SC-FDMA Uplink System," International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, pp. 2411-2415, March 2014.

[12] G. Abshir, M. M. Abdullahi, M. A. Samad," A Comparative Study of Carrier Frequency Offset (CFO) Estimation Techniques for OFDM Systems," IOSR Journal of Electronics and Communication Engineering, vol. 9, no.4, pp. 1-6, July-August 2014.

Performance analysis of Low Density Parity Check Codes implemented in second generations of Digital Video Broadcasting standards

Gr. Y. Mihaylov*, T. B. Iliev*, E. P. Ivanova*, I. S. Stoyanov** and L. Iliev**
*University of Ruse, Department of Telecommunication, Ruse, Bulgaria
**University of Ruse, Department of Electrical and Electronic Engineering, Ruse, Bulgaria
e-mail addresses: gmihaylov@uni-ruse.bg, tiliev@uni-ruse.bg, epivanova@uni-ruse.bg, stoyanov@uni-ruse.bg, lailiev@uni-ruse.bg

Abstract - The newly developed second generations standards for digital video broadcasting (DVB-T2, DVB-S2) emerges as a significant upgrade over its first generation predecessor. Low-Density Parity-Check (LDPC) codes received a lot of attention by their superb error-correcting capability and have been adopted as main error correct coding scheme by second generation Digital Video Broadcasting standards. The goal of this paper is to present the main constructive principle of low-density parity check codes, as well as a design methodology, that gives good performance for many communication systems like latest DVB satellite communications standard. In this paper we analyze the structure of LDPC codes and the iterative algorithms that are used to decode them. An analysis and simulation of the LDPC codes has been conducted using various code rates. The simulation results show that LDPC codes are a powerful error correcting coding technique under SNR environments. It has achieved near Shannon capacity. However, there are many factors needed to be considered in the LDPC code design.

I. INTRODUCTION

DVB-T is the most widely used standard for broadcast of digital television. With its introduction in 1997, it is used in more than 70 countries. Almost all of these countries adopted the new standard for digital terrestrial television broadcasting – DVB-T2. [1]

DVB-T2 is the best technology for digital television broadcasting, which offers better signal robustness, flexibility and more than 50% more efficiency, compared with other digital terrestrial television broadcast systems. In Single frequency networks (*SFN*) it can be achieved even more efficiency, because of guard intervals, used in these networks. DVB-T2 supports SD, HD, UHD, mobile television, radio broadcasting and any combinations thereof, furthermore Internet access can be offered as well.

Like its predecessor, DVB-T2 uses OFDM (orthogonal frequency division multiplex) modulation with large number of subcarriers. Thus, signal robustness can be achieved and various transmission modes are available, which makes this standard very flexible. In DVB-T2 are used the same coding and error correction as in DVB-S2 and DVB-C2 standards – Low Density Parity Check (*LDPC*) in combination with BCH (Bose-Chaudhuri-Hocquengham). It can be adjusted any number of carriers, size of guard intervals and pilot signals, in order to avoid overflow. Thus, signal transmission is optimized for every broadcast channel. [2]

T2-Lite is the first additional transmission profile, introduced in 2011. It's intended for mobile and portable transmission/reception of digital television and to reduce implementation costs. This profile is defined as a part of DVB-T2 standard with two additional LDPC code rates. Since in T2-Lite only elements, relevant for mobile and portable reception are included, data rate is restricted to 4 Mbit/s per PLP, the implementation complexity is reduced by 50%. FEF (Future Extension Frames) mechanisms allows T2-Lite to be broadcasted in one channel along with DVB-T2 signal, even when the two profiles are with different FFT sizes or guard intervals. [3]

Additional new technologies, used in DVB-T2 are:

- ***Multiple Physical Layer Pipes*** – allow separate adjustment of the signal robustness for every service. Thus, easily can be achieved the conditions for signal reception (e.g. indoor- or roof antenna). This technology allows receivers to save energy, by decoding only one (desired) service, instead to decode the whole service multiplex.
- ***Constellation Rotation*** – provides additional signal robustness in modulations with lower constellation order.
- ***Extended interleaving*** – includes bit-, cell-, time- and frequency interleaving.
- ***Future Extension Frames (FEF)*** – this technology allows DVB-T2 standard to be compatibly enhanced in the future.

In summary, DVB-T2 offers much higher broadcasting data rates than DVB-T standard, as well as offers great signal robustness. [4]

LDPC codes, discovered by Gallager in 1962 [5], were rediscovered and shown to form a class of Shannon-limit-approaching codes in the late 1990 [6]. These codes, decoded with iterative decoding based on belief propagation, such as the sum-product algorithm (*SPA*) [7,

8], achieve amazingly good error performance. Ever since their rediscovery, design, construction, decoding, efficient encoding, performance analysis, and applications of these codes in digital communication and storage systems have become the focal points of research. Various methods for constructing LDPC codes have been proposed. Based on the methods of construction, LDPC codes can be classified into two general categories: 1) random (or random-like) codes generated by computer search based on certain design guidelines and required structural properties of their Tanner graphs [9], such as the girth and degree distributions and 2) structured codes constructed based on algebraic and combinatorial methods, such as those given in [7, 8], and many others.

The rest of the paper is organized as follows. Section 2 gives a brief overview of LDPC codes. Section 3 describes the standardized LDPC code. Section 4 introduces message passing decoding algorithm. Section 5 presents the conducted simulation and the received results. Section 6 concludes the paper.

II. LDPC BASICS

Low-Density Parity-Check (*LDPC*) codes and turbo codes are among the known codes nearing the Shannon limit. These codes can achieve very low bit error rates for low signal-to-noise ratio (*SNR*) applications [10]. Compared to the decoding algorithm of Turbo codes, LDPC decoding algorithm has more parallelization, low implementation complexity, low decoding latency and no error-floors at high SNRs as well. LDPC codes are considered for all next generation communication standards.

The performance of Low Density Parity Check (*LDPC*) codes, combined with their relatively simple decoding algorithm, make these codes very attractive for next generation of satellite and digital radio transmission systems. [9]

The LDPC codes are linear block codes, whose codewords satisfy a set of linear parity-check constraints [1]. These constraints are typically defined by an m-by-n parity-check matrix H. The m rows specify each of the m constraints (the number of parity checks) and n represents the length of a codeword. H is also characterized by W_r and W_c, which represent the number of 1's in the rows and columns, respectively.

A LDPC code can be represented by a bipartite graph, which consists of two types of nodes – Variable Nodes (*VN*) and Check Nodes (*CN*). Check node i is connected to variable node j, whenever h_{ij} of H is non-zero.

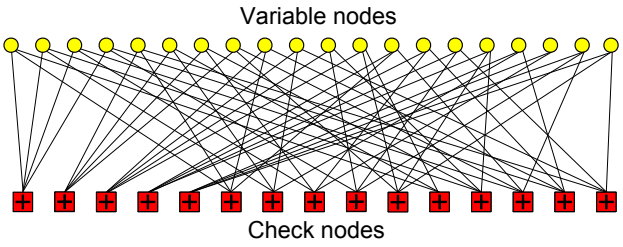


Figure 1. Bipartite graph for a (3,4) regular LDPC code with parity check matrix H

III. STANDARDIZED LDPC CODES

All current standardized LDPC codes combine a common set of features, which made feasible implementation possible in the first place: Accumulator-based matrices, allowing linear encoding complexity, structured matrices to be mapped on partly parallel architectures and permuted identity sub-matrices to ease network implementation.

Satellite video broadcasting standard DVB-S2 [2] was designed for an exceptional error performance at very low SNR rates (up to FER$\leq 10^{-7}$ at -2.35dB E_S/N_0). Thus, the specified LDPC codes use a large block length of 64800 bits with 11 different code rates, ranging from 1/4 to 9/10. The result is large storage requirements for up to 285000 messages and demands high code rate flexibility at the same time to support all specified node degrees [11].

In terms of modulation, rates from BPSK up to 64QAM are supported. Coding via convolution codes is mandatory with optional support for LDPC and turbo coding. Furthermore, more advanced method of turbo coding known as Shorted Block Turbo Code (*SBTC*) is expected to be supported as well.

IV. MESSAGE PASSING DECODING ALGORITHMS

It was stated in previous section, the principal advantage of low-density parity check codes is that these codes can be decoded using an iterative decoding algorithm, whose complexity grows linearly with the block length of the code. Belief propagation is one instance of a broad class of message passing algorithms on graphs as discussed in [7, 8]. Similar decoding algorithms, for binary erasure channel and the binary symmetric channel, are discussed in [12]. All message passing algorithms must, however, respect the following rule, which was introduced with turbo codes.

Rule (Extrinsic Information Principle) *A message sent from a node* ***n*** *along an edge* ***e*** *cannot depend on any message previously received on edge* ***e***.

The codeword x is transmitted across the additive white Gaussian noise (*AWGN*) channel, using BPSK modulation. The result is received sequence $r=(2x-1)+n$. Optimal decoder for the AWGN channel is the maximum a posteriori (*MAP*) decoder that computes the log-likelihood ratio (*LLR*):

$$\Lambda(x_r) = \log\left(\frac{P(x_r = 1|y)}{P(x_r = 0|y)}\right) \quad (1)$$

and makes a decision, comparing this LLR to the threshold zero. Belief propagation on a graph with cycles can closely approximate the MAP algorithm, and we can state the decoding algorithm for LDPC codes on the AWGN channel, using these results.

Let A={−1, +1} represent message alphabet, let $r_i \in R$ represent received symbol at variable node i, and let $\Lambda_i \in R$ represent decision at variable node i. A message

from variable node i to check node j is denoted by $\mu_{i\to j} \in R$ and a message from check node j to variable node i is $\beta_{j\to i} \in R$. Let $C_{j\backslash i}$ be the set of variable nodes, which connect to check node j, excluding variable node i. Similarly, let $V_{i\backslash j}$ be the set of check nodes, which connect to variable node i, excluding check node j. (Fig. 2)

In the first step the algorithm initialize $\Lambda_i = \frac{2}{\sigma^2} r_i$ for each variable node ($\sigma^2=N_0/2$). The variable nodes in step 2 send $\mu_i \to j=\lambda_i$ to each check node $j \in V_i$. In the third step check nodes connected to variable node i and send:

$$\beta_{j\to i} = 2\tanh^{-1}\left(\prod_{l\in C_{j/i}} \tanh\left(\frac{\Lambda_l}{2}\right)\right) \quad (2)$$

Variable nodes connected to check nodes j send:

$$\mu_{i\to j} = \sum_{l\in V_{i/j}} \beta_{l\to i} \quad (3)$$

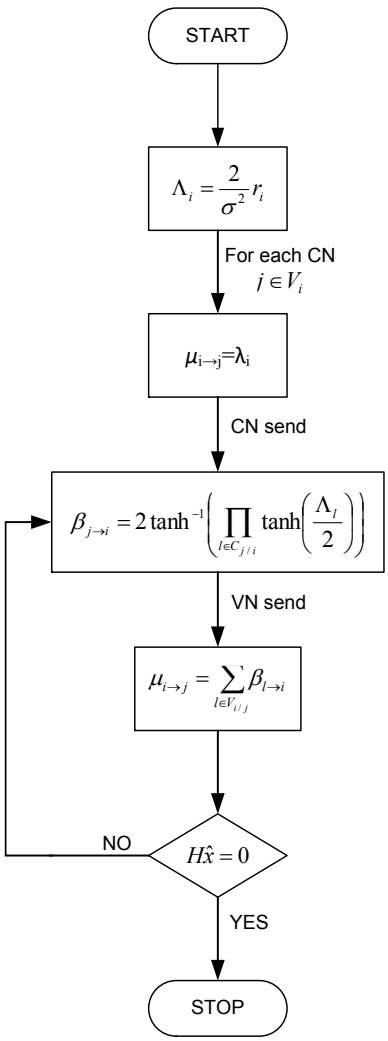


Figure 2. Message passing decoding algorithm

In the final step of the algorithm a fixed number of iterations have been completed or the estimated code word $\hat{x}$ satisfies the syndrome constraint $H\hat{x} = 0$, stop. Otherwise the process returns to step 3.

The check node's rule is fairly complex. But for quantized messages, it is possible to map LLRs to messages in such way, that the check-node rule can be implemented with some extra combinational logic. These decoders provide really nice efficiency with only few bits of accuracy.

V. SIMULATION RESULTS

To estimate the LDPC codes, simulations are performed using binary phase-shift keying (*BPSK*) and transmission over the additive white Gaussian noise (*AWGN*) channel. All simulations have been developed in Matlab. We estimate the bit-error rate (*BER*)/frame-error rate (*FER*) versus E_b/N_0. The E_b represents the energy per bit and the N_0 represents the noise power spectral density ratio.

The performance of LDPC codes for DVB-T2 are shown on the figures given below. In each case the proposed message passing decoding algorithm was used. The decoder stops in case of valid codeword is found or reaching 100 decoding iterations.

Fig. 3 represent the results from conducted simulation of BER rate versus E_b/N_0 of the LDPC code with input frame sizes of 16 200 bits for various code rates. It can be seen that the LDPC code with smaller code rate (R=1/2) has better performance. The gain for a certain value of BER 10^{-4} at the decoder is approximately 2,4 dB.

On Fig. 4 are shown results from same simulation of BER rate versus E_b/N_0 of the LDPC code with input frame sizes of 64 800 bits for various code rates from R=1/2 to R=8/9. Fig. 4 shows that the slope of curves and coding gain are improved by decreasing the code rate of the LDPC code. At 10^{-4} about 2,8 dB coding gain can be obtained by using code R=1/2 instead code rate R=8/9.

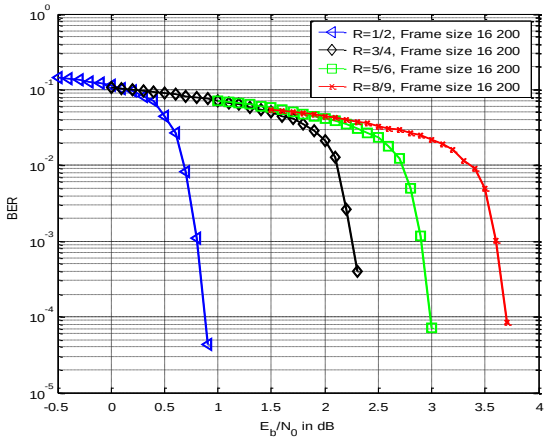


Figure 3. Comparison of the performance of LDPC codes for different code rates, using the same frame sizes = 16 200 bits

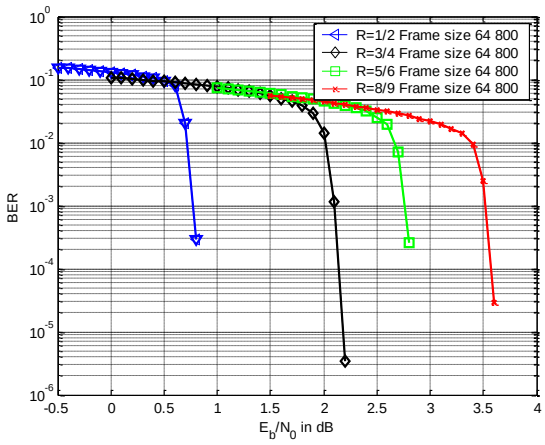


Figure 4. Comparison of the performance of LDPC codes for different code rates, using the same frame sizes = 64 800 bits

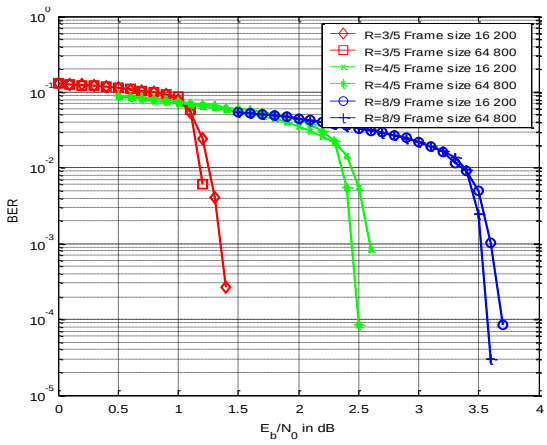


Figure 5. Bit – error performance of the LDPC code in DVB-T2 for various code rates and different frame sizes

Fig. 5 shows the performance of LDPC code for various code rates and two input frame sizes (16 200 bits and 64 800 bits).

From the conducted simulation presented on Fig. 5 we can conclude that for constant value of BER the needed SNR for achieving this probability is lowest for larger input frame size.

VI. CONLUSION

Analysis of techniques, formulated for LDPC codes are generic and can be used for other coding schemes based on iterative decoding methods. Each LDPC code is characterized by a number of fixed parameters that include the following: check node degree, variable node degree and block size. These parameters are used to determine the nodes in the LDPC factor graph and a collection of permissible edges between the nodes.

The conducted simulations show that LDPC codes are a major error correcting coding scheme in AWGN channel. These codes achieved near Shannon capacity. However, there are many factors in LDPC code design, to be considered. One of them is a trade-off between BER and number of iterations used in the decoding algorithm. A great number of iteration will provide a lower BER, but also it will increase the time for decoding process. Second, the frame size impacts on the performance – with larger frame size there is a better performance, but the output delay is again longer. Another factor to be considered is the code rate. The higher code rate needs more bandwidth.

REFERENCES

[1] EBU Tech 3348 r4, Frequency & Network Planning Aspects of DVB-T2, 2014

[2] European Telecommunications Standards Institute (ETSI). Digital Video Broadcasting (DVB) Second generation framing structure for broadband satellite applications; EN 302 307 V1.1.1. www.dvb.org.

[3] A. Sugaris, I Reljin, DVB-T2 technology improvements challenge current strategic planning of ubiquitous media networks. EURASIP Journal on Wireless Communications and Networking 2012, pp. 1-14.

[4] L. Vangelista, N Benvenuto, S Tomasin, et. al., Key technologies for next-generation terrestrial digital television standard DVB-T2, IEEE Commun. Mag. 2009, pp. 146–153.

[5] R. G. Gallager, Low density parity check codes, IRE Trans. Inf. Theory, vol. IT-8, no. 1, pp. 21–28, Jan. 1962.

[6] D. J. C. MacKay and R. M. Neal, Near-Shannon-limit performance of low density parity check codes, Electron. Lett., vol. 32, pp. 1645–1646, Aug. 1996.

[7] F. Guilloud; E. Boutillon; J. L. Danger, λ-min decoding algorithm of regular and irregular LDPC codes, 3rd International Symposium on Turbo Codes & related topics, September 2003.

[8] M. M. Mansour; N. R. Shanbhag, High-throughput LDPC decoders, IEEE Transactions on VLSI Systems, vol. 11, no.6, pp. 976–996, December 2003

[9] T. Iliev, Construction and simulation analysis of Low Density Parity Check Codes. Proceeding of Papers ICEST 2007, Ohrid, Macedonia, pp. 35 – 39, 2007

[10] T. Iliev, G. Hristov, P. Zahariev, M. Iliev, Application and evaluation of the LDPC codes for the next generation communication systems, Novel Algorithms and Techniques in Telecommunications, Automation and Industrial Electronics, Springer, pp 532-536, 2008

[11] F. Tsvetanov, K. Tsvetanov, E. Ivanova, I. Georgieva, Modelling and Simulation of Communication Efficiency in Low-Speed Networks, 22nd Telecommunications Forum Telfor, TELFOR 2014, Proceedings of Papers, pp 178-182

[12] T. Richardson and R. Urbanke, The Capacity of low-density parity check codes under message-passing decoding, IEEE Trans. Inf. Theory, vol. 47, no. 2, pp. 599–618, Feb. 2001

DATA AND IMAGE ANALYSIS

Iterative Denoising of Sparse Images

Isidora Stanković, Irena Orović, Srdjan Stanković, Miloš Daković
Faculty of Electrical Engineering, University of Montenegro
Dzordza Vasingtona bb, 20 000 Podgorica, Montenegro
isidoras@ac.me, irenao@ac.me, srdjan@ac.me, milos@ac.me

***Abstract*—The paper examines an application of the gradient-based algorithm to image denoising with noise values being in the range of the available (non-noisy) pixel values. The analyzed image is considered to be sparse in the 2D-DCT domain. The presented algorithm is a generalization of the previous results on denoising images when the noisy pixels can be detected and eliminated using the L-statistics. The algorithm is based on the recently developed technique used for denoising of one-dimensional signals. A significant advantage of the presented algorithm is that it does not use any a priori knowledge about the positions, values or distribution of the noisy pixels. It is assumed that the positions of noisy pixels cannot be determined using the methods like the L-statistics based ones. Hence, the proposed approach reconstructs the pixels values iteratively using the highest gradient as pixel selection criterion, thus performing blind denoising on a pixel-by-pixel basis. The examples with synthetic two-dimensional signal and a test image are presented. Quality of the image reconstruction is measured using the structural similarity index and the mean absolute error (MAE).**

***Keywords*—compressive sensing, recovery, noise, gradient algorithm, denoising, image processing**

I. Introduction

Compressive sensing (CS) is the field used for the reconstruction of signals which are sparse in a transformation domain. A sparse signal is a signal with very few transform coefficients. Knowing that a signal is sparse, it can be reconstructed using less number of samples than in the conventional way. The main two conditions that have to be met for the successful reconstruction are sparsity and incoherence of the signal. There are many methods and algorithms developed since the introduction of CS, classified into two large groups, one being greedy algorithms and the other based on convex relaxation algorithms. Since many signals are sparse in the transformation domain, the application of the CS algorithms is widely spread in all areas of digital signal processing, such as multimedia, biomedicine, radars, etc. The advantage in using CS algorithms is that it is memory efficient and, as long as the conditions are met, it gives a unique solution to the reconstruction [1]-[7]. The algorithm used in this paper is a modified version of the gradient-based algoritjm [8], [9], from the convex relaxation group. The algorithm showed good results in image denoising in the presence of different kind of noise. The advantage of this algorithm is that the image does not have to be strictly sparse. So far, the cases of salt and pepper and Gaussian noise were considered. In this paper, we use the uniform random noise, which means that the noise values are within the range of the uncorrupted pixels. That is, the detection is different and much harder in this case than in the previous ones. Another complex circumstance is that we do not have the knowledge of the positions of corrupted pixels.

In this paper, processing of sparse signals will be used in a more specific way. The reduced set of samples will be unchanged and the unavailable samples will be presented as the uniform random noise, so that their values are within the range of the available pixels. In comparison to the previous results of the algorithm, there will be additional steps in the reconstruction. The assumption is that we have missing/corrupted pixels in the spatial domain, and that we know that the original image (i.e. the image with no corrupted pixels) is sparse in the transformation domain. In contrast to the greedy algorithms, the values of the uncorrupted (available) samples will remain unchanged.

The paper is organised as follows: after the introduction part, basic theoretical background about the image reconstruction and proof of the detection of the positions using the algorithm are shown in Section 2. In Section 3, the reconstruction algorithm and its explanation are presented. In Section 4, the numerical results for two-dimensional synthetic signals are shown, as well as for a test image.

II. Theoretical Background

Let us consider an $N \times N$ 8-bit grayscale sparse image with impulsive noise. The image can be defined as

$$x_a^{(0)}(m,n) = \begin{cases} x(m,n), & \text{for } (m,n) \in \mathbb{M} \\ 0 \text{ (or } 255), & \text{elsewhere} \end{cases} \quad (1)$$

where 0 is black and 255 is white colour, with M number of uncorrupted (available) pixels and $\mathbb{M} = \{(m_1,n_1),(m_2,n_2),...,(m_M,n_M)\}$ presents set of the uncorrupted pixels. The image is sparse in the DCT domain and the sparsity is K. For the gradient algorithm, the reconstruction is done in the spatial domain. When the available pixels are known, or the image has impulsive noise so the corrupted ones are distinguishable from the available ones, the image does not have to be strictly sparse.

The approach considered in this paper deals with an image with some number of pixels corrupted with uniform random noise and with a reduced set of available pixels. The aim is to

This work is supported by the Montenegrin Ministry of Science, project grant funded by the World Bank loan: CS-ICT "New ICT Compressive sensing based trends applied to: multimedia, biomedicine and communications".

Authors are with the University of Montenegro, Faculty of Electrical Engineering, Dzordza Vasingtona bb, 20 000 Podgorica, Montenegro. Corresponding author is Isidora Stanković (e mail: isidoras@ac.me).

reconstruct the pixels that are corrupted without knowing the positions and the number of the corrupted samples. To make the reconstruction possible, the sparsity in the transformation domain should be implicitly assumed.

Before the introduction of the algorithm, we will present why the gradient is high for the unavailable/corrupted samples. Consider a sparse signal $x(m,n)$ as in equation (1). The signal with one corrupted sample at position (m_0,n_0) can be defined as $x_a(m_0,n_0)=x(m_0,n_0)+z$ with z being the uniform noise. Due to the change of z, form the signals

$$\begin{aligned} x_a^+(m,n) &= x(m,n)+(z+\Delta)\,\delta(m-m_0,n-n_0) \\ x_a^-(m,n) &= x(m,n)+(z-\Delta)\,\delta(m-m_0,n-n_0) \end{aligned} \tag{2}$$

where Δ is the gradient parameter. The sparisty measure is defined as

$$g(m_0,n_0)=\frac{\|\mathbf{X}_a^+\|_1+\|\mathbf{X}_a^-\|_1}{2N\Delta} \tag{3}$$

where $\mathbf{X}_a^+$ and $\mathbf{X}_a^-$ are the vectors of the transformation domains of the signals in (2). Let us consider that the transformation domain is the two-dimensional Discrete Cosine Tranform (2D-DCT). The proof for 1D-DFT can be found in [10]. So, the transformation domains of signals $x_a^+(m,n)$ and $x_a^-(m,n)$ can be written as

$$\begin{aligned} X_a^+(k,l) &= X(k,l)+(z+\Delta)\,\varphi(k,l) \\ X_a^-(k,l) &= X(k,l)+(z-\Delta)\,\varphi(k,l) \end{aligned} \tag{4}$$

where

$$\varphi(k,l)=\cos\left(\frac{2\pi(2m_0+1)k}{4N}\right)\cos\left(\frac{2\pi(2n_0+1)l}{4N}\right). \tag{5}$$

The 2D-DCT of the known samples is $X(k,l)$ and $(z\pm\Delta)\,\varphi(k,l)$ is the DCT of the missing (corrupted) sample. The measures can be defined as the addition of the measure of the original (non-corrupted samples) and the corrupted one (with the Δ shifts)

$$\begin{aligned} \|\mathbf{X}_a^+\|_1 &= \textstyle\sum_{k,l=0}^{N-1}|X_a^+(k,l)| \cong \mu+|z+\Delta|\,C \\ \|\mathbf{X}_a^-\|_1 &= \textstyle\sum_{k,l=0}^{N-1}|X_a^-(k,l)| \cong \mu+|z-\Delta|\,C \end{aligned} \tag{6}$$

where $\mu=\|\mathbf{X}\|_1$ is the sparsity measure of the original signal $x(n)$ and C is a constant dependent on number of samples N and is close to $C\cong\sqrt{\frac{\pi}{2}}$. For example, when $N=64$, the constant $C=1.2214$ and when $N=512$, $C=1.2321$. The sparsity measure can be written as

$$\begin{aligned} g(m_0,n_0) &= \frac{\|\mathbf{X}_a^+\|_1+\|\mathbf{X}_a^-\|_1}{2N\Delta}\cong \\ &\cong \frac{|z+\Delta|\,C-|z-\Delta|\,C}{2N\Delta}. \end{aligned} \tag{7}$$

For deviations from the true signal value smaller than the step $|z|<\Delta$ we get

$$g(m_0,n_0)\cong\frac{Cz}{\Delta}\sim Cz, \tag{8}$$

which means that the gradient is proportional to the shift of the corrupted samples at its position, and the gradient is equal to zero at the positions of the available samples.

The sparsification is done using the quantization matrix of the JPEG standard. The quantization is applied on the 8×8 DCT blocks with 50% quality factor. In this paper, the standard was assumed with quality factor being 25%, which defines the number of the DCT components considered in a block [11].

III. Reconstruction Algorithm

The gradient algorithm is based on the efficient minimisation of the missing/corrupted samples. For the beginning of the description, assume that the number and positions of the available pixels are known. For this scenario, the one-dimensional algorithm is introduced in [8] and two-dimensional in [12], [13]. Let us consider a two-dimensional signal $x(m,n)$, which is sparse in the 2D-DCT domain. Missing values are considered as variables. Then the signal with only M available samples in the first iteration can be represented as in equation (1).

A. Algorithm

Step 1: Firstly, we add and subtract a value Δ to the corrupted/missing samples. The new signals are

$$\begin{aligned} x_a^+(m,n) &= x^{(p)}(m,n)+\Delta\delta(m-m_i,n-n_i) \\ x_a^-(m,n) &= x^{(p)}(m,n)-\Delta\delta(m-m_i,n-n_i) \end{aligned} \tag{9}$$

Step 2: The transforms of such signals are calculated and the gradient value is estimated as a finite difference of ℓ_1-norms of these transforms. Based on the gradient value, the missing signal sample $x(m_i,n_i)$ is increased or decreased. For each $(m_i,n_i)\notin\{(m_1,n_1),(m_2,n_2),...,(m_M,n_M)\}$ (i.e. the missing samples) the gradient value is approximately:

$$g(m_i,n_i)=\frac{\|X_a^+(k,l)\|_1-\|X_a^-(k,l)\|_1}{2\Delta} \tag{10}$$

where $X_a^+(k,l)=T\{x_a^+(m,n)\}$, $Xa^-(k,l)=T\{x_a^-(m,n)\}$ and Δ is the missing samples change step. In our case, $T\{.\}$ is the 2D-DCT.

Step 3: The available samples are not changed. Each missing pixel value $(m_i,n_i)\notin\{(m_1,n_1),(m_2,n_2),...,(m_M,n_M)\}$ is changed in the direction opposite of the gradient for a step μ

$$x_a^{(p)}(m_i,n_i)=x_a^{(p-1)}(m_i,n_i)-\mu g(m_i,n_i). \tag{11}$$

Step 4: Since ℓ_1-norm is convex but of specific form, gradient based procedure will approach the solution of the problem up to the value proportional to the steps Δ and μ. When the values are close to the true signal values, they will oscillate around the solution with magnitude proportional to the step. When the oscillation is detected, the step size is reduced, for example as $\Delta=\Delta/3$ and $\mu=\mu/3$. These new

parameters continue approach to the true signal values until the new precision is reached. The procedure is repeated until the desired reconstruction accuracy is achieved.

Step 5: The algorithm is stopped when the change in two successive iterations is smaller than the desired accuracy ε,

$$\max_{m,n} \left| x_a^{(p)}(m,n) - x_a^{(p-1)}(m,n) \right| < \varepsilon. \tag{12}$$

Note that only missing samples contribute to this difference, since the available samples are not changed. The proof of the L_1-norm convexity with respect to the corrupted pixels as variables can be found in [14]. This is the basic reconstruction algorithm with the knowledge of the position of the missing samples or, at least, with very high impulsive noise so that the positions of the corrupted samples/pixels are distinguishable from the non-corrupted ones. In the next subsection, a method for selection of the corrupted pixels is explained.

B. Pixel selection

There are two ways in which the corrupted sample can be selected: single-step and the iterative way. For both ways we repeat Step 1 and 2 (i.e. equations (9) and (10)) for all samples since we do not know the positions of missing ones:

$$x_a^+(m,n) = x^{(p)}(m,n) + \Delta\delta(m-m_i, n-n_i)$$
$$x_a^-(m,n) = x^{(p)}(m,n) - \Delta\delta(m-m_i, n-n_i)$$

and

$$g(m_i,n_i) = \frac{\|x_a^+(k,l)\|_1 - \|x_a^-(k,l)\|_1}{2\Delta}.$$

Note that the value of Δ here is different to the value of Δ in the algorithm itself. For this calculation, Δ must be a high number i.e. $\Delta > \max_{m,n} = x(m,n)$.

For the single-step estimation, we sort the values of the gradient and set a threshold to find largest M values. The detected positions are considered as positions of the corrupted samples. Other values are considered as the available pixels. A drawback in this calculation is that the information about the number of the missing/corrupted samples (or at least approximate number) is crucial.

For the iterative way, we do not need the information about the number of corrupted samples nor the positions of them. The procedure is similar: each time we take the pixel with the largest gradient, reconstruct it (Note: use the algorithm explained above for only one missing sample with a known position), eliminate it from the array of possible values, and repeat the whole procedure. This will be repeated until the error of two successful iterations is below an acceptable level.

IV. Numerical Results

The algorithm was already tested on one-dimensional signals in [15]. The extension to two-dimensional signals is presented here. In the first subsection, a synthetic two-dimensional signal is shown to explain how the algorithm was improved in the two-dimensional case. In the second subsection, the image reconstruction using a grayscale image "Lena" is shown.

A. Two-dimensional signals

In this section, a synthetic two-dimensional signal is introduced. The signal is given in the form:

$$x(m,n) = \sum_{i=1}^{K} A_{p_i q_i} \varphi(k_{p_i}, k_{q_i}) \tag{13}$$

where

$$\varphi(k_{p_i}, k_{q_i}) = \cos\left(\frac{2\pi(2m-1)k_{p_i}}{4N}\right)\cos\left(\frac{2\pi(2n-1)k_{q_i}}{4N}\right).$$

$A_{p_i q_i}$ and k_{p_i,q_i} are random amplitudes and random frequency positions of the signal. The image is a $N \times N$ signal where $N = 16$, $M = 196$ signal samples are available and the sparsity (i.e. number of DCT components) is $K = 6$ (consider 6 components in a $16 \times 16 = 256$ samples signal). From this it follows that approximately 25% $(196/256)$ of the image is missing. Fig. 1 shows the result for the case when missing samples are set to zero (and the positions are known). In Fig. 1(top) we present the original image and its DCT, Fig. 1(middle) is the image with the available values (the unavailable are the blue pixels) and Fig. 1(bottom) is the reconstructed image.

The improvement achieved using the introduced algorithm is that we can reconstruct signals whose corrupted samples are within the range of the available pixel values. It is hard to distinguish the difference between the corrupted and available pixels and to find the unavailable ones (as it is seen in Fig. 2). In the first case (explained above), if we do not know the positions of the available ones, the reconstruction will not be successful. In this example, we will take 50% of samples with largest gradient and set them to zero. This will make the algorithm straightforward. The original image, the one with available pixels and the reconstruction image are shown in Fig. 2. The order of images is as in the previous one.

B. Image reconstruction

Unlike the classical reconstruction of the image using the gradient-based algorithm, we need to assure the image is sparse in the DCT domain. We will represent it with the standard JPEG image reconstructed with 8×8 reconstruction blocks.

The 512×512 image "Lena" image is tested using the algorithm presented. As the sparsity should be as high as possible (meaning that we use as few DCT components as possible), without destroying the visual effect too much, it is seen that the quality of image using quality factor $QF = 25$ is very similar to the quality of 50% and with less components (i.e. sparser). Because of that, the analysis is done for the $QF = 25$ (higher sparsity will give higher probability for reconstruction using CS algorithms). The original image and the sparsified image are shown in Fig. 3. The sparse image is then split into blocks and each block has 12.5% of corrupted samples. The block size is 8×8. The noisy and reconstructed images are shown in Fig. 4.

C. Comparison

The structural similarity (SSIM) index value is defined as a function of luminance, contrast and structure of two images

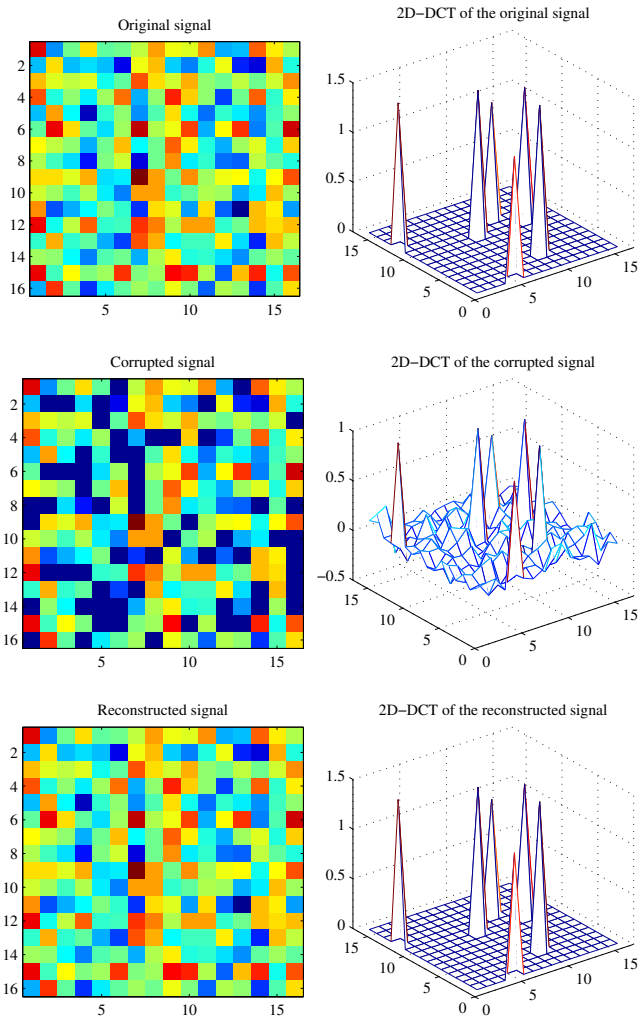


Fig. 1. Original signal and its 2D-DCT (top); Signal with corrupted samples and its 2D-DCT (middle); Reconstructed signal and its 2D-DCT (bottom)

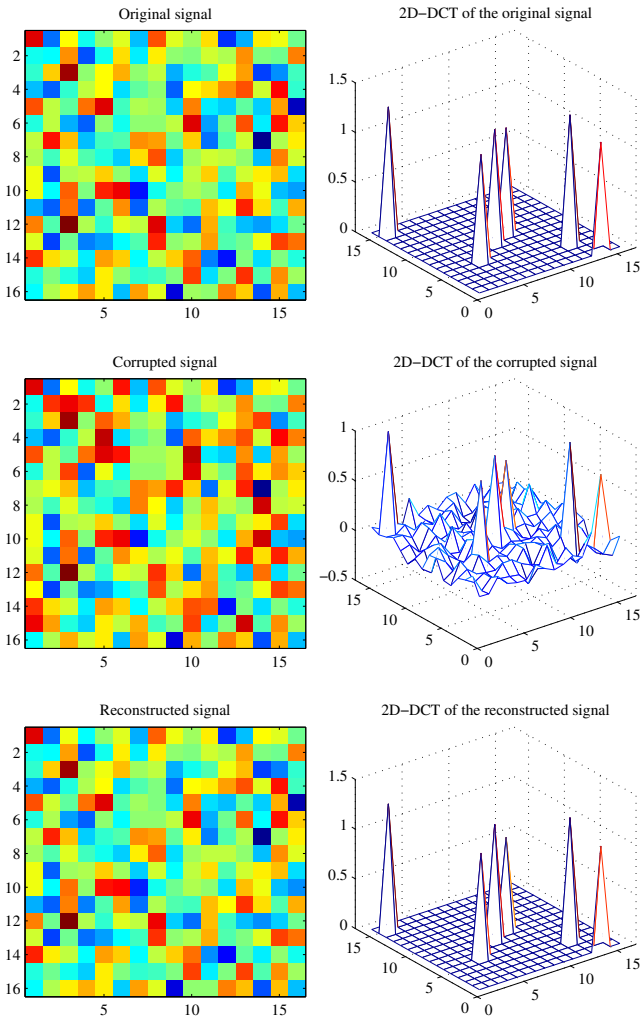


Fig. 2. Original signal and its 2D-DCT (top); Signal with corrupted samples and its 2D-DCT (middle); Reconstructed signal and its 2D-DCT (bottom)

[16], and can be written as

$$\text{SSIM}(x_1, x_2) = \frac{(2\mu_{x_1}\mu_{x_1} + c_1)(2\sigma_{x_1x_2} + c_2)}{(\mu_{x_1}^2 + \mu_{x_2}^2 + c_1)(\sigma_{x_1}^2 + \sigma_{x_2}^2 + c_2)} \quad (14)$$

where x_1, x_2 are the two considered images (in our case x_1 is the original sparse image and x_2 is the reconstructed image), μ_{x_1}, μ_{x_2}, and $\sigma_{x_1}, \sigma_{x_2}$, and $\sigma_{x_1x_2}$ represent the mean, variance and covariance of the two images respectively and c_1, c_2 are stabilization variables. If the SSIM index is close to 1, the images are highly similar, and if the index is close to 0, it means that the images are not similar. In our case, the SSIM index calculated for the reconstructed image is 0.9858. The SSIM index for the median filter of size 3×3 is calculated 0.9825 and for the 5×5 is 0.9486. The MAE is calculated as $MAE(x_1, x_2) = mean(mean(|x_1 - x_2|))$. For the gradient algorithm, it is calculated to be 0.5693, for the median 3×3 is 1.9726, and for 5×5 is 3.3173. The disadvantage of median filtering is that it degrades the whole image. The gradient algorithm reconstructs only the pixels which are found to be corrupted, and the pixels which are set as uncorrupted/available remain unchanged.

V. Conclusions

The gradient-based algorithm for reconstruction of noisy images with noise pixel values being within the range of the available pixel values is shown in this paper. The reconstruction is done on a synthetic two-dimensional image and on a grayscale test image. On the synthetic two-dimensional signals, the reconstruction using the gradient-based algorithm in a conventional way is shown as well, to explain the difference between what was done before and what are the features of the modified algorithm. When the number and the positions of the corrupted samples are unknown (and additionally the corrupted samples are in the range of the available ones), we use the algorithm presented here. It is shown that the reconstruction can be done successfully without the knowledge of corrupted sample positions. There are few parts that we have to consider in the reconstruction of a noisy image. In contrast to the conventional gradient (i.e. when we have the knowledge of the available pixels), the image has to be sparse in the transformation domain.





Fig. 3. Original image (top); Sparse image (bottom)

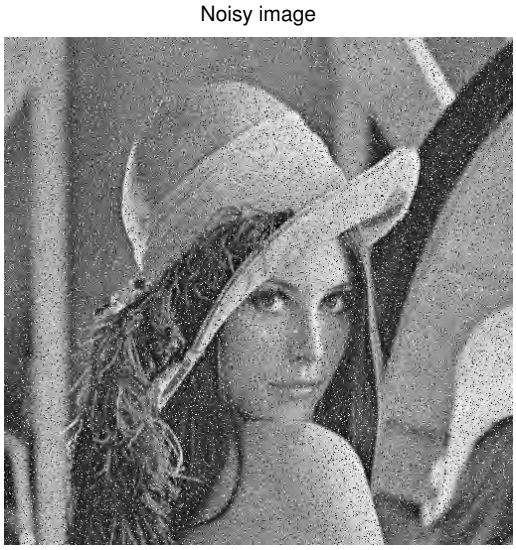




Fig. 4. Noisy image (top); Reconstructed image (bottom)

REFERENCES

[1] R. Baraniuk, "Compressive Sensing," *IEEE Signal Processing Magazine*, vol. 24, no. 4, pp. 118-121, July 2007.

[2] E.J. Candes, and M. Wakin, "An Introduction to Compressive Sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21-30, March 2008.

[3] E.J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489-509, February 2006.

[4] D.L. Donoho, "Compressive Sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289-1306., April 2006.

[5] S. Stanković, I. Orović, LJ. Stanković, "An Automated Signal Reconstruction Method based on Analysis of Compressive Sensed Signals in Noisy Environment," *Signal Processing*, vol. 104, Nov 2014, pp. 43 - 50, 2014.

[6] LJ. Stanković, S. Stanković, M. Amin, "Missing Samples Analysis in Signals for Applications to L-estimation and Compressive Sensing," *IEEE Signal Processing Letters*, vol. 94, Jan 2014, pp. 401-408, 2014.

[7] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, 2010.

[8] LJ. Stanković, M. Daković, and S. Vujović, "Adaptive Variable Step Algorithm for Missing Samples Recovery in Sparse Signals," *IET Signal Processing*, vol. 8, no. 3, pp. 246 -256, 2014. (arXiv:1309.5749v1).

[9] M.A.T. Figueiredo, R. D. Nowak, S. J. Wright, "Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems." *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, no. 4, pp. 586 - 597, December 2007.

[10] LJ. Stanković, *Digital Signal Processing with Selected Topics*. CreateSpace Independent Publishing Platform, An Amazon.com Company, November 4, 2015.

[11] S. Stanković, I. Orović, and E. Sejdić, *Multimedia signals and systems*. Springer - Verlag, 2012.

[12] I. Stanković, "Recovery of Images with Missing Pixels using a Gradient Compressive Sensing Algorithm", arXiv preprint, arXiv: 1407.3695

[13] I. Stanković, I. Orović, and S. Stanković, "Image Reconstruction from a Reduced Set of Pixels using a Simplified Gradient Algorithm," *22nd Telecommunications Forum TELFOR 2014*, Belgrade, Serbia

[14] M. Brajović M. Daković, L. Stanković, "Convexity of the l1-norm based Sparsity Measure with Respect to the Missing Samples as Variables," *Mediterranean Conference on Embedded Computing (MECO) 2016*, *accepted.*

[15] LJ. Stanković, M. Daković, and S. Vujović, "Reconstruction of Sparse Signals in Impulsive Disturbance Environments," *submitted.*

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, April 2004.

Compressive Sensing based image processing in TrapView pest monitoring system

Milan Marić*, Irena Orović** and Srdjan Stanković **
* S&T Crna Gora d.o.o, Podgorica, Montenegro
** University of Montenegro, Faculty of Electrical Engineering, Podgorica, Montenegro
irenao@ac.me

Abstract – Missing information in image can be recovered by using the principles of lower sample rate methods, such as Compressive Sensing. This method can, at the same time, recover the missing information in the signal and do the compression of the original data. Lowering the sample rate is especially suitable for natural images in applications where minor visually loss of fidelity is acceptable. The goal is to achieve a substantial reduction in bit rate and image size. In this paper we analyze the performance and quality of Compressive Sensing approach applied on images captured by the TrapView automated camera station for pest detection. The reconstruction at the decoder side, if only small number of image samples is available, is tested in the paper. This is done with the goal to test different approach in image capturing – acquisition of only part of digital data, and then the reconstruction of the uncaptured/missing part, in order to obtain the original signal. This leads to decreasing the bit rate and transferred data volume through the mobile network, from the station to the TrapView cloud centre. It is shown that CS can provide a good quality image reconstruction with significantly reduced number of samples. The theory is tested using real images, obtained by the TrapView camera.

I. INTRODUCTION

The alternative ways for signal sampling are intensively studied in the recent years. Traditional sampling based on the Shannon-Nyquist theorem is time consuming and requires large number of signal samples to be stored and transmitted. This can be a limiting factor in applications that deal with high resolution images. Compressive Sensing (CS) [1]-[4] approach can, under certain conditions, ensure successful reconstruction of such images using significantly smaller number of samples, compared to the conventional approach. CS uses random sampling strategy and powerful mathematical algorithms to reconstruct an under-sampled signal [5]-[10].

Special CS conditions that need to be satisfied in order to provide high accuracy signal reconstruction are signal sparsity and incoherent acquisition procedure [1]-[3]. Namely, the signal has to be sparse in a certain transform domain, meaning that the information about the signal needs to be concentrated within a small number of coefficients. Incoherence property should provide linearly independent measurements and assures signal reconstruction from a small number of acquired samples.

From the acquired measurements, signal is reconstructed by using different optimization techniques, which can be based on various norms minimization. The optimal solution in a large number of applications is provided by using ℓ_1-norm minimization, for 1D signals, while in image processing applications, the commonly used optimization technique is the Total Variation (TV) minimization [3], [6]-[11]. Here, we will apply the CS approach with TV minimization in pest monitoring system, in order to speed up processing and decrease memory requirements for storing pest images.

TrapView automated pest monitoring system [12] utilizes automated traps that regularly send captured images of caught insects to the TrapView cloud. Therefore, TrapView is a platform for an early warning of the occupancy of traps. System also provides automated pest recognition, pest occurrence statistics and manual review of the taken pictures.

In the fields/orchards, where matrix monitoring concept is being implemented, the effort to increase the number of monitoring points leads to the Trapview automated traps being combined with conventional traps. This results in a need for a tool that allows the end users to optimize captured image prior sending over mobile network, thus reducing cost for communication and storing captured images.

Standard TrapView captured image is around 2MB large, and one system in field with 1000 automated camera stations normally generates 1TB volume of captured data per year. All the captured data should be sent over mobile network to the TrapView cloud in real time, which needs performances and proper bandwidth of mobile network, which is not always achievable out of the urban centers.

In this paper, we have explored the possibility to use CS to decrease the amount of relevant image data, but still to keep the quality of the final reconstructed image, in order to be successfully post-processed. Post-processing of the images is, in fact, counting the number of insect specimens and it is based on algorithms specialized for counting. Therefore, it is of great interest to save the quality of the reconstructed image as better as is possible, in order to minimize the possibility of error occurrence during counting. CS, as an emerging approach

for reconstruction of the under-sampled signal, has been used as a tool for achieving both, compression and reconstruction of the images captured by the system.

The paper is organized as follows: Second part is theoretical background on the CS approach. Third part describes CS approach in image reconstruction, and its application in the TrapView pest monitoring system. Concrete results, using images captured by the TrapView system, are shown in the fourth part. Concluding remarks are given next, as well as literature overview.

II. THEORETICAL BACKGROUND

Signal acquisition process in real applications is mainly done according to the sampling theorem. In order to be transmitted and stored, the data are compressed by using algorithms for data compression that can be computationally very complex. CS approach brings an idea to achieve compression in sense of the acquisition, and to deal with far less data than we usually do. Consequently, it is important to develop algorithms for the reconstruction of compressive sensed data, and certain solutions have already been proposed in the literature.

In CS scenarios, the acquisition procedure is performed by using small set of randomly chosen signal samples. This means that the signal acquisition rate is much smaller than that required by the Shannon-Nyquist theorem. However, certain conditions have to be satisfied in order to apply CS approach, such as incoherence of measurement process and signal sparsity. The sparse representation means that the information about the signal is condensed into few non-zero coefficients in the transform domain. An *N*-dimensional signal could be written in terms of its transform domain representation, as [3], [13], [14]:

$$\mathbf{x} = \sum_{i=1}^{N} \mathbf{X}_i \psi_i = \mathbf{\Psi X}, \tag{1}$$

where $\mathbf{X}_i$ is a transform domain coefficient, ψ_i is a basis vector, $\mathbf{\Psi}$ denotes $N \times N$ transform matrix and $\mathbf{X}$ is the signal in $\mathbf{\Psi}$ domain. The acquired measurements are stored in vector $\mathbf{y}$. Hence, we can write:

$$\mathbf{y} = \mathbf{\Phi x} = \mathbf{\Phi \Psi X} = \mathbf{AX}, \tag{2}$$

where $\mathbf{\Phi}$ is a measurement matrix and $\mathbf{A}$ is called a CS matrix.

The signal is reconstructed by solving the set of linear equations defined by (2). As there are *M* linear equations with *N* unknowns, this system is undetermined and can have infinitely many solutions. Therefore, the optimization algorithms should be used to find the solution of the problem, which is in fact, finding the sparsest solution of the system (2).

Commonly used algorithms for reconstruction of the under-sampled signals are based on minimization of different norms (commonly used is ℓ_1 - norm minimization) [2], [3], [13]-[16]. From the class of greedy algorithms, the Orthogonal Matching Pursuit (OMP) is the most commonly used. OMP is an iterative procedure that, in each iteration, searches for the maximum correlation between the measurements and the matrix [17]. There are also gradient based algorithms, such as [8], [18], [19] and threshold based algorithms [20], [21]. Regarding the reconstruction of natural images, different approaches are used. One of the commonly used approaches is known as the TV minimization, and it is based on the optimization problem defined using image gradient [11], [22]-[24]. Therefore, in this paper we will use TV minimization approach for pest images reconstruction.

III. RECONSTRUCTION OF THE TRAPVIEW IMAGES USING THE CS APPROACH

Having in mind that natural images do not show sparsity property neither in the frequency nor in the spatial domain, for its reconstruction from a relatively small number of available samples, ℓ_1 - norm minimization is not very suitable. There are an alternative methods for image reconstruction and they are based on the fact that the image gradient is sparse. Therefore, minimization is reduced to the ℓ_1 - norm of the gradient minimization (TV minimization). For most real signals, TV minimization shows better results compared to the minimization based on the ℓ_2 - norm. TV method proved to be very efficient in the regularization of the image, without damaging the edges of objects in the image (i.e., keeping the edges of objects).

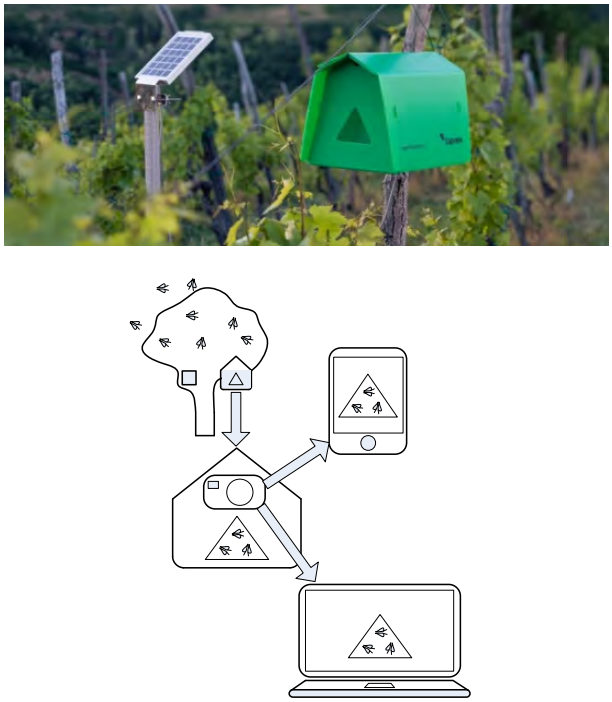


Figure 1: TrapView system

Starting from (2), and having available only the vector of measurements **y**, the goal is to estimate/reconstruct the entire image from the available samples in **y**. The measurements in real applications may also contain certain amount of noise, i.e.:

$$\mathbf{y} = \mathbf{A}\mathbf{X} + \mathbf{n}, \tag{3}$$

where **n** is an additive noise. In order to solve (3), the regularization function can be defined as:

$$\Psi_{\lambda}(\varepsilon) = \arg\min_{\mathbf{X}} \frac{1}{\mu}\Phi_{reg}(\mathbf{X}) + \frac{1}{2}\left\|\mathbf{y} - \mathbf{A}\mathbf{X}\right\|^2, \tag{4}$$

with $\Phi_{reg}(\mathbf{X}) = \sum_i \left\|\nabla_i \mathbf{X}\right\|$ for TVL1 problems. The TV of the signal **X**: $\sum_i \left\|\nabla_i \mathbf{X}\right\|$ represents a sum of the magnitudes of discrete gradient at each point, and can be defined as:

$$\nabla_{i,j}\mathbf{X} = \begin{bmatrix} \mathbf{X}(i+1, j) - \mathbf{X}(i, j) \\ \mathbf{X}(i, j+1) - \mathbf{X}(i, j) \end{bmatrix}. \tag{5}$$

Let us now define an acquisition procedure for the pest monitoring system TrapView. System is illustrated in Figure 1. A trap is placed in the field (orchard, vineyard) in order to catch targeted insects. At the same time, the system takes pictures of what was caught in the trap and send them to the cloud. The pictures are fed to the end user either via mobile or via web. Therefore, we have original image which can be under-sampled and such under-sampled image is sent to the web/mobile application.

The samples from the collected images are taken from the DCT domain, and are consisted of the two parts: low frequency coefficients of length K_1 and coefficients from the rest of the DCT plane K_2:

$$\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2, \tag{6}$$

where $\mathbf{y}_1$ are low frequency DCT coefficients, and $\mathbf{y}_2$ are the rest of the coefficients. During the acquisition procedure, low frequency coefficients are collected first, as they contain most of the image energy which is of great importance for preserving image quality. The TV minimization problem to be solved, is then defined as:

$$\min_{\mathbf{X}} \mathrm{TV}(\mathbf{X}) \;\; \textit{subject to} \;\; \mathbf{y}_1 + \mathbf{y}_2 = \mathbf{A}\mathbf{X}, \tag{7}$$

Or, in discrete form:

$$\mathrm{TV}(\mathbf{X}) = \sum_{i,j} \sqrt{(\mathbf{X}_{i+1,j} - \mathbf{X}_{i,j})^2 + (\mathbf{X}_{i,j+1} - \mathbf{X}_{i,j})^2}. \tag{8}$$

Such automated insect monitoring approach reduces travel and time costs, especially in cases when there is need for monitoring large areas, consisting of several hundreds of square kilometers. Images are sent on daily basis to the end user. However, as images are of high resolution, it would be of great importance to reduce image size in order to speed up upload and download procedure by the end user.

IV. EXPERIMENTAL RESULTS

In the sequel, under-sampling and reconstruction of the images recorded by the TrapView camera, at the end user point, is done by using the CS approach.

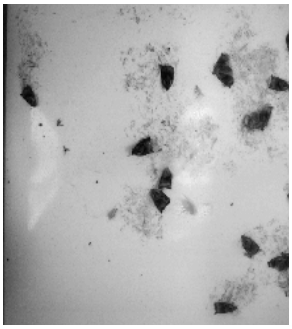


Figure 2: Original image

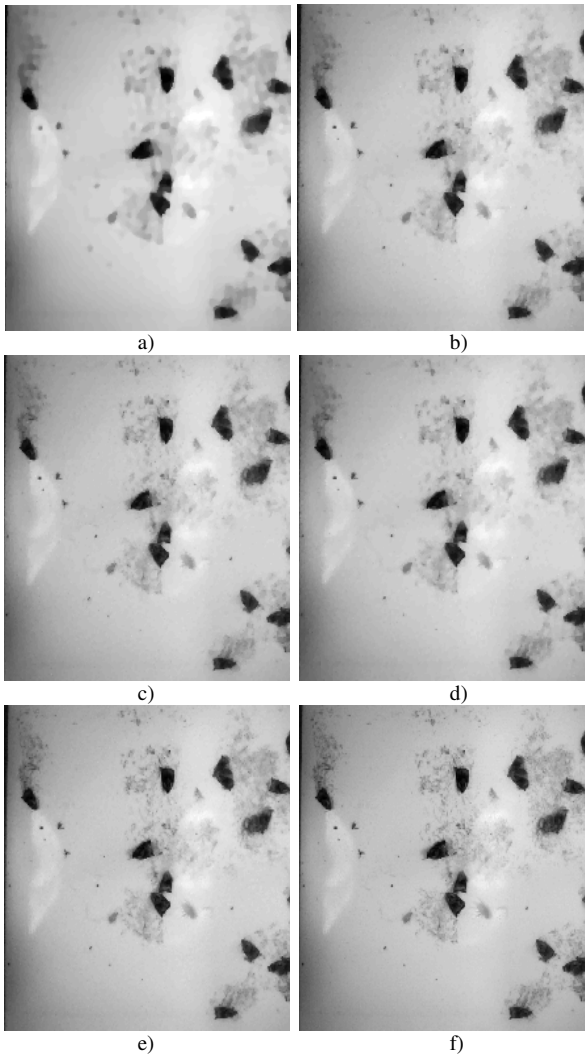


Figure 3: Reconstructed images from the beginning phase, using different number of measurements: a) 2.7%, b) 10.68%, c) 15%, d) 21%, e) 30.14%, f) 39.67% of the total number of samples

We have observed the images collected at different phases. Firstly, the beginning phase is tested. Figure 2 shows original image recorded in this phase.

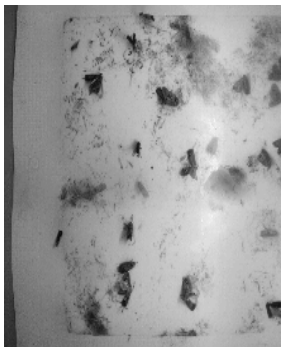


Figure 4: Original image

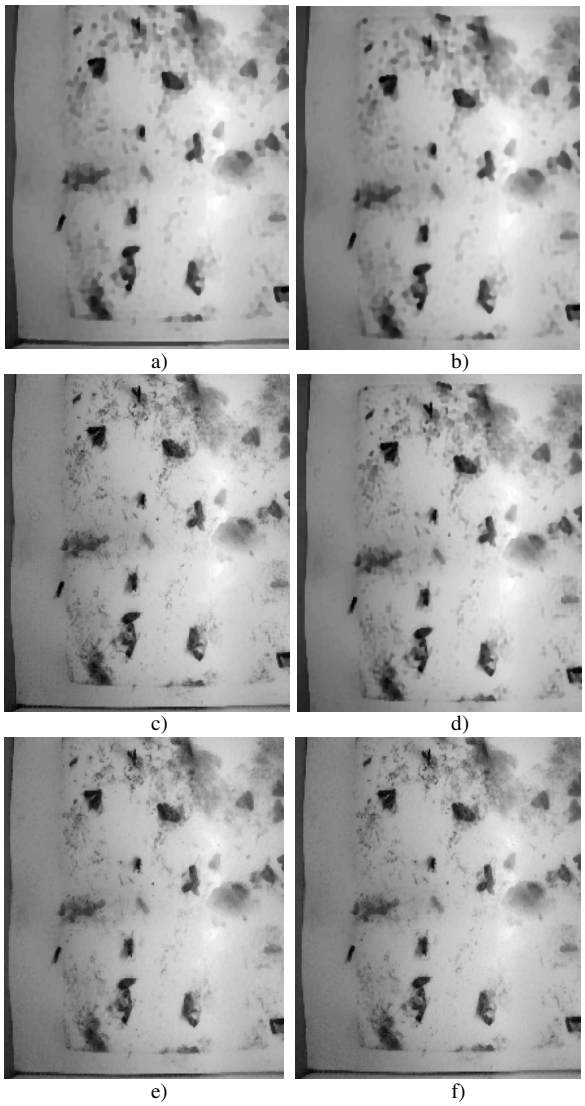


Figure 5: Reconstructed images using different number of measurements: a) 2.7%, b) 10.68%, c)15%, d) 21%, e) 30.14%, f) 39.67% of the total number of samples

It can be seen that the adhesive tape is relatively clear and there is small number of collected insect specimens. The reconstruction is done by using different number of available samples: from around 3% to around 40% of the total number of samples. Certain number of the low-frequency coefficients is taken in every considered case, having in mind that important information about the image is concentrated mostly in these coefficients. The rest of the coefficients are randomly selected from the rest of the image samples. The reconstructed images are shown in Figure 3. It can be seen that, even with around 3% of the available samples, the image can be reconstructed with high quality.

The same procedure is repeated with an image from the mature phase, i.e. from the phase when the adhesive tape is densely filled with insect specimens. Original image is shown in Figure 4, while Figure 5 shows the reconstructed images. Number of measurements is from around 3% to around 40% of the total number of samples.

Quality of the images is tested by visually observing original and reconstructed image, and by measuring peak signal to noise ratio (PSNR). Values of the PSNR, for different number of measurements used and for images from the both phases, are shown in the Table 1. It can be seen that PSNR for the first image (beginning phase), is slightly higher compared to the second image (mature phase). The reason for this is smaller number of insect specimens present in the first image, which leads to its higher sparsity.

It can be seen that the image can be reconstructed with almost the same quality as the original image, by using relatively small set of available information. The PSNR around 33 dB is obtained by using 3% of the total number of samples for the images from the beginning phase, and around 11% for the images from the mature phase. Having in mind high PSNR, and by visually comparing the original and reconstructed image, it can be concluded that CS approach will not affect post-processing of the images, i.e., the process of specimens counting. In this paper, we have used frequency domain measurements. It is important to note that measurements can be also used from the spatial domain, depending on the optimization algorithm that is applied.

Table 1: Number of measurements used for image reconstruction and measures of image quality (PSNR) for the images from the both phases

Number of measurements (%)	K1	K2	PSNR (dB) – image 1	PSNR (dB) – image 2
2.7	4000	3000	33.08	31.14
10.68	4000	24000	36.02	33.91
15	4000	35000	36.94	34.88
21	4000	52000	38.08	35.98
30.14	4000	75000	39.30	37.19
39.67	4000	100000	40.56	38.47

V. CONLUSION

The possibility to apply CS reconstruction approach on the images used in pest monitoring system TrapView, is tested in the paper. The reduction of number of acquired samples, and transmission of only small set of collected measurements to the end user, is of great importance in reducing the memory requirements for storing these images. The main idea is to capture not the whole image (i.e. full set of samples), but only a small percent of image samples at the random positions. The image reconstruction by using only this small number of captured information and based on the CS algoritms, is tested. Therefore, the proposed method is not an compression method, but is related to the recovering of the missing information. The images captured at different phases are tested, and they are reconstructed by using different number of available samples (frequency domain measurements are used): from 3% to 40% of the total number of samples. It is shown that the quality of the reconstructed images, i.e. PSNR, is high. In all considered cases, the obtained PSNR is above 30 dB. The CS approach can also be applied in situations when image is spatially under-sampled, which is the topic of our future research.

ACKNOWLEDGMENT

This work is supported by the Montenegrin Ministry of Science, project grant funded by the World Bank loan: CS-ICT "New ICT Compressive sensing based trends applied to: multimedia, biomedicine and communications". The authors are thankful to Mr. Matej Štefančič, the director of EFOS, for providing us test images and support for writing the paper.

REFERENCES

[1] E. J. Candes, and M. B. Wakin, "An Introduction to Compressive Sampling," IEEE Signal Processing Magazine, vol. 25, no. 2, pp. 21, 30, March 2008.

[2] R. Baraniuk, "Compressive Sensing", Lecture Notes in IEEE Signal Processing Magazine, vol. 24, July 2007.

[3] S. Stankovic, I. Orovic, and E. Sejdic, "Multimedia Signals and Systems," Springer-Verlag, New York, 2012.

[4] L. Stankovic, I. Orovic, S. Stankovic, and M. Amin, "Compressive Sensing Based Separation of Non-Stationary and Stationary Signals Overlapping in Time-Frequency," IEEE Transactions on Signal Processing, Vol. 61, no. 18, pp. 4562-4572, September, 2013.

[5] R. Monteiro, and I. Adler, "Interior path following primal-dual algorithms. Part I: Linear programming, Mathematical Programming," 44 (1989), pp. 27–41.

[6] P. Boyd, and L. Vandenberghe, "Convex Optimization," Cambridge University Press, March 8, 2004 - Business & Economics - 716 pages.

[7] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," IEEE Transactions on Information Theory, vol.50, no.10, pp.2231-2242, October 2004.

[8] T. Blumensath, and M. E. Davies, "Gradient Pursuits," IEEE Transactions on Signal Processing, vol.56, no.6, pp.2370-2382, June 2008.

[9] T. Blumensath, and M. E. Davies, "Iterative Thresholding for Sparse Approximations," Journal of Fourier Analysis and Applications, vol. 14, no. 5-6, pp 629-654, December 2008.

[10] S. Stankovic, I. Orovic, and L. Stankovic, "Single-iteration algorithm for compressive sensing reconstruction," 21st Telecommunications Forum (TELFOR), 2013, 447-450.

[11] J. M. Boiucas-Dias, and M. A. T. Figueiredo, "A New TwIST: Two-Step Iterative Shrinkage/Tresholding Algorithms for Image Restoration," IEEE Transactions on Image Processing, Vol. 16, December 2007.

[12] http://www.trapview.com.

[13] L. Stankovic, S. Stankovic, I. Orovic, and M. Amin, "Robust Time Frequency Analysis based on the L-estimation and Compressive Sensing," IEEE Signal Processing Letters, vol. 20, No. 5, pp. 499-502, 2013.

[14] S. Stankovic, I. Orovic, and M. Amin, "L-statistics based Modification of Reconstruction Algorithms for Compressive Sensing in the Presence of Impulse Noise," Signal Processing, vol.93, no.11, pp. 2927-2931, 2013.

[15] D. Angelosante, G.B. Giannakis, and E. Grossi, "Compressed sensing of time-varying signals," 16th International Conference on Digital Signal Processing, pp.1-8, July 2009.

[16] I. Orovic, S. Stankovic, and M. Amin, "Compressive Sensing for Sparse Time-Frequency Representation of Nonstationary Signals in the Presence of Impulsive Noise," SPIE Defense, vol. 8717, United States, 2013.

[17] M. A. Davenport, and M. B. Wakin, "Analysis of Orthogonal Matching Pursuit Using the Restricted Isometry Property," IEEE Transactions on Information Theory, vol.56, no.9, pp. 4395-4401, September 2010.

[18] E. J. Candes, and T. Tao, "Near optimal signal recovery from random projections: universal encoding strategies?," IEEE Transactions on Information Theory 52(12), 5406-5425 (2006).

[19] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems," IEEE Journal of Selected Topics in Signal Processing, vol.1, no.4, pp.586-597, Dec. 2007.

[20] S. Stankovic, I. Orovic, and L. Stankovic, "An Automated Signal Reconstruction Method based on Analysis of Compressive Sensed Signals in Noisy Environment," Signal Processing, vol. 104, pp. 43 - 50, 2014.

[21] L. Stankovic, S. Stankovic, and M. Amin, "Missing Samples Analysis in Signals for Applications to L-estimation and Compressive Sensing," Signal Processing, 94: 401-408, 2014, doi: 10.1016/j.sigpro.2013.07.002.

[22] M. Lustig, D. L. Donoho, J. M. Santos, and J. M. Pauly, "Compressed Sensing MRI," IEEE Signal Processing Magazine, March 2008.

[23] J. Romberg, "Imaging via Compressive Sampling," IEEE Signal Processing Magazine, March 2008.

[24] I. Orovic, A. Draganic, and S. Stankovic, "Compressive Sensing as a Watermarking Attack," 21st Telecommunications Forum TELFOR 2013, Novembar, 2013.

Big Data Analytics for Communication Service Providers

Danijel Šipuš
Ericsson Nikola Tesla, Zagreb, Croatia
Danijel.Sipus@Ericsson.com

I. ABSTRACT

Today, processing huge amount of various data in various forms, incremented on daily basis is no longer challenge for big data systems. The challenge is to get insight from big data analytics so we can get business value out of it.

Customer experience is the main driver of changes in a way telecommunication operators are providing services to their customers and perceiving themselves as organization. The main obstacle for delivering good customer experience for telecommunication operators is mainly not identifying individual customer problems on time. The results are longer customer service response periods, badly targeted promotions, campaigns, and services. As a result customer experience is decreasing.

The latest discoveries in predictive big data analytics in telecommunication services enable predictions of customer experience in real time by correlating numerous parameters gathered from various network elements and getting better insight into customer's activities and habits. Using holistic approach extended customer session record is created regardless of provided service, whether the customer was watching video, downloading data, or having internet correspondence. Today it is possible to recognize customer behavioral patterns and evaluate their satisfaction with provided services.

II. INTRODUCTION

The mere mention of Big Data Analytics it can seem very complex or even astronomically huge from one standpoint. Essentially, it is the process of examining enormous data in various forms with goal to uncover hidden patterns, known/unknown correlations, market trends, customer activities, preferences, habits, influence and satisfaction by applying various data models and algorithms.

The analysis of big data mainly involves analytical methods for traditional data and big data, analytical architecture for big data, and software used for mining and analysis of big data. Traditional data analysis is using statistical methods to analyze massive data, to concentrate, extract, and refine useful data hidden in vast chaotic data collections, and to identify inherit law of the subject matter, so as to develop function of data to the greatest extent and maximize the value of data.

Today's big data analytics in communication services have provided measurable evidences it can help and even drive customer experience management and business efficiency. They are reported as a reduced churn percentage, reduced number of service request for customer care services, shorter response times when dealing with customer requests, better customer satisfaction and loyalty, increased income from marketing promotions and campaigns, and better revenues from provided services.

Future big data analytics will reshape the analytical concepts for communication service providers (CSPs) and it will encapsulate many different data models and algorithms as well as data integration components. Most likely it will be consisted of following components:

- Advanced probes and adapters for retrieving data from all network nodes in real-time
- Advanced adapters for pulling relevant customer data from traditional big data and data warehouse systems
- Real-time analytics for customer activities
- Quality indexes for sentiment analysis
- Opinion mining in real-time
- Social network analysis and customer profiling
- Enhanced customer activity (session) analysis
- Device profiling and device generated quality indicators
- Key Performance Indicators (KPIs) for service operations, marketing and customer experience
- Business rules and business policies for customer support, service operations and marketing process automation
- Business processes integration and automation

III. CSP BIG DATA ANALYTICS CAPABILITIES

Current big data analytics capabilities for CSPs are illustrated in figure 1. Today's big data analytics systems are using real-time analysis of correlated events gathered from various network nodes in order to determine most probable cause in network performance. If this process is running smoothly it can increase customer experience and business efficiency and as result it will increase revenues.

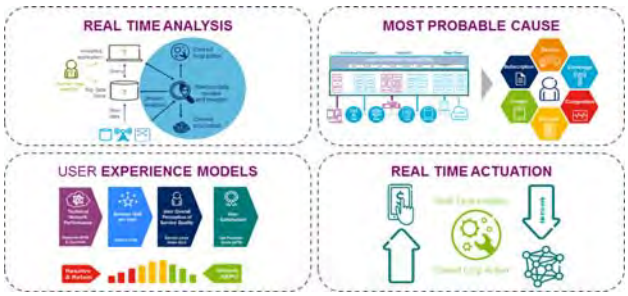


Figure 1 - Big Data Analytics current CSP capabilities

There are various problems in current big data analytics implementations and all of them share similar issues. Some of them shall be described since they should be eliminated in the next years to come. Our experience in dealing with underlying problems is growing by each day, while hardware and software performance is increasing on even faster pace.

Term "real-time" is often exaggerated and it usually means we can analyze data in near real-time periods, i.e. with delays up to 10 seconds depending on hardware infrastructure and software optimizations. The problem resides in fast event and data history correlations across multiple network nodes in order to get quick quality insight from network session analysis while dealing with thousands of events per node per second.

Most probable cause is our "best-effort" algorithm for trying to guess where could be root-cause problem in network services. It uses all accumulated experience up to this date for correlating many events which are occurring simultaneously in the network in order to assist the process of detecting the problem in network performance thus enabling business efficiency.

The biggest challenge is to evaluate customer satisfaction, i.e. sentiment. There are various user experience models and artificial intelligence algorithms trying to predict future customer behavior while trying to increase customer satisfaction. Currently there is no such solution which can guarantee data quality which will reflect real customer sentiment at given moment. However, in this document there is a concept which may be on track to be "as-good-as it gets" in discovering customer sentiment.

IV. CSP Big Data Analytics Business Areas

Currently big data analytics for CSPs is established in several business areas and it's proven its business values by enhancing overall business performance. Those areas are: customer experience management (CEM), business efficiency as a result of fast detection of possible problems in network performance, and social network analysis. Sentiment analysis will be described in the next chapter since it encapsulates all the business areas to some extent and it uses some newly developed data models and algorithms.

A. Customer Experience Management (CEM)

Achieving excellent customer experience is the main strategic goal for many communications service providers. With increasingly competitive market situation it means that everyone is struggling to keep their customers and attract someone else's. Thus it is a matter of survival to address what matters to their customers the most. This starts with knowledge, the knowledge about your customer.

Big data analytics start with collecting all the data from various data sources in various data formats and correlating them using holistic approach. Various predefined and custom made alghoritms and data models are used to get information or insight about customer activity. When all informations are put together they can bring knowledge about specific customer activites, preferences, habits, influence and satisfaction thus enable communication service provider to adjust business accordingly.

In order to efficiently improve customer experience it is critical to act as it happens in real-time. Real-time big data system can remove a lot of the complexity of making sense out of data that is there in telecom environment today as its processes data in real-time and correlates into an exclusive record per session per user. In addition to the technology challenge there is also an interpretation challenge in order to base key decisions on accurate insights. This is one of the most common problems identified in the industry.

B. Business Efficiency

Understanding customer usage, spend, activities, and problems is the key for enabling targeted value propositions to be created and achieve intimacy and closer bond with end customers.

It is imperative to improve efficiency by proactively detecting and pin-point experience incidents. Real-time, or near real-time information can off-load customer care service by using proactive measures and enable it to properly respond if needed. It can assist marketing activities to create better targeted promotions or campaings towards customers. Real-time analytics can also improve user satisfaction, increase customer loyalty and reduce churn.

C. Social Network Analysis (SNA)

Social network analysis is specific approach in data analysis to determine different aspects of social relations, i.e. subject cohesion, centrality, power, role, position, community, and subgroup. It characterizes networked structures in terms how individual customers are making relationships or interactions among them.

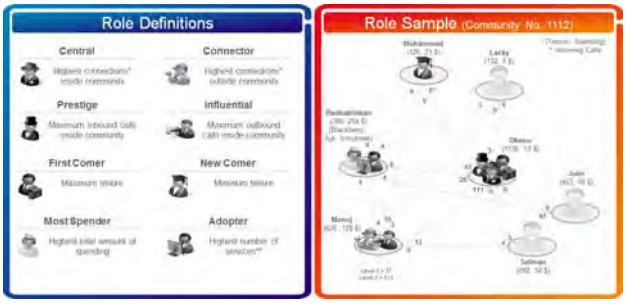


Figure 2 - Social Network Analysis (SNA)

SNA is used mostly in the market segmentation and in understanding customers influence in their communities.

Like illustrated on figure 2, one customer could generate very little network traffic or it could be low spender, but it could have big prestige inside his social network and therefore it can have big influence on other customers. Some customers may be big spenders, early adopters, new comers or those with big prestige and influence. All of them have different personal preferences and should be approached and managed using different marketing strategies.

V. AUTOMATED REAL-TIME BIG DATA ANALYTICS FOR CSPS IS THE FUTURE

Real-time big data analytics for CSPs has the new architecture which will allow communication service providers to predict customer satisfaction and enable different organizations to work more efficient reducing costs and delivering better customer service leading to increased revenues. It is still evolving and current architecture is illustrated in figure 3.

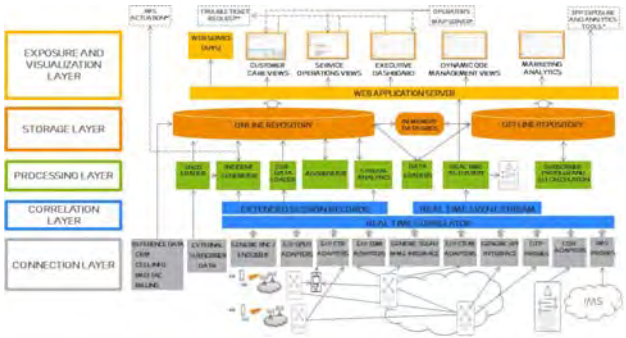


Figure 3 – Real-time Big Data Analytics Architecture

Idea behind it is to automate all business processes since many telecom operators still have manual analytical steps to support their business, for example applying an analytics unit to process data collected from networks. Real-time automated analytics enables closed-loop action, making decisions much quicker and more efficient.

New big data analytics architecture is designed to gather metrics and events from the network and through complex event correlation into the Extended Session Record (ESR) for each user and for each session.

A. Extended Session Record (ESR)

Quality indicators are calculated using the data retrieved from various network sources through network probes and adapters. Some quality indicators are derived from data originating in the subscriber's device (e.g. smartphone, tablet, pc), rather than from network nodes and probes.

Extended Session Records (ESRs) is a quality indicator which gathers data from network nodes and probes and then is further used in analysis processes (in combination with reference data).

The data provided by the network nodes and probes are being correlated and aggregated into extended session records. Each ESR contains information about one subscriber active device for defined time period. Big data system keeps track of all customers, all the time, and preserves key information.

ESR has a dedicated multi-vendor adaption layer with different adapters per interface:

- Generic KPI adapters, e.g. probes
- Generic Network Adapters
- OSS and BSS real-time adapters
- Terminal and Survey adapters
- Reference data adapters, e.g. CRM systems or other external systems
- Big Data adapters, e.g. data ware house systems

The ESR records are further used in newly developed data models and algorithms, i.e. in USP concept, sentiment analysis and subscriber (customer) profiling.

B. User-experienced Service Performance (USP)

One essential aspect of Customer Experience Management is the customer experience of service performance. Few businesses do actually have the data about how customers get and use the services. Measurements of service performance belong to the service assurance space, but traditional assurance systems, such as Service Quality Management (SQM) systems, have been focused on measuring parameters within networks. Examples of such metrics include bit error rates, signal-to-noise ratios, latency, jitter and delay.

In a CEM context these metrics are "inside-out" metrics. Inside-out metrics are still important for network and service management, but there is an increasing interest in measuring the service delivery from the customers' perspective. In a CEM context ''outside-in'' metrics are used to measure service quality as experienced by the customer; these are complex measurements and therefore more difficult to quantify, requiring clear service definitions, well-selected metrics and insights into customer expectations from the service responses.

Those demands for adequate, accurate and actionable metrics resulted with a new user-focused design implemented in new architecture. New data models encapsulate the User-experienced Service Performance (USP) concept.

The User-experienced Service Performance (USP) concept provides a service, vendor technology neutral approach to service quality measurement. This is based on 10 years of research and the USP concept includes Key Performance Indicators (KPI) that measures quality as perceived by the user per session. A System Service is a single usage of a service such as the download of a specific webpage, a voice-call or an individual video session such as watching a YouTube clip. These S-KPIs are based on extensive research by Ericsson Consumer Labs to validate the relevant threshold with real feedback from users in many different markets. During studies with real users mapping between user perception rating and user generated data has been done to find the most relevant KPI's from a user perspective. These studies tell

not only which KPIs are important but at what levels do customers start to become dissatisfied.

The USP concept is based on:

- Definitions of system and resource services, system service key performance indicators (S-KPIs) and resource service key performance indicators (R-KPIs)
- A structural relationship between system services and resource services, and between S-KPIs and R-KPIs
- A methodology to identify and prioritize S-KPIs from a user perspective
- A methodology to drill down and identify which resource services are responsible for S-KPI degradation.

The S-KPI's are independent implementation and its definition has no reference to technology choice or realization of the delivery network which makes it transparent across access-technologies. S-KPI's give objective facts of user experience per session on how easy it is to get on to a service, how to maintain the quality and how to continue to use the service.

This end user, which can be either a person or a machine, consumes the service through a terminal such as a mobile phone, a screen or a camera. Resource services, on the other hand, are logical or physical entities that combine to deliver system services, and are based on resources such as bearers, links and nodes. Separating the two allows operators to set performance requirements and fully appreciate the service quality delivered to the end user.

The performance of both system services and resource services is characterized by key performance indicators. S-KPIs and R-KPIs offer an insight into one of three quality metric groups: accessibility (the system's ability to provide a service upon the user's request); integrity (the quality of the service as perceived by the user); and retainability (the system's ability to provide a service session that is as long as the user needs it to be).

C. Service Level Index (SLI)

Service level index is a newly developed concept which calculates customer satisfaction score. It continuously captures and measures the experience for each individual user (customer). It has been validated during trials and the SLI has proven as valuable indicator of loyalty, as a measure of willingness to spend, and as a predictor of future customer behaviour. It encapsulates new data models and algorithms as well as specialized metrics for predicting customer satisfaction, deepening on our understanding of customer behavior and shifting focus on customer centricity with new insights about customer experience.

The SLI score is a specialized subscriber attribute, computed by the new big data analytics system. This score is based on key performance indicators (KPIs) and quality indicators collected for the subscriber and represent the customer experience for an individual subscriber on a scale of 0 to 10.



Figure 4 - Service Level Index (SLI)

The SLI score computation has to be highly customizable due the fact different social groups have different parameters that can affect their sentiment. Parameters used in SLI calculation have to be tuned differently in different countries, i.e. ethnic groups since same parameters have different weight factors in their individual sentiment, i.e. SLI calculation. The KPIs and other data points have to be tuned and considered in determining the SLI score and adjust their relative weightings.

SLI sentiment calculation uses inputs from Extended Session Record (ESR) stream with Key Performance Indicators (KPIs) about customer operation. A new, fairly complex mathematical calculation model has been created and only very high level details are provided here since SLI is under patent pending process.

The SLI value can be further used in business logic and it can support customer care, marketing and services operations actions and decisions.

D. Subscriber (customer) profile

New big data analytics architecture include subscriber profiling. Subscriber profiling is a very complex set of operations and it's currently performed in offline big data systems. All data stored in the near real-time repository is also moved and stored in the offline repository in a canonical format.

The profiler algorithms generate a set of profile attributes for each subscriber usually on daily basis. Raw profile attributes are computed periodically (in defined time intervals) for each subscriber, based on the external data received for the subscriber, and on reference data (e.g. CRM, cell location, etc.). These attributes are computed using a built-in set of statistical operations such as sum, average, max, and last.

Raw profile attributes are stored for calculation of final profile attributes. Final profile attributes are computed for each subscriber based on raw profile attributes or previously computed final profile attributes. Unlike raw attributes, which are computed based on a single day of subscriber data, final attributes are computed over different time periods.

Final profile attributes for subscribers are stored in the offline repository and they're used as customer activity indicators for various business purposes and real-time computation.

VI. CONCLUSION

Today's big data analytics in communication industry has taken different approaches in order to discover customers' attitude towards provided services, their level of influence in community, their behavioral model, and their level of sentiment or satisfaction.

The future big data analytics for communication service providers will most certainly shift towards real-time analytics and process automation. It will be crucial to react as soon as possible and the reaction time will differentiate CSPs how they understand their customers and how they react on their customer needs and problems.

Those who will be able to deliver quality architecture and get the most out of big data analysis using new methods and technologies will be able to properly determine customers influence in the community, their sentiment and their satisfaction with consumed products and services. It will be imperative to provide customer care (CC) and service operations departments with the quality insight into the performance of specific system services, for instance web browsing, data download, etc. It will be mandatory to identify how well or how poorly a customer is experiencing specific consumed services over a period of time.

It will be important to automate some business decisions by applying business policies and rules. For example, automatically notify customer that we're aware of his problems with watching high-definition (HD) video and that problem is with his bandwidth limitation or include him in promotional or campaign activity or offer him a new product or service.

New automated real-time big data analytics architecture is still evolving, but it can already address numerous problems mentioned in the document in the most efficient manner. Extended Session Records as quality indicators of network performance will be key input for different concepts of understanding customer behavior. User-experienced Service Performance (USP) concept will improve service quality measurement by polishing data models, underlying processes and relevant service-KPIs. USP concept will be one of the cornerstones for business process automation and future automation will be triggered upon breaching defined KPI levels using predefined business policies and processes. Current offline subscriber (customer) profiling which usually occurs on daily basis will probably compute more often since it is important input for sentiment analysis. Since data warehouse systems are evolving at rapid rate it won't be unusual if the future subscriber profiling occur on hourly base or even in matter of minutes.

Service Level Index after successful patenting will be explained in more details since its complex mathematical model and advanced algorithms of predicting customer's sentiment was successfully validated by real users in last several years.

Role of Data Analytics in Utilities Transformation

Vanesa Čačković, Željko Popović
Ericsson Nikola Tesla d.d, Zagreb, Croatia
vanesa.cackovic@ericsson.com
zeljko.popovic@ericsson.com

Abstract - The value of data is on the rise since the new technologies are collecting more data than ever before. Use of data is maturing, disconnecting itself from the IT sphere and blending with other technology areas like the Internet of Things (IoT). Consequently data analytics is gradually turning into an important part of many industries in helping with challenges faced by implementation of new technologies. Similar to for example transport systems or smart city systems, data analytics is used also by utility companies today as support for mass rollout of smart metering systems as well as for increasing the network visibility and for simplifying network management. This paper outlines how utility companies can benefit from processing existing and new data and data analytics methods.

I. INTRODUCTION

Analytics methods and big data have gone from hype to work. It is becoming a foundation for decision-making and a business priority among the world's organizations. Big data is a term describing large amounts of data, collected from a variety of sources, analyzed with the purpose of creating new intelligence and building business advantages. It brings faster results, a new depth to analytics and an unprecedented predictive power. Big data is primarily about three things: customer experience, efficiency and growth. It is used as powerful preventive mechanism as well as corrective mechanism for real-time and near real-time control and management. A data-enhanced customer experience provides a deeper understanding of users and an increased customer experience through high performance services, fast feedback and customized offerings. Data-driven efficiency means taking advantage of the information available within the organization, in order to work smarter and reduce costs. Growth, finally, is about innovations and new revenue streams sparked by big data [1].

The utilities industry is transforming. Electrical grids, water and gas pipes are examples of infrastructures that are expensive to build and maintain, and are often heavily regulated and considered 'natural monopolies'. Market conditions vary significantly between countries in terms of competitiveness and private/public ownership, but in general, the sector has particular ways of conducting business. Simply put, utility businesses have been built around energy resources and have had generation, transmission and energy retail flowing from the center in a one-way, linear build-up. Consumers have had little influence on service and pricing. Today, there is a new situation for utilities companies: market mechanisms are changing, competition is increasing and pricing is becoming increasingly complex. There are a number of factors that contribute to the market transformation and call for big data solutions. Some of the most important drivers include:

1. Regulatory pressure: Governments are under pressure to improve energy efficiency, and their actions have consequences for energy companies. Regulations like the European Union's 2012 Energy Efficiency Directive and the US Energy Policy Act of 2005 put pressure on energy companies to become more efficient and reduce their environmental impact. Utilities are also incentivized to utilize their assets in more effective ways. This is done through value-based assets management, including deployment of sensors and analysis of the financial impact of outages.

2. Consumers become producers: The relationship between utility companies and consumers is becoming increasingly complex. The combination of decentralized generation and storage options with demand side flexibility can further enable consumers to become their own suppliers and managers for (a part of) their energy needs, becoming producers and consumers and reduce their energy bills. Decentralized renewable energy generation, whether used by consumers for their own use or supplied to the system, can usefully complement centralized generation sources. Where self-consumption exhibits a good match between production and load, it can help reducing grid losses and congestion, saving network costs in the long-term that would otherwise have to be paid by consumers. If consumers generate their own electricity from onsite renewable energy systems, they consume less electricity from the grid. This will affect how network tariffs are calculated. Network tariffs should be designed in a cost-reflective and fair manner while supporting energy efficiency and the renewable energy objectives and being simple and transparent for consumers Those dramatic changes in market mechanisms push energy companies to find ways to manage distributed generation and net metering using the data analytics to help manage all new relations, power flow and business models.

3. Increase in price volatility: The increase of intermittent energy sources like wind and solar has resulted in more fluctuations in energy access over time. The volatility is elevated further with the growing influence of power markets. By introducing smart meters, consumers can adjust their energy use in relation to the

fluctuating prices. These fluctuations are now becoming normal; but handling those demands sophisticated data analysis.

4. Threats against the service delivery: Utilities are among the most critical infrastructures in society. Major incidents such as geopolitical events and natural disasters can cause severe damage to vital systems, in turn threatening markets, security, and people's lives. We all rely on energy infrastructure, and the increased level of these threats calls for improved resilience and intelligence.

5. Aging infrastructure: Around the globe, utility infrastructure is growing, and with aging technology comes increased vulnerability and a lack of interoperability with newer systems. Old infrastructure requires ongoing optimization to remain functional and cost effective [2].

II. TODAY'S UTILITIES APPROACH TO DATA

Today utilities only make use of data to a limited extend. Decisions are based on historical data from internal data sources and sometimes weather data. Usage of real time data is limited to pilot projects and small scale initiatives. Utilities often do not use all valuable data they have access to. Major analysts estimate that around 5% of existing data is being used for some specific purpose, while the rest is not processed at all. For example, a smart meter can be seen as a sensor that can help to visualize "the health" of power grid and communication. Only small percent of data one smart meter is capable of storing and sending is being fetched and that is primarily for billing purposes. The fact that a utility collects data does not necessarily means that it is making use of the data. Everyday operation and maintenance routines often mean handling of various excel files coming from different data sources and is not being done in real-time manner. Databases are also often updated by hand.

Utilities approach can be described as "one problem/one system" approach which means that single problems are handled using single dedicated applications. Links between systems are loose or missing. This means that many utilities still lacks synchronized databases and use many redundant databases (data silos). However, efforts to integrate those separate data systems can be already seen for most of energy companies.

Today's utilities usually have some business intelligence tools in place and have insight into some areas of their operation. Those tools used by utilities are mostly standard reporting tools and obsolete forecasting models.

However, utilities are aware of the changes in the energy landscape and available data sources. They see that conversion of operation and information technologies brings massive volume of data. Some utilities have started pilot projects to investigate the use cases for big data other than accurate billing. It is clear that utilities will have to invest in new technologies for data management to unlock true potential of new data sources. They need to re-build their data architecture model and redefine data use cases.

Utilities have always dealt with data management (for example traditional Supervisory Control and Data Acquisition (SCADA) systems [10], but the volume and velocity of data in smart grid is something which they are not prepared for. The main difficulty that they will have to overcome to start implementing big data technologies refers to expertise, infrastructure, and perception of risk (business value recognition).

Handling a tremendous amount of data from the grid, smart homes, in-house data bases or social networks requires big data processors requires enormous storage capacity and specialized ICT services. Most utilities do not have that. They also complain about the shortage of people specialized both in energy industry and big data.

Business model transformation will mean step by step adoption of new data management technologies for more informed decision making. The transformation will be based on 3 pillars:

- Emergence and adoption of new data sources
- Integration of new types of data
- Changes in data consumption model

With the emergence of smart grids, utilities will start exploiting new data sources (grid sensors, smart meters, and electric vehicles) to optimize their business and provide better customer services. Utilities will be also forced to improve customer interaction building on social media and various web data.

Utilities are used to dealing with simple and structured data limited in volume and velocity. Collection, storage, and processing of M2M data from sensors and meters will pose a challenge for energy companies who have no in-home big data capabilities. However, the greatest difficulty will be to integrate those structured data with unstructured one which is totally new to utilities [3], [4].

Business model transformation will be based on connecting traditional and new data sources, tapping new data, and integrating smart meter, smart grid, transaction, and geospatial data as shown on Figure 1:

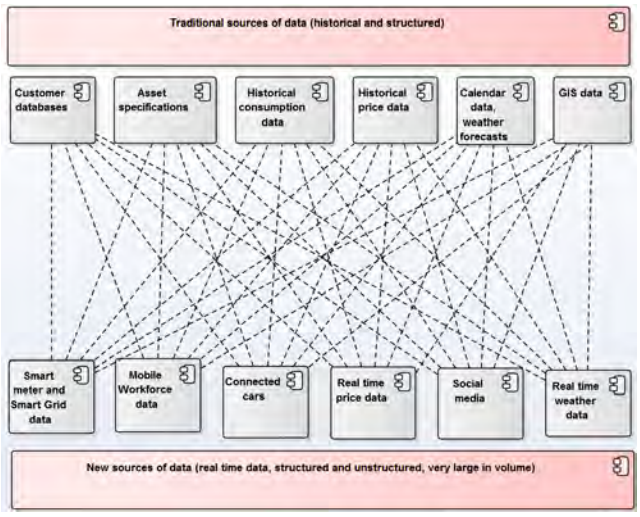


Figure 1 Data-business model transformation

Big data in the utilities sector will be dependent on a much larger number of nodes than for example in telecoms, demanding lots of machine power and powerful analytic models. Smart meters are not the only components of smart grids that can generate useful data for utilities companies.

In the transformation, period utilities will not be using the full potential of available data. Utilities will leverage data potential only to some extent, mainly using just smart meter data on commercial scale. Starting from reporting, billing and settlement, they will test new technologies and use cases. After choosing the most successful ones, they will start adding new sources of data to their data management systems and build up their analytics capabilities.

The development pace of data management systems will differ by country, energy market participants and technology.

Utilities will start with analyzing structured data. Many pilot projects, especially involving smart meter data have already been kicked off. Utilities will lag behind with adoption of analytical tools for processing unstructured data. It seems that cost approach will be the biggest challenge which will have to be overcome to enable moving to a truly data-driven model.

Cloud computing may be the answer for utilities hesitancy in deploying advanced analytics solutions. Some utilities have already move part of their processes to cloud environment, however, a majority of those cases are non-mission critical processes.

III. METER DATA MANAGEMENT AND ANALYTICS

Meter Data Analytics (MDA) is an upcoming trend where Meter Data Management (MDM) vendors or ICT vendors are trying to expand their offerings into in order to provide utilities add-on capabilities for data processing, analyzing, and reporting trends that add value rather than just uploading data. MDA enables utilities to avoid forecast failures rather than respond to them.

MDM is much more than metering. It forms a critical component that holds a crucial stance in deriving benefits from advanced metering. These systems consist of a database repository, automating and streamlining the complex process of collection of large amounts of interval-based meter data from various collection technologies and delivering the same in an appropriate format to a utility billing system [18].

New methods for storing, managing, and integrating new data into existing systems are possible with MDM, which is a flexible and scalable solution of Automated Metering Infrastructure. Time-of-usage of energy in real-time can be measured with the help of smart meters, but it is also critical to understand optimization, prediction, and forecasting of the transactional value of data. Business challenges faced by utilities can be solved with proper data analysis.

With an effective back-office and enterprise level MDM integration, standardizing applications and modifying key business processes, utilities can achieve operational efficiencies, lower costs, and improve customer satisfaction.

Major MDM challenges include utilities handling increased volume of data flooding into their systems, storage of data, data analysis and forecasting, information security, future proof technology, and software. High load transaction handling is a major challenge. Hence, careful planning and selection of appropriate communication technologies and MDM vendors are critical for realizing the potential of Advanced Metering Infrastructure (AMI) [9].

With increasing large-scale AMI installations, tracking physical distribution and location of new meters has become critical. Hence, it becomes crucial for utilities to determine and decide the type of data their MDM system will handle. Better data mining capabilities, storage, and retention becomes an issue. Regulators would require the utilities to provide information over an extended period. Such a huge influx of data can result in overloading the system.

The utilities implement meter data management systems for optimizing the meter-to-cash process. It is also important for the utilities to offer broader analytics and convert data into actionable information. It is important for utilities to gain more insight into the grid by fully capitalizing on the value of smart meters and capturing this information to operate the grid in a better way. Although master records of meter data or billing systems traditionally exist, the challenge facing utilities is when data needs to be accessed by various departments and users such as interval data needed for outage management, asset sizing, or load forecasting rather than monthly totals. Such high volumes of data need to be managed and manipulated.

Key players active in the MDM systems and smart metering applications space include enterprise software industry leaders such as Oracle and SAP, leading meter and energy infrastructure vendors such as Landis+Gyr, Iskraemeco, Elster, Siemens, Schneider Electric and a number of ICT and software specialists focused on the energy market [6].

IV. BALTIC UTILITY TRANSFORMATION CASE – IMPLEMENTING ANALYTICS

Utility company is one of the largest electricity distribution companies on the Baltic countries, owning over 640.000 metering points (residential and industrial), 61.000 km of power lines and 22,000 substations. Main motivation for transformation the utility was the fact that all electricity meters need to be remotely readable by 2020 according to EU directive, opening of the electricity market 01.01.2013 (for trading period of 1h) and aging equipment with high maintenance cost. Two primary goals set for the transformation project were to improve the quality of the electricity supply and to improve the quality of the service [5], [18].

The utility company started an eight-year long transformation project in year 2011 within which integration of new Meter Data Management System was

included as well as rollout of smart meters and integration of Analytics platform with existing systems.

Smart meters help modernize the power grid by providing the ability to remotely monitor and transfer energy consumption and power quality information. This allows utilities to manage the distribution network more efficiently. By collecting near real-time energy consumption information, Smart metering infrastructures empower suppliers to:

- Reduce OPEX by managing manual operations (meter reading, service disconnection,..) remotely
- Improve forecasting and smooth power-consumption
- Improve customer service through profiling and segmentation
- Reduce energy losses
- Simplify micro-generation monitoring
- Increase end user satisfaction

Smart metering also benefits end customers in the form of energy and cost savings, reduced carbon emissions, improved service reliability and improved customer service and responsiveness. To make all of these benefits a reality, utilities need a secure and flexible two-way communications infrastructure to connect and communicate with smart meters.

Collection and transmission of energy consumption data is a continuous process that is done automatically. As the network grows, efficient operation of Smart meter communication network becomes a critical process. Ability to address critical issues in network and also capability to address problems before a failure require innovative tools to support communication personal in their daily work. Analytics platform is a solution for enabling visualization and context development for smart metering data to support power grid management [11], [12], [13], [14], [15], [16].

Within the project, the Analytics platform (Smart Meter Manager, SMM) was integrated with of the existing system as shown in Figure 2.

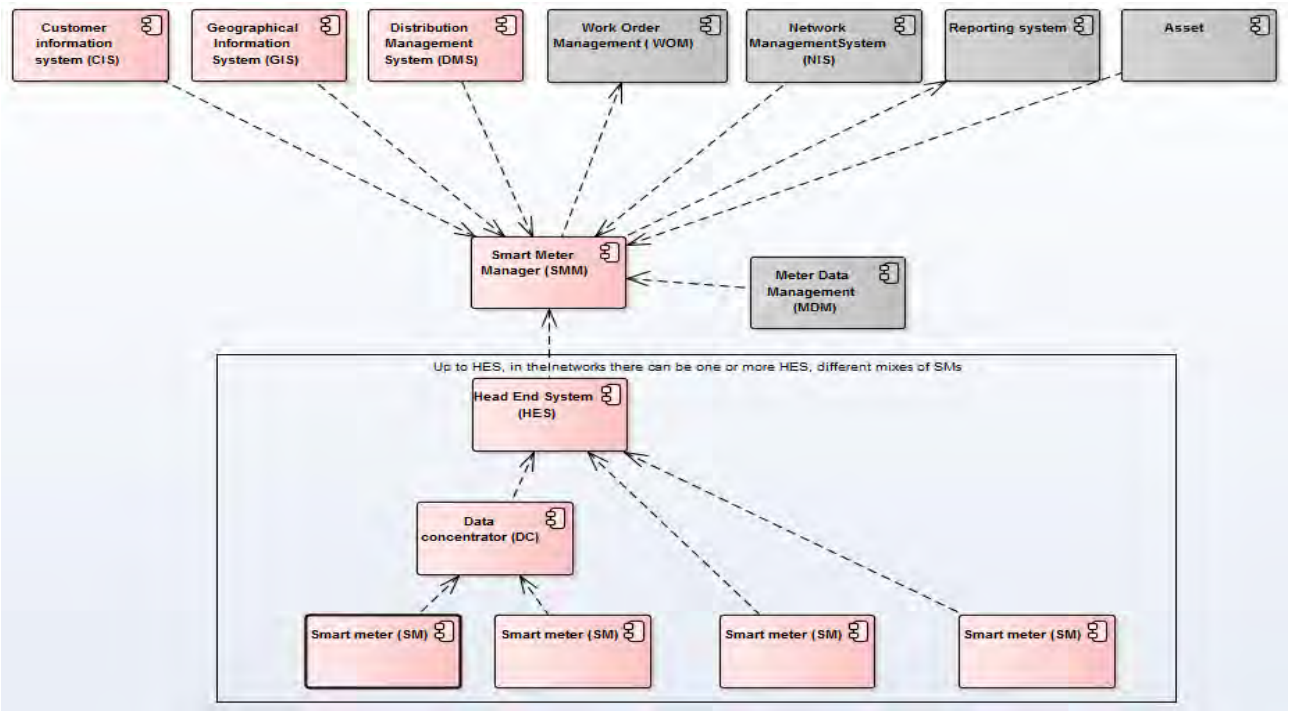


Figure 2 Integration of the analytics platform with existing systems

One of the keys to long-term return on investment on smart grid deployment is managing the big data assets associated with the smart grid roll out. This data comes from an increasing number of new and enhanced internal enabling technologies including Advanced Metering Reading (AMR), Meter Data Management (MDM), Distribution Management Systems (DMS), Asset Management Systems etc. Big data assets are also coming from third-party sources such as off-grid data sets coming from various sources such as pricing details for demand response and forecasting information for renewable energy. Other big data comes from additional software and service vendors for asset management of additional smart devices and associated operating systems.

Analytics platform provides functions to support different applications with data storage, processing, querying, machine learning, visualization and other services. The Analytics Platform gathers the data from various data sources in online mode and does the modelling and processing of data in order to act on newly processed information. During the implementation over 50 candidates for analysis and implementation were identified (so called use cases). Today's use cases are divided into three groups: power network operations, smart meter operations and fraud&loss related use cases.

Examples of the use cases are:

1. Use Case (Communication supervision): Loss of meter data during the collection process may cause issues with billing, as well as disturbances in the distributed energy generation. Upon a data collection failure, it is possible to quickly analyze possible causes by drilling down in the communications network, from a geographical network map down to an individual meter in an apartment house. Regardless where the information come from (multiple systems), it is presented in the same user interface which uses latest usability principles to increase user efficiency and effectiveness. The majority of the faults can be identified and solved either with remote reconfiguration or by deploying field personnel by issuing a work order. Network planning issues are found and can be corrected during normal operation. The method can be used to discover causes for income drops, users complaints, users churn, loss of control of distributed generation figures etc. There might even be regulations which put requirement on fault identification and fault repair which forces the utility company to have an efficient tool to localize and correct faults within a predefined period implying that every saved second means saved money.
2. Use Case (Power Quality Visualization): Power quality determines the fitness of electric power to consumer devices. Without the proper power, an electrical device (or load) may malfunction, fail prematurely or not operate at all. The complexity of the system to move electric energy from the point of production to point of consumption combined with weather variation, generation, demand and other factors provide many opportunities for the quality of supply to be compromised. Smart meters provide simple power quality measures. Analysis of this information can aid in the detection of anomalies and other emerging adverse conditions in the distribution network, allowing the utility to proactively initiate equipment maintenance, reducing the likelihood of an outage. Power quality monitoring helps in preventive system repairs/ maintenance by identifying areas of poor quality and subsequently acting on those, additionally life of assets could be extended and unplanned service interruptions could be avoided via predictive and pro-active maintenance
3. Use Case (Energy Loss Detection): Fraud amongst the subscribers is always a risk, including loss of income to unbalance in the power grid. Aggregation and correlation of multiple metering points (e.g. substation vs. aggregated feeder line consumption) makes it possible for the utility to understand magnitude of, and identify possible fraud activity. All information available from different sources are gathered in the same user interface to optimize the user experience and increase the efficiency. An alarm as well as a work order can be issued upon detected (suspected) fraud. The method can be used to reduce the loss of income but also prohibit unwanted unbalance in the power grid. Additionally it may be a safety issue, someone has tampered with high current equipment, possibly lethal.
4. Use Case (Outage management visualization): Smart meter events such as 'last gasp' and 'power restore' that provide meter off/on status can be used for improving outage management. Being real time or near real time events, these have an obvious advantage over outage information coming from customers and field staff. Examples of scenarios which could be visualized: Momentary outages and restoration-related events; communication and network interface issue-related events; Events due to planned outages, outages at the lateral, feeder or transformer level, customer disconnects, etc. Effective events processing & analytics engine shall be used to detect outages at the right device level, create proactive tickets, as well as 'power restore' to identify nested outages after large-scale outage restoration. Aim is to help make outages shorter and restoration more reliable. It can also reduce the customer complaints if information about the outage can be handled and in some cases distributed, even before the customers notices the outage.

It has been seen that it would be beneficial to for use this type of Analytics platform also in other parts of utility organizations such as call center and general performance monitoring of the company [5], [7].

As an outcome of the integrating the Analytics platform with existing systems and various data streams, it is possible to visualize the network, get communication quality, drill down to root cause, read configuration of devices and have near real time data available. The platform enables proactive and/or fast actions in order to secure needed network quality and meter data availability, supervise network rollout with higher quality, handling big number or smart meters and reduce operating costs. The platform provides a way to collect, aggregate and analyze data in real-time, allowing utilities companies to prevent issues and manage the customer experience.

V. CONCLUSION

Big data makes it possible to collect real-time data from existing grid elements to gain better control of the infrastructural assets. Ongoing monitoring of resources improves proactive maintenance actions and operations. Data from smart meters and other smart assets helps us to understand asset usage and provides a means of asset lifecycle management. The ability to anticipate issues that could interrupt the provision of service is a value that big data brings to both customer experience and operations. Like many other aspects of running a utilities operation, fault management is an area that is more difficult to handle today without big data solutions. Predictive

modeling of events that cause outages can be deployed to trigger automated mitigating actions. This is an effective tool to reduce costs and improve customer experience.

Energy grids were not originally designed to cater for the volatility and complexity created by the influence of intermittent energy forms. Nor were they built to withstand the dawn of an energy marketplace, or the growing needs of a society that is increasingly dependent on digital tools. Nevertheless, they still stand, despite ongoing transformations. The development of data-driven ways of managing operations will be necessary for them to remain the backbone of infrastructure going into the future. Big data helps companies to develop more accurate forecasting models and distribute capacity and manage assets according to shifting needs. However, utility companies are under pressure to keep up with the pace of transformation, and to improve quality of service (QoS) in the midst of this change. Big data helps them with a new generation of communication solutions where QoS can be controlled. In this setting, processes are homogenized and pricing models ensure the most cost-effective service possible.

LITERATURA

[1] Industry Transformation – Horizon Scan: ICT & the Future of Utilities; http://www.ericsson.com/news/141211-horizon-scan-ict-and-the-future-of-utilities_244099435_c

[2] Actionable intelligence Transforming Utilities; Ericsson; http://www.ericsson.com/res/docs/2015/actionable-intelligence-transforming-utilities.pdf

[3] Smart metering in Europe; Tobias Ryberg, Berg insight research team; www.berginsight.com

[4] Data-driven Utilities: Business Intelligence and Analytics to become the Top Investment Area for Energy Companies; Frost&Sullivan; October 2013; http://www.frost.com/prod/servlet/press-release.pag?docid=288057433

[5] http://www.ericsson.com/industry-transformation/energizing-estonia/

[6] Rob Lockhart: "Smart Grid Data Analytics: a New Approach for Utilities"; Gigaom Research, 2015

[7] http://www.metering.com/smart-meter-europe-estonia-utility-deploys-half-of-630k-target/

[8] GTM Research: "The soft grid 2013-2020: Big Data & Utility Analytics for Smart Grid"

[9] https://www.smartgrid.gov/recovery_act/deployment_status/sdgp_ami_systems.html

http://whatis.techtarget.com/definition/SCADA-supervisory-control-and-data-acquisition

[10] J. Kwac, J. Flora,and R. Rajagopal, "Household Energy Consumption Segmentation Using Hourly Data," IEEE Transactions on Smart Grid, vol. 5, no. 1, January 2014, pp. 420-430.

[11] G. Chicco, "Overview and Performance Assessment of the Clustering Methods for Electrical Load Pattern Grouping," Energy, vol. 42, no. 1, June 2012, pp. 68-80.

[12] L. Hernandez, C. Baladron, J. Aguiar, B. Carro, A. Sanchez-Esguevillas, J. Lloret, and J. Massana, "A Survey on Electric Power Demand Forecasting: Future Trends in Smart Grids, Microgrids and Smart Buildings," IEEE Communications Surveys & Tutorials, vol. 16, no. 3, Third Quarter 2014, pp. 1460-1495.

[13] I. S.Bayram, G. Michailidis, M. Devetsikiotis, and F. Granelli,"Electric Power Allocation in a Network of Fast Charging Stations," IEEE Journal on Selected Areas in Communications, vol. 31, no. 7, July 2013, pp. 1235-1246.

[14] A.G. Phadke and J.S. Thorp, "Synchronized Phasor Measurements and Their Applications (Power Electronics and Power Systems)," Springer, 2008, ISBN 978-0-3877-6535-8.

[15] Y. Zhang, P. Markham, T. Xia, L. Chen, Y. Ye, Z. Wu, Z. Yuan,L. Wang,J. Bank, J. Burgett, R. W. Conners, and Y. Liu, "Wide-Area Frequency Monitoring Network (FNET) Architecture and Applications," IEEE Transactions on Smart Grid, vol. 1, no. 2, September 2010, pp. 159-167.

[16] G. Giannakis, V.Kekatos, N. Gatsis, S. J. Kim, H. Zhu, and B.Wollenberg, "Monitoring and Optimization for Power Grids: A Signal Processing Perspective," IEEE Signal Processing Magazine, vol. 30, no. 5, September 2013, pp. 107-128.

[17] Frost & Sullivan "Global Advanced Metering Infrastructure (AMI) Market Strong Global Growth as the Smart Market Accelerates", Novemner 2013

[18] http://ec.europa.eu/competition/sectors/energy/electricity/electricity_en.html

Using MEAN stack for development of GUI in real-time big data architecture

Marko Štajcer, Marko Štajcer i Dražen Oreščanin
Poslovna Inteligencija d.o.o., Zagreb, Croatia
{marko.stajcer, stajcer.marko, drazen.orescanin}@inteligencija.com

Abstract – There are several key criteria for successful software development branch and one of them is architecture. We describe in detail architecture, design and development of graphic user interface for real-time fraud detection system in the telecom industry. Our focus is on a web application architecture, which includes data model component, technical infrastructure component, and components that interact or are associated in any way with the system (users, third party components).

I. INTRODUCTION

Complex event processing, or CEP, is event processing that combines data from multiple sources [1] to infer events or patterns that suggest more complicated circumstances. The goal of complex event processing is to identify meaningful events (such as opportunities or threats) [2] and respond to them as quickly as possible. Many of today's big data technologies were built on the idea of batch processing human generated unstructured and predominantly persistent data, stored in distributed file systems. However, data volumes generated from machine-to-machine interactions (M2M) surpass by far the amount of data generated by humans. M2M data streams must be processed in real-time or near real-time, they are predominantly transient and naturally distributed, and are much more structured in nature. FERARI is a collaborative project within the European Commission's FP7 ICT Work Programme, whose goal is to pave the way for efficient, real-time big data technologies of the future. It will enable business users to express complex analytics tasks through a high-level declarative language that supports distributed complex event processing and sophisticated machine learning operators as an integral part of the system architecture. Effective, real-time execution at scale will be achieved by making the sensor layer a first-class citizen in distributed streaming architectures and leveraging in-situ data processing as a first (and, in the long run, the only realistic) choice for realizing planetary-scale big data systems.

To give a new potential user a good idea of the capabilities provided by the FERARI platform, we created an example application. This application demonstrates a basic mobile fraud detection scenario, modeled after our mobile fraud use case scenario. Nevertheless, the prototype is built upon the FERARI open source architecture and strongly interacts with it. We start with the system architecture that we are using for the application development. In the next section we continue with technology that we are using including their advantages in modern web development and with the description of the graphic user interface.

II. ARCHITECTURAL INTEGRATION OF FRAUD MINING PROTOTYPE

A. System architecture

Our focus is on the web application architecture, which includes the following:

- data model component
- technical infrastructure component
- components that interact or are associated in any way with the system (users, third party components)

To build a stable ecosystem for web application, we have to consider the needs of the functional graphical user interface:

- collecting data from data source
- handling data
- sending data to the front-end
- getting data and displaying on the front-end

In order to achieve the previously mentioned, we need a solution, or more, to develop an application with a structure broken down to database, server and client. There are several benefits we can achieve through breaking down the whole application into smaller and modular components:

- easier application maintenance and expansion
- reducing the number of bugs
- reusable components (using individual components in many different applications), parallel programming
- minimal impact when replacing the technology
- security significance. [3]

Because of that, the application is built in three-layer architecture, as shown in Figure 1.

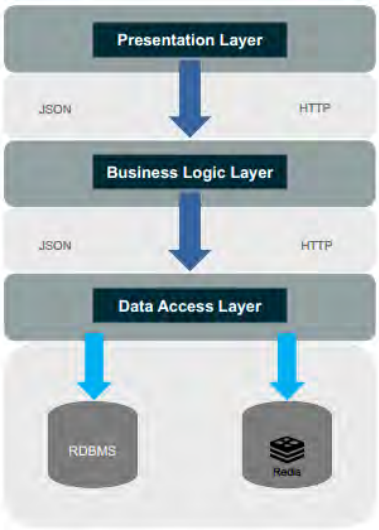


Figure 1. The application architecture

B. *Presentation layer*

The presentation layer gives a user the possibility to interact with the application through elements such as navigation, buttons, pop-up windows and other elements for managing user interaction which have been implemented by user interface components and presentation logic.

User interface elements/components actually include information which we want to show in a particular form regarding specification of the user interface for the application. Some of these elements are used to display information to the user and to accept user input. In some architectural patterns there are UI components called views, whose role is to show the presentation to the user, regarding underlying data and the application logic. The user interface components can be extended with logic which is common in the user interface implementation. The extension of the UI components should not be exaggerated with previously mentioned logic.

The application code which cares about the behavior and structure of the application (user interface code) is placed in the presentation logic. The behavior can be defined through units called user stories, and we have logic components for their implementation. There are two more roles of the logic components:

- controlling the user's interactions with the underlying logic and state of the application
- formatting data from the business layer for the UI components.

C. *Application business logic layer*

The core application logic and the bridge between the presentation and the data layer represent the business logic layer. When we say bridge, it means that it gets data from the data layer and sends the results to the presentation layer. Within the context of the FERARI real-time processing, the actual business logic is expressed in terms of CEP expressions that are evaluated in real-time, in a distributed manner. In the considered use case, the FERARI processing system is generating alerts that relate to the detection of possibly fraudulent events. This "business logic" is of course outside of the GUI application. However, the end-users requested to have the possibility to store these alarms, work on the history of alarms, and investigate fraudulent events in an interactive manner, after the corresponding alarm has been raised. For this purpose, we added a data access layer for storing alarms and short term history of events. The business logic for working on stored alarms and the short term history is implemented in the GUI application.

D. *Data access layer*

The data access layer (the DAL) is responsible for tasks related to reading and writing data, which uses technologies and programs (e.g., SQL, JSON parser, ODBC, etc.) for the data access logic implementation. There is, the presentation layer which defines the user interface, and the business layer in which the behavior of the application is defined, and at the end there is the data access layer for data flow control. Each of these layers communicates only with layer above or below it and because of that, there is no connection between data and presentation layer.

E. *MEAN stack*

MEAN is a free and open source powerful JavaScript framework which simplifies and accelerates full stack web application development. MEAN stands for:

- MongoDB – very popular NoSQL database solution designed with web application in mind the user's interactions with the underlying logic and state of the application
- Express – web server framework for Node.js
- AngularJS – frontend JavaScript MVC framework for single-page application development
- Node.js – JavaScript server platform built on V8 JavaScript engine

The most interesting fact is that the MEAN Stack is using only one programming language (JavaScript) on each layer of the application. We looked for a modular solution that will give us flexibility needed for the project. We have chosen MEAN, but we are not using original MEAN Stack with the mentioned technology. We have replaced MongoDB with MySQL and there is a M*EAN (*MySQL, Express, AngularJS, Node.js) stack. [4]

F. *Single page applications — modern web applications*

In web applications development, there are different approaches based on business logic placement:

- multi-page application
- single-page application (SPA).

Multi-page application: Back-end side contains most of the business logic with a little bit of development on the client-side (JavaScript). The example where the most of the business logic is implemented on back-end side is the Ruby on Rails applications (MVC architecture). In this case, when user changes a state through the web page navigation, every change sends a HTTP request.

Single-page application (SPA): Client side contains most of the business logic which is often developed by front-end MVC architecture. In this case, there is a front-end router whose role is to deliver AJAX requests and make API calls to the back end. [6]

SPA is an application that keeps the web page from reloading during use. The purpose of the application is to help the user to complete a task, for example; write a document, or administer a web server. SPA can be considered a thick client loaded from the web server. SPA delivers a desktop application in the browser; which results in a highly responsive experience, and a bigger number of satisfied users. There is a difference between SPA clients and traditional web pages. The replacement of a traditional website with a SPA causes the change of the entire software; from the database server to the HTML templating. The companies that aimed at a successful transition from the traditional website to the SPA, have understood that the change of the old structures and practices is necessary, so they decided to refocus the discipline, the engineering and the testing on the client. The server does not lose its importance, but focuses on providing JSON data services. Here are some important advantages of JavaScript SPAs over Java and Flash:

- No plugin required – There is no need for plugin when users want to access the application, and there is also no need for the developers to worry about a separate security model, which facilitates the use.
- Less bloat – An SPA using JavaScript and HTML usually uses fewer resources and does not require an additional run-time environment, which is necessary when it comes to plugin.
- One client language – Web architects and most developers are obliged to be familiar with many languages and data formats – HTML, CSS, JSON, XML, JavaScript, SQL, PHP, Java, Ruby, Perl, and so on. There is no need for writing applets in Java, or Flash applications in ActionScript, when JavaScript is already in use elsewhere on our pages. By using a single programming language for everything on the client, we significantly reduce complexity.
- A more fluid and interactive page – A Flash or Java application has already been seen many times on a web page. There is often a big difference in details between the HTML elements that surround it, and the application itself, which is usually displayed somewhere in the box. The said implies that some elements like the graphical widgets, the right click, and the sounds are very much different, while the interaction with the rest of the page is limited. With a JavaScript SPA, the entire browser window is the application interface. [11]

With the maturation of JavaScript [11], most of its weaknesses have been either fixed or mitigated and, causing the increase in value of its advantages:

- The web browser is the world's most widely used application – The majority of people leave the browser window open throughout the day, to make sure they can use it whenever they want. The access to a JavaScript application is only a bookmark click away.
- JavaScript in the browser is one of the world's most widely distributed execution environments. In the last 3 years, the increase in the activation of android and iOS mobile phones caused a significant increase in shipping of robust JavaScript implementations on phones, tablets, laptops, and desktop computers around the world.
- JavaScript deployment is trivial – By hosting it on a HTTP server, we could make the JavaScript application available for more than a billion web users.
- JavaScript is useful for cross-platform development – Now SPAs can be created using Windows, Mac OS X, or Linux, and a single application can be deployed not only to all desktop machines but also to tablets and smart phones. Mature libraries such as jQuery and PhoneGap that smooth over inconsistencies and the convergence of the implementations of standards across browsers have contributed to the mentioned.
- JavaScript has become surprisingly fast and can, at times, rival compiled languages – For its speed up, we can thank to an ongoing and heated competition between Mozilla Firefox, Google Chrome, Opera, and Microsoft. Modern JavaScript implementations have experienced advanced optimizations such as JIT compilation to native machine code, branch prediction, type-inference, and multi-threading.
- The evolution of JavaScript has enabled it to include advanced features – These features include the JSON native object, native jQuery-style selectors, and more consistent AJAX capabilities. Mature libraries like Strophe and Socket.IO have made push messaging far easier.
- There has been a significant advance in HTML5, SVG, and CSS3 standards and support – thanks to these advancements, the rendering of pixel-perfect graphics that can rival the speed and quality produced by Java or Flash, is allowed.
- JavaScript can be used throughout a web project – Now the excellent Node.js web server, and data stores such as CouchDB or MongoDB can be used, both of which communicate in JSON. Even libraries can be shared between the server and the browser.

- The power of the desktop, laptop, and even mobile devices has increased – The omnipresence of multi-core processors and gigabytes of RAM allows the processing that used to be accomplished on the server to be distributed to the client browsers.

The popularity of JavaScript SPAs has increased due to these advantages, and the demand for experienced JavaScript developers and architects has also increased. Applications that were previously developed for many operating systems (or for Java or Flash) are nowadays delivered as a single JavaScript SPA. Startups have chosen Node.js as the web server of choice, while JavaScript and PhoneGap are used by mobile applications developers to create "native" applications for multiple mobile platforms using a single code base. [11]

G. *MVC design pattern*

For better organization of the web application architecture there is a widely adopted structure Model View Controller (MVC), popular architectural pattern whose idea is to separate code into three components of the application.

- **Model** stores data retrieved from the controller and displayed in the view
- **View** represents a presentation based on the data in the model
- **Controller** contains business logic and defines the behavior of the application

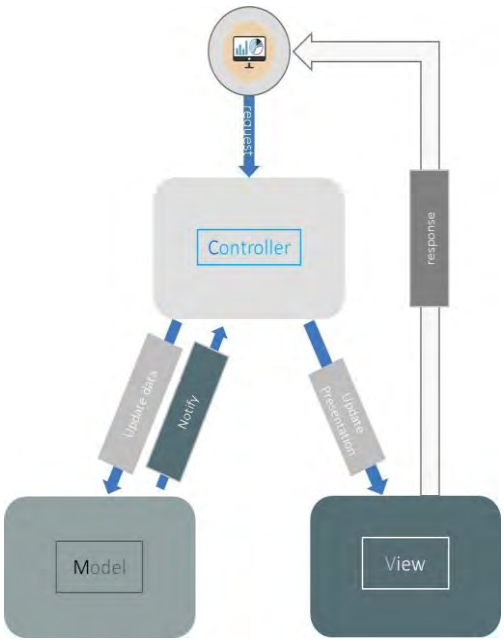


Figure 2. MVC design pattern

The controller is activated when end user interacts with the application. At this moment, the data in the model are changed by controller and the view can get these data to display them to the end user. The view can make an update when it knows that a change has occurred by notification from the model. [9]

AngularJS is an open source JavaScript library, whose development and maintenance is Google's responsibility. Some of the largest and most complex web apps around have been using it. The MVC pattern is followed by the AngularJS applications, while the development process itself relies on a wider range of building blocks. There are the headline building blocks - the model, the views, and the controllers - but there are also numerous moving parts in an AngularJS app, as well as modules, directives, filters, factories, and services. [8]

H. *Application flow*

We are going more in detail with the architecture of the application and we will explain the connection among technologies used in separate layers. With separation of our application we achieve the possibility that each layer can handle specific tasks. On the back-end side there is a server (node) whose role is to handle data layer and provide information to the front-end on the client side (angular). As shown in Figure 3, on the front-end we have angular which makes a request (API call) to the server. Our server stands alone with API from which angular can get all data. In order to access a database we need to have the application connected to it. When the node has access, we can build routes and define API and at the end JSON information is provided to the angular. The whole concept (shown in Figure 3) allows us to separate the front-end part of the application from the data (API).

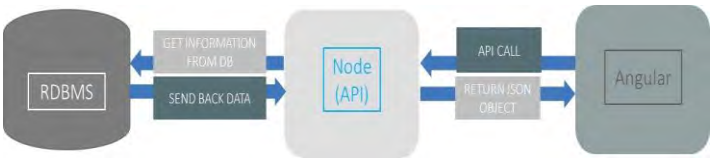


Figure 3. Application flow

I. *Realtime architecture*

If we want to show some data to the front-end in real-time we need a solution which includes pushing the same data from the back-end to the front-end side when something new (fraud event in our case) happens. In this case, the most of the logic is on the front-end side, because the client has to retrieve and load data dynamically, and change the user interface. Figure 4 depicts the overview on the real time push architecture we made, based on the FERARI open source architecture.

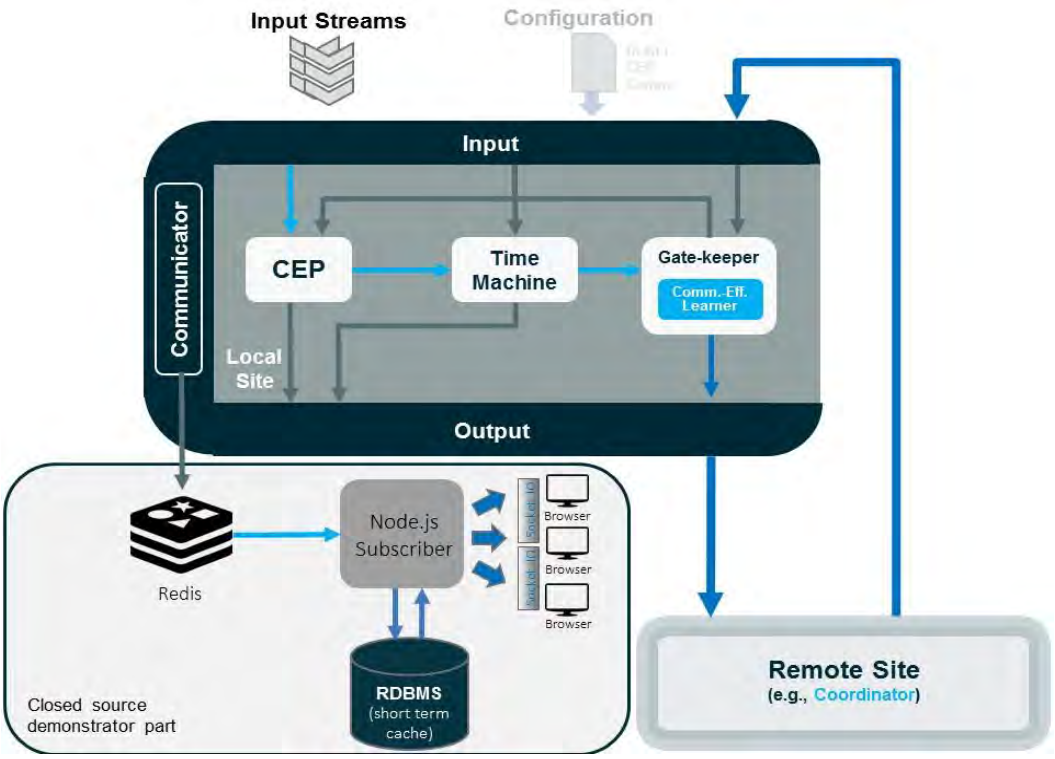


Figure 4. Realtime architecture

Inside of the red container the story begins with external data stream source which emits data tuples to the DetailOut bolt. DetailOut bolt's role is to cache a limited amount of detailed data in RDBMS (short term history). This information is needed for the purpose of forensic data analysis and validation of this specific use case. We want to store data only for this specific cases to make validation of detected fraud events and previously mentioned action will provide better insight into events. Apart from those details only derived events are stored until they are processed by the fraud mining expert. This caching mechanism for details is necessary as it is impossible for the streaming real time implementation to store all call details. Nevertheless, the capability to process all incoming stream items is not influenced by this reduction.

Next we have the components of the FERARI architecture: CEP (complex event processing) engine that is helping with the development of complex event processing applications. It is used for real-time event data analyzing and generating immediate insight. CEP (including its technology and implementations) also includes enabling instant response to changing conditions and providing resources to define and maintain the event processing logic of the application. Flexibility of using CEP comes from its runtime that can meet functional and non-functional requirements without affecting application performance and there is also no need for the developers and system managers to worry about the issue, which facilitates the use. Complex event processing component is built on and extends the IBM Proactive Technology Online (PROTON) in the FERARI project. The CEP is responsible for filtering and combining data from external sources and at the end for serving meaningful results. These results are processed in the FraudEvent which publishes data to a dedicated channel in Redis, as the event occurs. As being subscribed to the same Redis channel, the Node.js receives all the events as they occur, as well as all the other subscribers subscribed to the Redis channel. In order to enable communication and data streams between the server and the browser we have to implement something that provides real-time updates. For Node.js there is a Socket.IO module which has communication methods both on the server and the client, and as the Figure 4 shows, we are using these methods to enable communication on both sides. Node.js is subscribed to Redis channel and it listens to the message [1]. When the message comes, Socket.IO allows the sending of the message back to the browser clients that are connected to the server and subscribed to the same channel. Browser will display the content of the message when FraudEvent bolt publishes the message using the communicator. In the current state of the prototype we use Redis to provide the actual communication channel. The benefits of that mechanism are recognized in connecting and providing subscribe from every third party system without any additional knowledge or change in the rest of the technology used in the architecture in order to receive a message. It is enough to have access to the Redis channel and subscribe to the same. If compared to a traditional messaging system this can be a disadvantage,

but it is advantageous where data has importance in real time and need not to be stored.

Redis is a NoSQL fast in-memory key-value store which is used when performance and scaling are important. It is currently used in the prototype as underlying communication layer. Access to this layer is modularized via the FERARI interfaces such that it can be replaced in a later stage with the advanced communication layer. For the purpose of caching recent fraud events and the limited amount of corresponding detail records, the in-memory storage system of Redis is well suited. Besides the mentioned, it can do many different things, for example: implementing a pub/sub system. As the keys can contain strings, lists, sets, hashes and other data structures, Redis is also known as a data structure server.

Pub/sub stands for publish-subscribe, a pattern where messages are not sent directly to specific receivers. The publishers send messages to different channels, and the subscribers receive these messages only if they are listening to a given channel. The pub/sub pattern is supported by Redis, responsible for the commands that allow the publishing of messages and the subscription to the channels. [7]

Ever since there has been an onset of web applications, developers have worked on different ways of getting duplex communication between the server and the browser. The aim of Java, Flash, Comet, and other workarounds, is the same. But for the first time, there is a specification to build a full-duplex communication system, by using HTML5 WebSockets. HTML5 specification contains an astonishing new communication system, a feature which can define a full-duplex communication channel operating over the Web through a single socket. Although the WebSockets RFC is published, the older browsers that are still in use are not, and will never be able to obtain it. Socket.io is an abstraction layer for WebSockets, which has Flash, XHR, JSONP, and HTML File fallbacks. An easy server and client library, which is used for making real-time, streaming updates between a web server and a browser client, is also provided by the Socket.IO. [12]

[1] Information used for communication between server and client

III. CONCLUSION

This article presents a big prototype platform for fraud detection, and gives an overview of the architectural integration in combination with multiple technologies, frameworks and techniques through three-layer architecture. It is shown how the M*EAN stack solution supports development of a graphical user interface with quantitative and other types of charts that give some general information about the fraudulent. Since fraud continues to be a very real problem in the telecom industry, presented big data architecture has great potential in identification of suspicious calls and users.

REFERENCES

[1] Schmerken, Ivy (May 15, 2008), Deciphering the Myths Around Complex Event Processing, Wall Street & Technology

[2] Bates, John, John Bates of Progress explains how complex event processing works and how it can simplify the use of algorithms for finding and capturing trading opportunities, Fix Global Trading, retrieved May 14, 2012

[3] I. S. Jacobs and Dafydd Stuttard and Marcus Pinto. The Web Application Hacker's Handbook: Discovering and Exploiting Security Flaws. John Wiley & Sons, Inc., New York, NY, USA, 2008. ISBN 9780470170779.

[4] Mean - full-stack javascript using mongodb, express, angularjs, and node.js. URL http://mean.io.

[5] Microsoft developer network: Component guidelines. URL https://msdn.microsoft.com/en-us/library/ee658121.aspx.

[6] Wojciech Bednarski. Learning JavaScriptMVC. Packt Publishing, 2013. ISBN 9781782160205.

[7] Lopes Tavares H. Dayvson Da Silva, M. Redis Essentials. Packt Publishing, 2015.

[8] A. Freeman. Pro AngularJS. Apress, 2014.

[9] Brad Green and Shyam Seshadri. AngularJS. O'Reilly Media, Inc, 2013. ISBN 978-1-4493-4485-6.

[10] R. Hailard. Recommended practice for architectural description of software-intensive systems, 2010. URL http://www.enterprise-architecture.info/Images/Documents/IEEE%201471-2000.pdf.

[11] Powell J. Mikowski, M. Single Page Web Applications. Manning Publications Co., 2014. ISBN9781617290756.

[12] Rohit Rai. Socket.IO Real-time Web Application Development. Packt Publishing, 2013. ISBN9781782160786.

NETWORK TECHNOLOGIES

A Survey on physical layer Impairments Aware Routing and Wavelength Assignment Algorithms in transparent wavelength routed optical networks

H. Dizdarević, S. Dizdarević, M. Škrbić, N. Hadžiahmetović
BH Telecom d.d. Sarajevo, Sarajevo, Bosna i Hercegovina
BH Telecom d.d. Sarajevo, Sarajevo, Bosna i Hercegovina
Elektrotehnički fakultet Univerziteta u Sarajevu, Sarajevo, Bosna i Hercegovina
Elektrotehnički fakultet Univerziteta u Sarajevu, Sarajevo, Bosna i Hercegovina

haris.dizdarević@bhtelecom.ba, sead.dizdarevic@bhtelecom.ba, mskrbic@etf.unsa.ba, nasuf.hadziahmetovic@etf.unsa.ba

Abstract - A wavelength-routed optical network consists of a number of cross-connects which are interconnected by optical fibers. Transparent optical networks allow signals to bypass extensive electronic signal processing at intermediate nodes. The quality of an optical signal degrades as it travels though several optical components along its lightpath. It is essential to incorporate imapact of the physical impairments in the routing and wavelength assignment (RWA) problem in transparent optical networks. This work compiles a comprehensive survey, from various research publications, of the proposed algorithms that address this issue. The physical layer impairments classification in optical networks are initially presented followed by physical layer impairments (PLI) RWA algorithms and algorithmic approaches for the PLI-RWA problem. Performance metrics for evaluating the performance of the proposed PLI-RWA algorithms are presented at the end of the work. Further research topics are presented in this study.

I. INTRODUCTION

Optical network architectures are evolving from traditional opaque networks via translucent networks toward transparent (i.e. all-optical) networks. In opaque optical networks, regenerators are available at each node and as the optical signal travels from its source to its destination, it undergoes signal regeneration at each intermediate node. OEO (optical-electronic-optica) conversion provide a larger range of optical signals, which leads to increased costs due to the large number of regenerators which are needed in the network. A network is said to be transparent if there are no regenerators used on any lightpath (i.e. end-to-end optical communication channel between two nodes,) established over the network. Translucent network architectures have been proposed as a compromise between opaque and transparent optical networks.

Once a set of lightpaths has been chosen or determined, we need to route each such lightpath in the network and assign a wavelength to it. This is referred to as the routing and wavelength assignment (RWA) problem. RWA problem is one of a very complex problem that needs to be addressed when designing wavelength routed optical networks. A large number of studies and research papers are devoted to the problem of RWA [1], which is known as NP-complete (Nondeterministic Polynomial time).

In transparent optical network, impairments due to unideal optical components at the physical layer can significantly affect network performance. It must be given special attention to impairments of the physical layer on network performance and accordingly develop advanced algorithms for RWA needs to overcome these influences. In most RWA proposals the optical level considered as a perfect system of transmission and therefore all results of RWA process considered as a feasible solution with respect to QoT (Quality of Transmission) requirements. However, the actual QoT of lightpaths may be unacceptable. Therefore, taking into account the impairments of the physical layer in the planning and operation of complete optical network has recently attracted considerable attention of the scientific community. PLI (Physical Layer Impairments) is considered as a limitation of RWA decisions (PLI constrained) or their influence integrated in RWA decisions (PLI aware). For simplicity is used a generic called IA-RWA, which refers to both approaches.

Routing and wavelength assignment (RWA) is the core component in planning and operation of optical networks. RWA is the fundamental control problem in optical WDM networks. We employ algorithms to solve the RWA problem. Since the performance of a network depends not only on its physical resources (e.g., optical cross-connects, converters, fiber links, number of wavelengths per fiber, etc.) but also on how it is controlled, the objective of an RWA algorithm is to achieve the best possible performance within the limits of

physical constraints. A large number of algorithms have been developed to solve the RWA problem. Taking into account the physical layer impairments in the RWA decisions adds a new dimension to this problem. It's very important to incorporate the effects of physical impairments in network algorithms. There are also some proposals that address the resilience and protection issues in the IA-RWA algorithms. IA-RWA algorithms can be incorporated in the control plane which is recognized as a necessary requirement for fast and flexible resource provisioning, easy network operation and high reliability and scalability of optical networks. IA-RWA algorithms based on network optimization theory are very important because network optimization methods are suitable for planning of network resources, as well as for operation and lightpath provisioning. IA-RWA proposal must take into account any constraints posed by the current network state. One big challenge for IA-RWA algorithms is that they must ensure that all lightpaths in the network meet a QoT (Quality of Transmission), e.g. BER, threshold. One of the objectives is to minimize the number of wavelengths that will be needed to carry a certain number of connections in the network for a given physical topology. Algorithms could be more efficient when more physical impairments are incorporated in the lightpath-assignment procedure. The physical size of the transparent network is mainly limited by the impairments. The impairment-aware algorithms could make connection provisioning more efficient and hence lead to more efficient resource utilization. It is therefore a clear need and importance of those RWA algorithms for efficient connection provisioning with signal-quality guarantees in an all-optical WDM network operating with high-speed wavelength channels. The impairment-aware algorithms efficiently provide signal-quality-guaranteed connections while significantly reducing connection-blocking probability, better utilizing network resources and having a reasonable computational requirement.

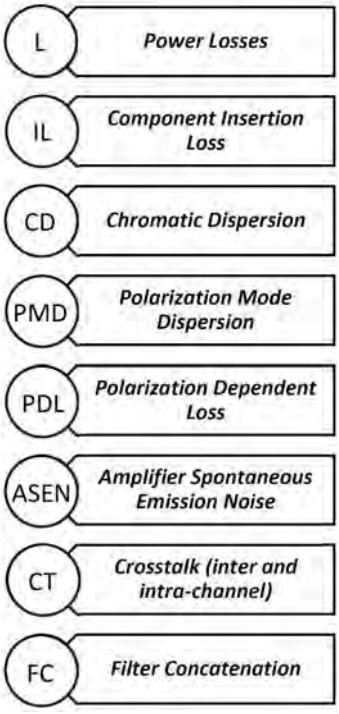


Figure 1. Linear physical layer impairments

It is very important for operators and network designers to know a different types of PLI, their effects on the feasibility of the light path, analytical modeling PLI as well as monitoring and mitigation techniques of their impact. This work compiles a comprehensive survey of the proposed algorithms that address this issues.

II. PHYSICAL LAYER IMPAIRMENTS

The effects of the physical layer impairments can not be ignored in an all-optical routing. Physical layer impairments are generally classified into two categories: linear and non-linear [8-11]. The terms linear and nonlinear in optics means independent and dependent of intensity of the signal power, respectively. The linear impairments are static, while the non-linear are dynamic. Non-linear attenuation strongly depends on the current distribution of routes and wavelengths, ie. of the current status of the allocated lightpaths.

A. Linear impairments

The most important linear impairments are: power losses (attenuation), component insertion loss (IL), Chromatic Dispersion (CD), Polarization Mode Dispersion (PMD), Polarization Dependent Loss (PDL), Amplifier Spontaneous Emission (ASE) Noise, crosstalk (inter- and intra-channel) (CT) and Filter Concatenation (FC).

Figure 1 depicts linear physical layer impairments.

Power loss can be defined as the optical loss that is accumulated from source to destination along fiber links. The loss introduced by the insertion of optical components is called insertion loss (IL) and is usually independent of wavelength. The most serious linear impairment for optical systems operating at bit-rates higher than 2.5 Gb/s is CD. Chromatic dispersion causes pulse broadening which affects the receiver performance. PMD becomes an issue at 40 Gbps or higher bit rates and for WDM systems designed for longer distances [15], [66], [82]. PDL is a measure of the peak-to-peak difference in transmission of an optical component or system with respect to all possible states of polarization. PDL affects the signal quality and system performance and the fluctuation of the signal-to-noise ratio. The primary source of additive noise in optically amplified systems is due to the ASE produced by the optical amplifiers. ASE emitted by the amplifier in the forward direction is a direct concern to system performance. Linear crosstalk arises due to incomplete isolation of WDM channels by optical components and depends on the ratio of the optical powers of two channels. Imperfect optical components introduce signal leakage either as inter-channel [25], [66] or intra-channel [25], [66] crosstalk in WDM transmission systems. The narrowing of spectral width of the optical signal as it traverses through a number of filters along a path is called filter concatenation (FC) effect.

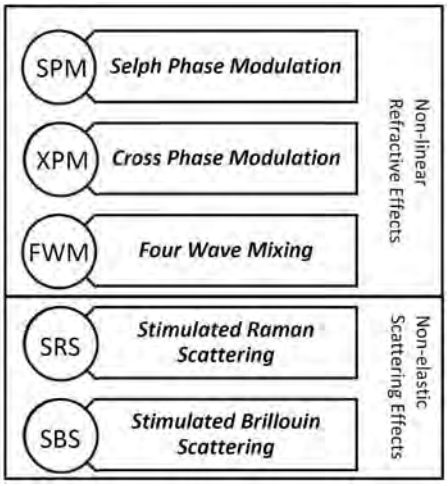


Figure 2. Non - linear physical layer impairments

B. Non-linear impairments

The optical nonlinearity represents a major obstacle for optical communication systems which must be overcome. The effects of non-linear impairments become critical with increasing the data transmission speed, transmission length, the number of wavelengths, the optical power levels and with decreasing spacing between channels.The nonlinear effects in optical fiber occur either due to intensity dependence of refractive index of the medium or due to phenomenon of non-elastic scattering. Nonlinear effects are much more complex than linear.

Classification of nonlinear effects in optical media is shown in Figure 2.

A simple physical model of the non-linear effects is presented in assessing the performance of optical networks with the aim of creating RWA algorithms with transmission impairments on each lightpath [12] and taking into account the static noise of optical components and nonlinear effects due to the current allocation and use of wavelengths. Nonlinearities rapidly degrade the quality of the transmission layer with encreasing the number of established lightpaths i.e. when the offered network load is large.

TABLE I. TECHNIQUES FOR PERFORMANCE EVALUATION

Techniques for performance evaluation	References
Analytical models	
Linear impairments	[15], [18], [20], [21], [26], [29], [30], [32], [33], [37], [40], [41], [42], [43], [45], [47]
Non-linear (and linear) impairments	[16], [21], [23], [24], [31]
Hibrid models	
Analytical and simulations	[2], [18], [19], [27], [35], [36]
Analytical and monitoring	[28], [46]

Analytical and hybrid models have been reported in order to encompass PLI effects in the RWA algorithms.

Table 1 summarizes the reported techniques for performance evaluation.

Most of today's RWA algorithms do not take into account the quality of the signal. For efficient management of optical networks we need new RWA algorithms that take into account the impairments of the physical layer (i.e. physically aware RWA).

III. PLI-RWA ALGORITHMS

The RWA problem in transparent optical networks with consideration of the impact of the impairments of the physical layer is currently in the focus of both industry and academic. This chapter provides an overview of previous research achievements in the field of routing and wavelength assignment in transparent optical networks through review of IA-RWA algorithms and their algorithmic approaches as well as description of performance metrics and methods adopted for the evaluation of the different proposals. RWA problem can be decomposed into two subproblems: a routing problem and a wavelength assignment problem. Simplification of IA-RWA problem can be achieved using a heuristic (or meta-heuristic) algorithm.

A. Heuristics

In the literature we can find a variety of heuristic algorithms in connection with a routing subproblem, a wavelength assignment subproblem, as well as heuristics intended for solving PLI-RWA problem jointly. In Table 2 are presented reported heuristic algorithms.

Regarding the routing sub-problem a number of IA-RWA algorithms used SP (shortest path) approach with a minimum number of hops [15], [16], [18], [19] , [20], [21] and [24]. In the literature are described the PI aware RWA algorithms that use different link costs [18], [20], [21] and [23].

TABLE II. PLI-RWA HEURISTICS ALGORITHMS

	References
Single path routing	
Hop-based, PI-aware and modified shortest path	[15], [16], [18], [19], [20], [21], [23], [24], [26], [27], [46], [83]
Multi path routing	
Hop-based and PI-aware k-shortest path route calculation	[28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [40]
Sequential and parallel route selection	[29], [30], [32], [38]
Wavelength Assignment	
First-Fit	[2], [16], [18], [19], [21], [27], [28], [30], [37], [40], [41], [42], [43], [45],
Best-Fit	[27]
Random	[15], [21]
PI-aware	[23], [84]
Least-Loaded	[26]
Routing and Wavelength Assignment	
Best-OSNR	[19]
Minimum Crosstalk	[47]

Some IA-RWA algorithms modify the shortest path routing algorithm [19], [23], [26], [27], [46], [83]. In the multi-path routing the cost metric can be related to the hop

count [28], [29], [30], [31], [32] or can be PI-aware [33], [34], [35], [36], [37], [40]. Sequential and parallel route selection algorithms are published in [29], [30], [32], [38].

Wavelength Assignment sub-problem can be performed either sequentially or in parallel. Sequential approach, i.e. First-Fit selection method, has been considered in next proposals [2], [16], [18], [19], [21], [27], [28], [30], [37], [40], [41], [42], [43], [45]. PLI-RWA algorithm using Best-Fit approach is reported in [27]. Random selection amongst the available wavelengths can be performed [15], [21]. Some of the proposed wavelength assignment algorithms make a decision based only on physical layer impairments. A dynamic RWA algorithm encompassing physical impairment due to Four-Wave Mixing (FWM) is proposed in [23]. A novel cost function is introduced based upon an impairment-constraint-based routing (ICBR) approach, taking into account the physical impairment due to FWM crosstalk. A wavelength assignment algorithm suitable for optical networks mainly impaired by physical layer effects, named the Intelligent Wavelength Assignment algorithm (iWA) is presented in [84]. Some different physical layer effects, such as four-wave mixing and residual dispersion, were considered. The new RWA algorithm for semitransparent networks, the PR-MTD (Prediction Routing according the Maximum Transmission Distance) is described in [26]. If two or more wavelengths of the shortest route accomplish the three defined conditions, the algorithm will select the least loaded wavelength.

There are some heuristic algorithms that intent to solve the PLI-RWA problem jointly. One of them is Best-OSNR [19] which will jointly assign to a given request a path and a corresponding wavelength. Alorithm explicitly tries to minimize the impact of physical impairments. Another algorithm, called Minimum Crosstalk, that solves this issue is presented in [47]. MC runs a shortest path algorithm for each wavelength (with constant link weights equal to the physical link lengths) to find candidate routes. For each candidate route, the number of crosstalk components along the route is calculated.

B. Metaheuristics and optimizations methods

In addition to heuristic-based algorithms, there is a class IA-RWA algorithms using metaheuristic methods. Metaheuristics enable convergence to an optimal solution. One of the very often used methods to solve a variety of complex combinatorial problems is based on metaheuristic called ACO (Ant Colony Optimization) [40, 45, 48, 83]. Traditionally applied metaheuristic methods for solving IA-RWA problems are genetic algorithms (GA) and tabu search (TS). Metaheuristic IA-RWA algorithm that uses a genetic algorithm was published in [49]. Two transmission impairments are added to the usual set of routing constraints: polarization mode dispersion (PMD) and amplified spontaneous emission (ASE). For solving the RWA is proposed tabu search algorithm, which takes into account the PMD and ASE noise [37]. In [26] authors for solving the PLI-RWA problem authors proposed Predictive Algorithm (PA) with consideration to the physical layer impairments using the concept of MTD (Maximum Transmission Distance).

TABLE III. PLI-RWA METAHEURISTICS ALGORITHMS

Metaheuristics	**References**
ACO	[40], [45], [48], [83]
GA	[49]
TS	[37]
PA	[26]
Optimization methods	
ILP	[33], [42], [52], [53], [56]

The algorithm selects the lightpaths based on previous requests for connections and can be used in a distributed routing. In Table 3 are presented reported metaheuristic algorithms and also Integer Linear Programming (ILP) optimization methods used by many authors for solving PLI-RWA problem. The network optimization methods are appropriate for off-line optimization of network resources. Authors in [52] proposed a binary ILP and sequential heuristic approach to minimize blocking the lightpaths and increase efficiency while meeting the large number of lightpaths requests. Authors in [53] proposed ILP and Polynomial Time Heuristic approach to provide adequate use of allocated bandwidth while preventing any bit errors during transmission of lightpath request. Mixed ILP and Tabu, Greedy Search Heuristic approach is proposed like a method which allows information passing from one layers to another in a network [56]. Authors from [42] are also used mixed ILP formulation for the RWA problem of multicast connections. In [33] is proposed a link-path formulation to solve an ILP problem of RWA.

IV. PERFORMANCE METRICS

In the literature the traditional way of evaluating the performance of the proposed PLI-RWA algorithms has been the percentage of blocked connections versus the offered traffic load for statyc and dynamic traffic. For static scenario like a metric is used amount of required resources (optical fibers, wavelengths, etc). For a static traffic scenario the objective is to route the maximum number of connections if the resources of the network are fixed. In this scenario the objective may be to minimize the number of necessary wavelengths. In a dynamic scenario the objective is to route the maximum number of connections. Another metric is the performance evaluation of the proposed PLI-RWA algorithms. Authors evaluated proposed algorithms using simulation studies, experimental and analytical approaches or hybrid method.

Evaluations of the proposed PLI-RWA algorithms are presented in Table 4.

TABLE IV. EVALUATIONS OF THE PROPOSED ALGORITHMS

	References
Simulations	[15], [16], [18], [19], [20], [21], [23], [24], [26], [27], [28], [29], [31], [33], [34], [35], [36], [37], [40], [41], [43], [45], [47],
Simulations and experimentsm analytical models or hybrid approach	[32], [38], [46],

V. CONCLUSION

In transparent wavelength routed optical networks physical layer impairments determine the feasibility or transmission quality of the lightpaths and it's possible that reserved resources and lightpath become useless. In this article, we discussed the importance of PLIs and surveyed the related work. Traditional RWA approaches assume an ideal physical layer medium and ignore the effects of physical impairments on the lightpath feasibility. One of the major problems facing optical networks is RWA of optical signals taking into consideration the impairments caused by the physical layer.

Various heuristics, meta-heuristics and optimization techniques are proposed as algorithmic approaches to solve PLI-RWA problems. We presented a survey of a number of PLI-RWA algorithms available in the literature to aid in better understanding of the research in optical WDM networks. As no existing studies considers all PLIs in the RWA process, future research should consider both linear and non-linear impairments at the same time. PLI-RWA algorithms play important roles in maximizing the performance of an transparent optical network design. New algorithms should try to reduce the computational complexity due to PLI evaluation. The challenges identified in this survey need further consideration in current and future research projects.

REFERENCES

[1] H. Zang, J.P. Jue, B. Mukherjee, A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks, SPIE Opt Netw Mag 1(1), pp. 47–59, 2000.

[2] S. Pachnicke, T. Paschenda, PM. Krummrich, Physical impairment based regeneratorplacement and routing in translucent optical networks. In: Proceedings of IEEE/OSA optical fiber communication conference (OFC), San Diego, CA, 2008.

[3] J. Berthold, A.A.M. Saleh, L. Blair, J.M. Simmons, Optical networking: past, present, and future, J. Lightw. Technol. 26(9), pp. 1104–1118, 2008;

[4] M.R.I. Mat Jusoh et al., A review and survey of routing techniques in transparent core optical networks: Evolution, perspectives and frontiers, 2nd International Conference on Electronic Design (ICED), pp. 368-373, 2014.

[5] R. Ramaswami, K.N. Sivarajan, Routing and wavelength assignment in all-optical networks. IEEE/ACM Trans Netw 5(3): 489–500, 1995.

[6] G.N. Rouskas, Routing and wavelength assignment in optical WDM networks. In: Proakis J (ed) Wiley encyclopedia of telecommunications. Wiley, New York, 2001.

[7] S. Azodolmolky, M. Klinkowski, A survey on physical layer impairments aware routing and wavelength assignment algorithms in optical networks, Elsevier Comput. Netw. 53(7), 926–944, 2009;

[8] G. P. Agrawal, Fiber Optic Communication Systems. Wiley Interscience, 1992.

[9] G. P. Agrawal, Nonlinear Fiber Optics. New York: Academic, 1989.

[10] B. Mukherjee, Optical WDM Networks. Springer, 2006.

[11] R. Ramaswami and K. Sivarajan, Optical Networks: A Practical Perspective. Morgan Kauffman, 2001.

[12] R. Cardillo, V. Curri, and M. Mellia, Considering transmissionimpairments in wavelength routed networks, Proceedings, 9th Conference on Optical Network Design and Modeling (ONDM), '05, Milan, Italy, pp. 421-429, Feb. 2005.

[13] H. Zang, J.P. Jue, B. Mukherjee, Capacity Allocation and Contention Resolution in a Photonic Slot Routing All-Optical WDM Mesh Network. IEEE/OSA Journal of Lightwave Technology, vol. 18, no. 12, pp. 1728-1741, 2000.

[14] S. Pachnicke, T. Paschenda, and P. M. Krummrich, Physical impairment based regenerator placement and routing in translucent optical networks. In: Proceedings of IEEE/OSA optical fiber communication conference (OFC), San Diego, CA, 2008

[15] B. Ramamurthy, D. Datta, H. Feng, J.P. Heritage, B. Mukherjee, Impact of transmission impairments on the teletraffic performance of wavelength-routed optical networks. IEEE/OSA J Lightwave Tech 17(10):1713–1723, 1999.

[16] H.A. Pereira, D.A.R. Chaves, C.J.A. Bastos-Filho, J.F. Martins-Filho, Impact of physical layer impairments in all-optical networks. In: Proceedings of IEEE international microwave and optoelectronics conference (IMOC), Salvador, Brazil, pp. 536–541, 2007.

[17] Y. Huang, J.P. Heritage, B. Mukherjee, Connection provisioning with transmission impairment consideration in optical WDM networks with high-speed channels. IEEE/OSA J Lightwave Tech 23(3):982–993, 2005.

[18] T.J. Carpenter, R.C. Menendez, D.F. Shallcross, J.W. Gannett, J. Jackel, A.C. Von Lehmen, Cost-conscious impairment-aware routing. In: Proceedings of IEEE/OSA optical fiber communication conference (OFC) 1, San Diego, CA, 2004.

[19] R. Cardillo, V. Curri, M. Mellia, Considering transmission impairments in configuring wavelength routed optical networks. In: Proceedings of IEEE/OSA optical fiber communication conference (OFC), Anaheim, CA, 2005.

[20] S. Zsigmond, G. Németh G, T. Cinkler, Mutual impact of physical impairments and grooming in multilayer networks. In: Proceedings of international conference on optical networking design and modeling (ONDM), Athens, Greece, pp. 38–47, 2007.

[21] J. He, M. Brandt-Pearce, Y. Pointurier, S. Subramaniam, QoT-Aware routing in impairment-constrained optical networks. In: Proceedings of IEEE global communications conference (GLOBECOM), Washington, DC, pp. 2269–2274, 2007.

[21] J. He, M. Brandt-Pearce, S. Subramaniam, QoS-aware wavelength assignment with ber and latency guarantees for crosstalk limited networks. In: Proceedings of IEEE international communication conference (ICC), Glasgow, Scotland, pp. 2336 – 2341, 2007.

[23] A. Marsden, A. Maruta, K. Kitayama, Routing and wavelength assignment encompassing FWM in WDM lightpath networks. In: Proceedings of international conference on optical networking design and modeling (ONDM), Vilanova i la Geltrú, Spain, pp. 1-6, 2008.

[24] J. He, M. Brandt-Pearce, Y. Pointurier, C.L. Brown, S. Subramaniam, Adaptive wavelength assignment using wavelength spectrum separation for distributed optical networks. In: Proceedings of IEEE international communication conference (ICC), Glasgow, Scotland, pp. 2406 - 2411, 2007.

[25] M. Farahmand, D. Awduche, S. Tibuleac, D. Atlas, Characterization and representation of impairments for routing and path control in all-optical networks. In: Proceedings IEEE/OSA national fiber optics engineers conference (NFOEC), Dallas, TX, 2002.

[26] E. Marin, S. Sánchez, X. Masip, J. Solé, G. Maier, W. Erangoli, S. Santoni, M. Quagliotti, Applying prediction concepts to routing on semi transparent optical transport networks. In: Proceedings of IEEE international conference on transparent optical networks (ICTON), Rome, Italy, pp. 32–36, 2007.

[27] R. Cardillo, V. Curri, M. Mellia, Considering transmission impairments in wavelength routed networks. In: Proceedings of international conference on optical networking design and modeling (ONDM), Milan, Italy, pp 421–429, 2005.

[28] M.A. Ezzahdi, S. Al Zahr, M. Koubaa, N. Puech, M. Gagnaire, LERP: A quality of transmission dependent heuristic for routing and wavelength assignment in hybrid WDM networks. In: Proceedings on international conference computer communication and networks (ICCCN), Arlington, VA, pp. 125–136, 2006.

[29] E. Salvadori, Ye Yabin, A. Zanardi, H. Woesner, M. Carcagni, G. Galimberti, G. Martinelli, A. Tanzi, D. La Fauci, Signaling-based architectures for impairment-aware lightpath set-up in GMPLS networks. In: Proceedings of IEEE global communications conference (GLOBECOM), Washington, DC, pp. 2263–2268, 2007.

[30] X. Yang, B. Ramamurthy, Dynamic routing in translucent WDM optical networks: the intradomain case. IEEE/OSA J Lightwave Tech 23(3), pp. 955–971, 2005.

[31] G. Markidis, S. Sygletos, A. Tzanakaki, I. Tomkos, Impairment aware based routing and wavelength assignment in transparent long haul networks. In: Proceedings of international conference on optical networking design and modeling (ONDM), Athens, Greece, pp 48–57, 2007.

[32] F. Cugini, N. Andriolli, L. Valcarenghi, P. Castoldi, A novel signaling approach to encompass physical impairments in GMPLS networks. In: Proceedings of IEEE global communications conference (GLOBECOM), Dallas, TX, pp. 369–373, 2004.

[33] I. Tomkos, D. Vogiatzis, C. Mas, I. Zacharopoulos, A. Tzanakaki, E. Varvarigos, Performance engineering of metropolitan area optical networks through impairment constraint routing. IEEE Commun Mag 42(8), S40–S47, 2004.

[34] T. Deng T, S.S. Subramaniam, Adaptive QoS routing in dynamic wavelength-routed optical networks. In: Proceedings of international conference on broadband communication networks and systems (Broadnets), vol 1, Boston, MA, pp. 184–193, 2005.

[35] P. Kulkarni, A. Tzanakaki, Mas C, I. Tomkos, C.M. Machuka, Benefits of Q-factor based routing in WDM Metro Networks. In: Proceedings of European conference on optical communications (ECOC), vol 4. Glasgow, Scotland, pp. 981–982, 2005.

[36] G. Markidis, S. Sygletos, A. Tzanakaki, I. Tomkos, Impairment-constraint-based routing in ultralong-haul optical networks with 2R regeneration. IEEE Photonics Tech Lett 19(6), pp. 420–422, 2007.

[37] X. Yang, L. Shen, B. Ramamurthy, Survivable lightpath provisioning in WDM mesh networks under shared path protection and signal quality constraints. IEEE/OSA J Lightwave Tech 23(4), pp. 1556–1567, 2005.

[38] F. Cugini, F. Paolucci, L. Valcarenghi, P. Castoldi, Implementing a path computation element (PCE) to encompass physical impairments in transparent networks. In: Proceedings of IEEE/OSA optical fiber communication conference (OFC/NFOEC), Anaheim, CA, pp. 1–3, 2007.

[39] J. He, M. Brandt-Pearce, RWA using wavelength ordering for crosstalk limited networks. In: Proceedings of IEEE/OSA optical fiber communication conference (OFC/NFOEC), Anaheim, CA, 2006.

[40] G.S. Pavani, L.G. Zuliani, H. Waldman, M. Magalhães, Distributed approaches for impairment-aware routing and wavelength assignment algorithms in GMPLS networks. Comput Netw 52(10), pp. 1905–1915, 2008.

[41] J.F. Martins-Filho, C.J.A. Bastos-Filho, E.A.J. Arantes, S.C. Oliveira, L.D. Coelho, J.P.G. de Oliveira, R.G. Dante, E. Fontana, F.D. Nunes, Novel routing algorithm for transparent optical networks based on noise figure and amplifier saturation. In: Proceedings of IEEE international microwave and optoelectronics conference (IMOC), vol 2. Iguaza Falls, Brazil, pp. 919–923, 2003.

[42] A.M. Hamad, A.E. Kamal, Routing and wavelength assignment with power aware multicasting in WDM networks. In: Proceedings of international conference on broadband communication networks and systems (Broadnets), vol 1. Boston, MA, pp. 31–40, 2005.

[43] E. Salvadori, Ye Yabin, A. Zanardi, H. Woesner, M. Carcagnì, G. Galimberti, G. Martinelli, A. Tanzi, D. La Fauci, A study of connection management approaches for an impairmentaware optical control plane. In: Proceedings of international conference on optical networking design and modeling (ONDM), Athens, Greece, pp. 229–238, 2007.

[45] G.S. Pavani, H. Waldman, Adaptive routing and wavelength assignment with power constraints using ant colony optimization. In: Proceedings of international telecommunication symposium, Fortaleza, Brazil, pp. 637–642, 2006.

[46] A. Morea, N. Brogard, F. Leplingard, J.C. Antona, T. Zami, B. Lavigne, D. Bayart, QoT function and A* Routing an optimized combination for connection search in translucent networks. OSA J Opt Netw 7(1), pp. 42–61, 2008.

[47] Y. Zhai, Y. Pointurier, S.S. Subramaniam, M. Brandt-Pearce, QoS-aware RWA algorithms for path-protected DWDM Networks. In: Proceedings of IEEE/OSA optical fiber communication conference (OFC/NFOEC), Anaheim, CA, pp. 1-3, 2007.

[48] K. Lee, M. Shayman, Optical network design with optical constraints in IP/WDM networks. IEICE Trans Commun E88(B5), pp. 1898–1905, 2005.

[49] M.A.C. Lima, AC. Cesar, A.F.R. Araujo, Optical network optimization with transmission impairments based on genetic algorithm. In: Proceedings of IEEE international microwave and optoelectronics conference (IMOC), vol 1. Foz do Iguaqu, Brazil, pp. 361–365, 2003.

[50] W. Zhang, J. Tang, K. Nygard, C. Wang, REPARE: Regenerator Placement and Routing Establishment in Translucent Networks. GLOBECOM 2009 - 2009 IEEE Global Telecommunications Conference, pp. 1-7, 2009.

[51] K. Manousakis, K. Christodoulopoulos, E. Kamitsas, I. Tomkos, E. Varvarigos, Offline Impairment-Aware Routing and Wavelength Assignment Algorithms in Translucent WDM Optical Networks. Journal of Lightwave Technology 27, 12, pp. 1866-1877, 2009.

[52] P. Pavon-Marino, S. Azodolmolky, R. Aparicio-Pardo, B. Garcia-Manrubia, Y. Pointurier, M. Angelou, J. Sole-Pareta, J. Garcia-Haro, I. Tomkos, Offline Impairment Aware RWA Algorithms for Cross-Layer Planning of Optical Networks. Journal of Lightwave Technology 27,12, pp. 1763-1775, 2009

[53] N. Sengezer, and E. Karasan, Static Lightpath Establishment in Multilayer Traffic Engineering Under Physical Layer Impairments. IEEE/OSA Journal of Optical Communications and Networking 2, 9, pp. 662-677, 2010.

[54] R. Aparicio-Pardo, B. Garcia-Manrubia, M. Klinkowski, P. Pavon-Marino, D. Careglio, Offline Impairment-Aware RWA and Regenerator Placement in Translucent Optical Networks. Journal of Lightwave Technology 29,3, pp. 265-277, 2011.

[55] K.M. Katrinis, and A. Tzanakaki, On the Dimensioning of WDM Optical Networks with Impairment-Aware Regeneration, IEEE/ACM Transactions 19, 3, pp. 735-746, 2011.

[56] N. Sengezer, E. Karasan, Multi-layer Virtual Topology Design in Optical Networks Under Physical Layer Impairments and Multi-hour Trafficc Demand, IEEE/OSA Journal of Optical Communications and Networking 4, 2, pp. 78-91, 2012.

[57] K. Christodoulopoulos, K. Manousakis, E. Varvarigos, M. Angelou, Considering Physical Layer Impairments in Offline RWA, IEEE Network, 23, 3, pp. 26-33, 2009.

[58] E. Varvarigos, K. Manousakis, K. Christodoulopoulos, Offline Routing and Wavelength Assignment in Transparent WDM Networks, IEEE/ACM Trans. on Networks 18, 5, pp. 1557-1570, 2010.

[59] J. Sole, S.S. Subramaniam, D. Careglio, S. Spadaro, Cross-Layer Approaches for Planning and Operating Impairment-Aware Optical Networks, Proceedings of the IEEE 100, 5, pp. 1118-1129, 2007.

[60]J. Y. Youe, S. W. Seo, An Algorithm for Virtual Topology Design in WDM Optical Networks under Physical Constraints, Proc. IEEE ICC 1999, vol. 3, pp. 1719-1723, 1999.

[61] M. Ali, L. Tancevski, Impact of Polarization Mode Dispersion on the Design of Wavelength-Routed Networks, IEEE Photonics Technol.Lett., vol. 14, No. 5, pp. 720-722, 2002.

[62] R. Sabella et al., Impact of Transmission Performance on Path Routingin All-Optical Transport Networks,IEEE/OSA J. Lightwave Technol.,vol. 16, no. 11, pp. 1965-1972, 1998.

[63] G. Bogliolo et al., "Considering Transmission Impairments in RWA Problem: Greedy and Metaheuristic solutions," Proc. OFC 2007.

[64] M. Ali et al., Routing and wavelength assignment with power considerations in optical networks, Computer Communications, no. 32, pp. 539-555, 2000.

[65] R. Martnez et al., On-line ICBR in a Transparent GMPLS Network: A Reality Check, WGN5: V Workshop in G/MPLS networks, 2006.

[66] J. Strand et al., Impairments and Other Constraints on Optical Layer Routing," RFC 4054, 2005.

[67] X. Yang and B. Ramamurthy, Inter-Domain Wavelength Routing in the Next-Generation Translucent Optical Internet

Backbones,OSA J. Optical Networking, vol. 3, no. 3, pp. 169-187, 2004.

[68] A. Farrel et al.,A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering, RFC 4726, 2006.

[69] A. Farrel et al., Inter domain Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Traffic Engineering - RSVP-TE Extensions, RFC 5151, 2008.

[70] J. P. Vasseur et al., A Per-domain path computation method for establishing Inter-domain Traffic Engineering (TE) Label Switched Paths (LSPs), RFC 5152, 2008.

[71] P. Pavarangkoon, E. Oki, Survivable lightpath provisioning in multi-domain optical networks, HPSR, 2014 IEEE 15th International Conference on, pp. 20-205, 2014.

[72] R. Lu et al., Domain-level-based routing algorithm in multidomain WDM optical network,IEEE/OSA Journal Journal of Optical Communications and Networking vol. 5, no. 5, pp 532-543, 2013.

[73] J. Zao et al., Inter-domain QoT-aware RWA for translucent optical networks, International Conference on Computing, Networking and Communications, San Diego, CA, pp. 1102-1106, 2013.

[74] JL. Le Roux,Path Computation Element Communication Protocol (PCECP) Specific Requirements for Inter-Area MPLS and GMPLS Traffic Engineering,RFC 4927, 2007.

[75] I. E. Fonseca et al., Algorithms for FWM-aware Routing and Wavelength Assignment,Proc. IEEE IMOC 2003, vol. 2, no. 2, pp. 707-712, 2003.

[76] T. Deng et al., Crosstalk-Aware Wavelength Assignment in Dynamic Wavelength-Routed Optical Networks,Proc. Broadnets 2004, pp. 140-149, 2004.

[77] J. He and M. Brandt-Pearce, Dynamic Wavelength Assignment using Wavelength Spectrum Separation for Crosstalk Limited Networks, Proc. IEEE Broadnets 2006, pp. 1-9, 2006.

[78] E. Martinelli et al., A heuristic approach for border node selection and traffic routing in multi-domain optical networks, Optical Switching and Networking, vol. 15, pp. 87-96, 2015.

[79] S.C. Tan, F.M. Abbou, The effect of Stimulated Raman Scattering (SRS) in Routing and Wavelength Assignment (RWA) in WDM mesh and ring topologies, Optik - International Journal for Light and Electron Optics, vol. 123, Issue 21, pp. 1934-1937, 2012.

[80] R.J. Pandya, V. Chandra, D. Chadha, Impairment aware routing and wavelength assignment algorithms for optical WDM networks and experimental validation of impairment aware automatic light-path switching, Optical Switching and Networking, vol. 11, Part A, pp. 16-28, 2014.

[81] A. Ebrahimzadeh, A.G. Rahbar, B. Alizadeh, QoS aware green routing and wavelength assignment in core WDM networks, Journal of Network and Computer Applications, vol. 53, pp. 1-13, 2015.

[82] J. Strand, A.L. Chiu, R. Tkach, ssues for routing in the optical layer, IEEE Communications Magazine, vol. 39, no. 2, pp. 81-87, 2002.

[83] A. Keleş, A. Yayimli, A.S. Uyar, Ant based Hyper Heuristic for Physical Impairment aware routing and wavelength assignment, IEEE Sarnoff Symposium, pp. 1-5, 2010.

A Survey on transition from GMPLS control plane for optical multilayer networks to SDN control plane

mr.sci. Sead Dizdarević, dipl.ing.el., mr.sci. Haris Dizdarević, dipl.ing.el., dr.sci. Mirko Škrbić, dipl.ing.el. and dr.sci. Nasuf Hadžiahmetović, dipl.ing.el
BH Telecom d.d. Sarajevo/ Executive Directorate for Investments, Sarajevo, BiH
BH Telecom d.d. Sarajevo/ Executive Directorate for Investments, Sarajevo, BiH
Faculty of Electrical Engineering/University of Sarajevo, Sarajevo, BiH
Faculty of Electrical Engineering/University of Sarajevo, Sarajevo, BiH

Abstract - Software-defined network (SDN)/OpenFlow research and development have attracted the attention of many researchers and companies. The key idea of an SDN is characterized by decoupling the control plane from the data plane. This paper surveys the state-of-the-art of SDN and latest developments in this active research area of SDN. This work also compiles a comprehensive survey of programmable networks from early ideas to recent developments. SDN architecture and the OpenFlow standard in particular are initially presented followed by current alternatives for implementation and SDN protocols and services. It is also discussed why OpenFlow/SDN can succeed where Generalized Multi-Protocol Label Switching (GMPLS) failed in terms of optimal strategy transition from GMPLS control plane for optical multilayer networks to SDN control plane. At the end of the work are presented SDN applications and further research topics with some suggested open research challenges.

I. INTRODUCTION

Today Internet applications require the underlying networks to be fast, carry large amounts of traffic, and to deploy a number of distinct applications and services.

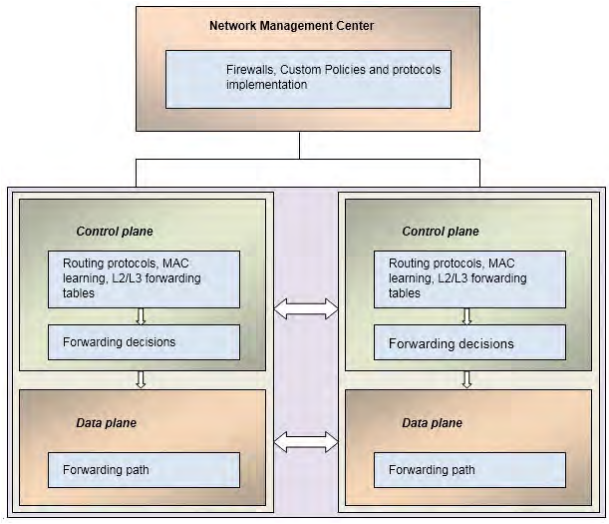


Figure 1. Inflexible legacy infrastructure.

In addition to various proprietary network hardware, distributed protocols, and software components, legacy networks are inundated with switching devices that decide on the route taken by each packet individually. The data paths and the decision making processes for switching or routing are collocated on the same device. This situation is shown in Fig. 1.

The decision making capability or network intelligence is distributed across the various network hardware components. This makes the introduction of any new network device or service a monotonous job because it requires reconfiguration of each of the numerous network nodes.

Legacy networks have become difficult to automate [1]. Networks today depend on IP addresses to identify and locate servers and applications. This approach works fine for static networks where each physical device is recognizable by an IP address, but is extremely uneasy for large virtual networks. Managing such complex environments using traditional networks is time-consuming and expensive. The rigid structure of legacy networks prohibits programmability to meet the variety of client requirements, sometimes forcing vendors into deploying complex programmable management systems.

Vast teams of network administrators are employed to make thousands of changes manually to network components [2]. The demand for services and network usage is growing rapidly. Operators require a network that is efficient, flexible, agile and scalable. Software-defined networking (SDN) and virtualization are poised to be the solutions that overcome the challenges described above.

The need for dynamically set up virtual tunnels in NSP (Network Service Provider) backbone networks, where QoS requests arrive one by one and traffic matrix is not known in advance, is caused to SDN is proposed as replacement for today's MPLS (Multi Protocol Label Switching) networks, in which none of the bigger NSP providers uses traffic engineering mechanisms due to instability and high management complexity. GMPLS is predominantly implemented as an overlay model, with separate control of IP and Transport and in practice is typically not available across different

vendors and there is no higher level awareness of the network.

In the current metro and backbone optical networks, optical nodes are controlled and managed through the EMS (Element Management System) and the NMS (Network Management System) in a manual and semistatic style for lightpath provisioning. The network carriers require a control plane technique for dynamic and intelligent control of wavelength paths in these optical networks. A well-known choice is the GMPLS for automatically provision end-to-end connections in a fully distributed manner. The GMPLS-based control plane has not been widely deployed. One of the most important reasons is that the GMPLS-based distributed control plane is too complex, especially for dynamic control of both IP and optical layers as a UCP (Unified Control Plane).

On the other hand, the SDN architecture, in particular, the OpenFlow protocol can provide satisfactory flexibility for operators to control a network (i.e., software defined) and can arguably match carriers' preference given its simplicity and manageability.

The control plane plays a key role in enabling an intelligent and dynamic WSON (Wavelength Switching Optical Networks), with a great reduction of operational expense and processing latency. In recent years, there has been much advance in this area, ranging from the traditional GMPLS to a PCE (Path Computation Element)/GMPLS-based architecture. The PCE/GMPLS decouples the path computation functionality from GMPLS controllers (GCs) to a dedicated PCE, which can bring several advantages for optical network control. The PCE/GMPLS-based control plane is also complex because of the introduction of the PCE and PCE Communication Protocol (PCEP).

On the other hand, the OpenFlow based control plane provides maximum flexibility and manageability for carriers because all the functionalities are integrated into a single OpenFlow controller. More importantly, the OpenFlow-based control plane is a natural choice for a UCP in IP/DWDM multilayer networks. The technical maturity of both GMPLS and PCE is high. On the other hand, the OpenFlow-based control plane is still at a starting stage for optical networks.

II. SOFTWARE-DEFINED NETWORKING

SDN operates on an aggregated and centralized control plane that might be a promising solution for network management and control problems. The main idea behind SDN is to separate the forwarding/data plane from the control plane while providing programmability on the control plane as illustrated in Fig. 2.

A precise definition of SDN is hard to find, or rather, it's difficult choose the best from the large number of definitions. Therefore, it is necessary to briefly explain the network devices and the way they work in the classical network environment. Each network component consists of three basic planes. These are the management plane, control plane and data plane.

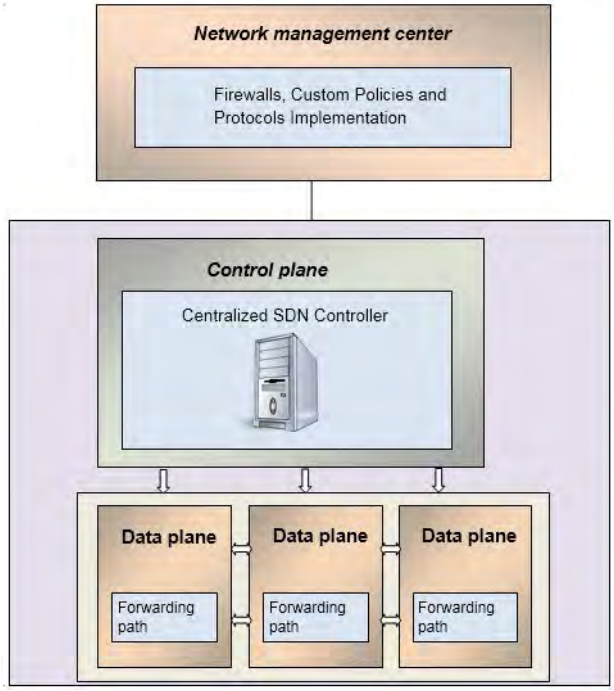


Figure 2. SDN architecture.

Routing protocol in the network calculated that the best route between the user and the Web server is: R1-R3-R5-R7, in terms of speed.

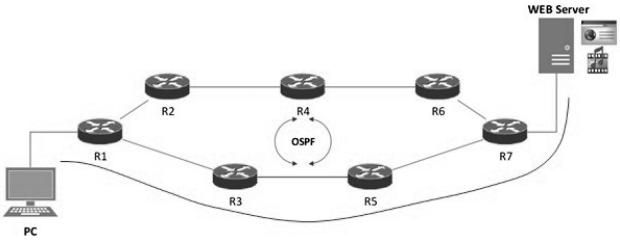


Figure 3. An example of the classic network environment [3].

We will assume that video content is located on the web server substantially burdens the path R1-R3-R5-R7, and thus makes it difficult to browsing websites, and at the same time is completely unencumbered path R1-R2-R4-R6-R7, which has a smaller capacity, but sufficient to browse web content on a server.

In classic network environment there is no simple and automatic way that would allow video content to travel one route, and web content through other route. The reason for this is that each router has its own control plane and it calculates for itself the best path in the network, considering the default parameters (speed connections - in this case).

The router cannot easily distinguish web content from the video content, which would allow him, according to the traffic type, make a decision on which interface it will send traffic. And most important of all, there is no possibility that router doing all this operations dynamically.

These problems can be solved so that the control plane from all the routers set aside on one component - the

controller, component that constantly communicates with all devices on the network and dynamically adjusted to any device on what way and in what time will send the traffic. So in our example sending video traffic can be enabled one route and web traffic to another route [3] (Figure 4).

Routes can be changed at any time, according to the needs of the user, network traffic, etc. So now it is possible to precisely define the term SDN.

SDN is the network technologies access that enables the separation of the control plane from data plane of the network device and switching control plane to a separate component called SDN controller.

An additional advantage of SDN is the possibility of using the network components from different vendors, without having to know how to work a certain device for the complete network environment managed from a single point, that is, SDN controller.

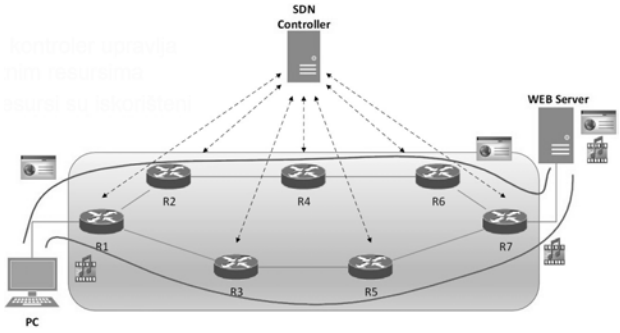


Figure 4. Network management by SDN controller.

So it is possible to further achieve significant operational savings in terms of costs of education and employment of workers, narrow specialization related to only one manufacturer, and capital, in terms of easier implementation of acceptable devices from different manufacturers in a single network environment. Every serious manufacturer who specialized in networking and virtualization technology developed its own SDN platform. In a virtualized environment, it should be mentioned VMware NSX, Citrix and Microsoft SDN solutions.

VMware NSX is the network virtualization and security platform for software defined data center. As regards developing SDN for hardware platforms it should be noted Cisco's ACI (Application Centric Infrastructure) solution and some HP solutions. There are large number of solutions based on OpenFlow developing environment (SDN development is initially started on it), and based on open source.

The Open Networking Foundation (ONF) [4] is a nonprofit consortium dedicated to development, standardization, and commercialization of SDN. The ONF defines the term of SDN as follows:

Software-Defined Networking (SDN) is an emerging network architecture where network control is decoupled from forwarding and is directly programmable [5].

Here, it should be mentioned SwitchWare [6], [7]. SwitchWare is an active networking solution, allowing packets flowing through a network to modify operations of the network dynamically. Similarly, software routing suites on conventional PC hardware, such as Click [8], XORP [9], Quagga [10], and BIRD [11], also attempt to create extensible software routers by making network devices programmable. Attempts of decoupling between control and data planes has been proliferated during the last decade.

Caesar et al. first presented a Routing Control Platform (RCP) in 2004 [12], in which Border Gateway Protocol (BGP) inter-domain routing is replaced by centralized routing control to reduce complexity of fully distributed path computation.

In the same year, IETF (Internet Engineering Task Force) released the Forwarding and Control Element Separation (ForCES) framework, which separates control and packet forwarding elements in a ForCES Network [13]-[15]. ForCES defines two logic entities called the Forwarding Element (FE) and the Control Element (CE), both of which implement the ForCES protocol to communicate. The FE is responsible for using the underlying hardware to provide per-packet handling. The CE executes control and signaling functions and employs the ForCES protocol to instruct FEs on how to handle packets. The protocol works based on a master-slave model, where FEs are slaves and CEs are masters. An important building block of the ForCES architecture is the LFB (Logical Function Block). The LFB is a well-defined functional block residing on the FEs that is controlled by CEs via the ForCES protocol. The LFB enables the CEs to control the FEs' configuration and how FEs process packets.

In 2005, Greenberg et al. proposed a 4D approach [16]-[17], introducing a clean slate design of the entire network architecture with four planes.

In 2006, the PCE architecture was presented to compute label switched paths separately from actual packet forwarding in MPLS and GMPLS networks [18].

In 2007, Casado et al. presented Ethane, where simple flow-based Ethernet switches are supplemented with a centralized controller to manage admittance and routing of flows [19].

Commercial networking devices have also adopted the idea of data/control plane separation. SDN provides programmability through decoupling of control and data planes. SDN offers simple programmable network devices rather than making networking devices more complex as in the case of active networking.

SDN proposes separation of control and data planes in the network architectural design. With this design, network control can be done separately on the control plane without affecting data flows. As such, network intelligence can be taken out of switching devices and placed on controllers. At the same time, switching devices can now be externally controlled by software without onboard intelligence.

Most current network devices have control and dataflow functionalities operating on the same device. The only control available to a network administrator is from the network management plane, which is used to configure each network node separately. The static nature of current network devices doesn't permit detailed control-plane configuration. This is exactly where SDN comes into the picture.

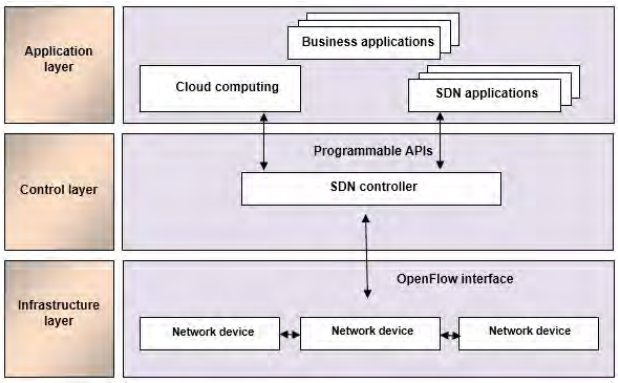


Figure 5. Basic SDN-based network architecture

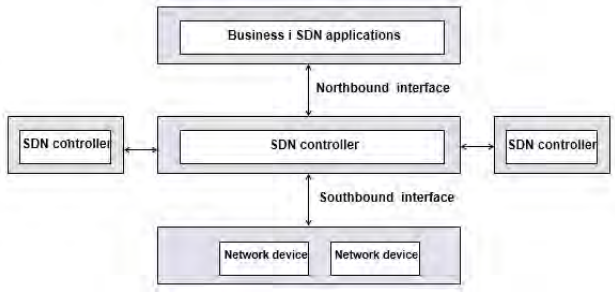


Figure 6. API directionality in SDN architecture.

The ultimate goal of SDN as defined in [20] is to "provide open user-controlled management of the forwarding hardware of a network element."

Compared to legacy networks, there are four additional components in SDN [21–24] (Fig. 5. and Fig. 6.)

The control plane/controller presents an abstract view of the complete network infrastructure, enabling the administrator to apply custom policies/protocols across the network hardware. The network operating system (NOX) controller is the most widely deployed controller.

The *"northbound"* application programming interfaces (APIs) represent the software interfaces between the software modules of the controller platform and the SDN applications running atop the network platform.

In the case of a multi-controller-based architecture, the East-West interface protocol manages interactions between the various controllers.

The data plane represents the forwarding hardware in the SDN network architecture. Because the controller needs to communicate with the network infrastructure, it requires certain protocols to control and manage the interface between various pieces of network equipment. The most popular *"southbound protocol"* is the OpenFlow protocol.

III. OPENFLOW ARCHITECTURE

OpenFlow is the protocol used for managing the southbound interface of the generalized SDN architecture. It is a communication protocol that gives access to the forwarding plane of a switch or router over the network. It is the first standard interface defined to facilitate interaction between the control and data planes of the SDN architecture.

OpenFlow was first proposed by Stanford University to enable the Flow Table application in Switch [25], and then application in MPLS networks with a simpler and extensible control plane [26]. On European Conference on Optical Fibre Communication 2011, a software defined packet over optical networks solution is implemented based on the OpenFlow and GMPLS control plane integration [27], and OpenFlow based wavelength path control is demonstrated for lightpath provisioning in transparent optical networks [28]. Interworking between OpenFlow and Path Computation Element (PCE) is presented for dynamic wavelength path control in multi-domain WSON [29]. Although OpenFlow has been designed for various applications in optical networks, numeric result comparison between GMPLS and OpenFlow based control over large scale optical networks has never been presented. Only Saurav Das from Stanford University gave the comparison between GMPLS and OpenFlow/SDN for IP over Optical networks without any numeric results [30].

OpenFlow provides software-based access to the flow tables that instruct switches and routers how to direct network traffic. Using these flow tables, administrators can quickly change network layout and traffic flow. OpenFlow protocol provides a basic set of management tools which can be used to control features such as topology changes and packet filtering.

The OpenFlow specification is controlled and defined by the non-profit Open Network Foundation (ONF). Open Networking Foundation is a user-driven organization dedicated to the promotion and adoption of SDN through open standards development. Most of the networking hardware vendors such as CISCO, HP, and IBM offer switches and routers that use the OpenFlow protocol [31].

OpenFlow-compliant switches come in two main types: OpenFlow-only and OpenFlow-hybrid. OpenFlow-only switches support only OpenFlow operations, i.e., all packets are processed by the OpenFlow pipeline. OpenFlow-hybrid switches support both OpenFlow operations and normal Ethernet switching operations, i.e., traditional L2 and L3 switching and routing.

The OpenFlow architecture consists of numerous pieces of OpenFlow-enabled switching equipment which are managed by one or more OpenFlow controllers, as shown in Fig. 7. Network traffic can be partitioned into flows, where a flow could be a transmission control protocol (TCP) connection, packets with the same MAC address or IP address, packets with the same virtual local

area network (VLAN) tag, or packets arriving from the same switch port [22].

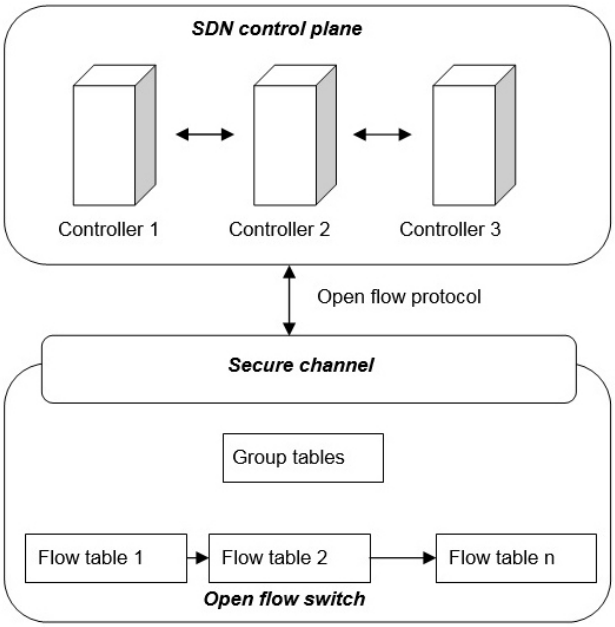


Figure 7. Basic architecture of OpenFlow.

An OpenFlow switch consists of one or more flow tables and a group table. It performs packet look-ups and forwarding.

Flow tables consist of flow entries, each of which determines how packets belonging to a flow will be processed and forwarded. Flow entries typically consist of: (1) match fields, or matching rules, used to match incoming packets; (2) counters, used to collect statistics for the particular flow, such as number of received packets, number of bytes and duration of the flow; and (3) a set of instructions, or actions, to be applied upon a match; they dictate how to handle matching packets [33].

The controller manages the OpenFlow-enabled switch using the OpenFlow protocol over a secure channel. The OpenFlow channel is the interface that connects each OpenFlow switch to a controller. The controller configures and manages the switch using this interface. The OpenFlow controller is responsible for maintaining all the network protocols and policies and distributing appropriate instructions to the network devices. It manages the switch flow table by adding and removing flow entries over the secure channel using the OpenFlow protocol. Typically, the controller runs on a network-attached server [34].

There are commercial controllers, such as ProgrammableFlow Controller from NEC and Big Network Controller from Big Switch, as well as open source controllers, such as OMNI [38], Trema [39], Ryu [40], Floodlight [41], NOX [35] and OpenDaylight [43].

Many OpenFlow software projects exist today. Among them NOX controller [42], and Mininet simulator [36], [37] are the most frequently used tools in SDN related researches. NOX is the first OpenFlow controller. It allows applications to be implemented based on a centralized network view using high level names as opposed to distributed algorithms over lowlevel addresses. Applications are written in either Python or C++ and are loaded dynamically. Core infrastructure and speedcritical functions of NOX are implemented in C++. Mininet is a network simulator for rapidly prototyping of a large OpenFlow network. Mininet hosts run standard Linux network software, and its switches support OpenFlow for highly flexible custom routing and Software-Defined Networking.

IV. WHY OPENFLOW/SDN CAN SUCCEED WHERE GMPLS FAILED

As a Unified Control Plane (UCP) technology, General Multi-Protocol Label Switching (GMPLS) has been deployed in current optical networks. GMPLS is ultimately just a collection of control protocols and merely extending them for the unique characteristics of the circuit-world. However, it has not been used widely because of some commercial reasons. In fact, GMPLS has failed in terms of actual deployment in wide area networks. It has undergone a lengthy standardization process at the ITU (International Telecommunication Union), IETF and OIF (Optical Internetworking Forum). All transport equipment vendors have implemented it in their switches. There have been many interoperability demonstrations and yet after a decade, there hasn't been even one significant commercial deployment of GMPLS as a UCP. GMPLS may have found use in a limited way as a tool to help the NMS/EMS (Network Management System/Element Management System) provision a circuit after being triggered manually by the management system.

Which is more suitable for the control over large scale optical networks, GMPLS or OpenFlow? This question is mainly answered in the analysis presented in the homonymous work [32]. Two testbeds based on GMPLS and OpenFlow are built respectively to validate their performance over large scale optical networks. Blocking probability, wavelength utilization and lightpath setup time are shown on the topology with 1000 nodes.

It should be noted that there are mainly two control architectures for multi-domain optical networks. One is implemented based on GMPLS and PCE, and the other is based on OpenFlow. Some typical network parameters are gained from the two testbeds [32]. Numeric results show that OpenFlow architecture shows better performance than GMPLS architecture in large scale multi-domain optical networks.

Recently, OpenFlow has become the research focus, which is considered as the core protocol of SDN. The IETF also defined three architectural-model models for GMPLS usage: peer, overlay and augmented. The peer model has never found favor. In principle, you could use a single instance of GMPLS protocols across packet and circuit domains, treating it as a UCP with full information sharing between routers and transport switches (peer model). What did find favor is the IETF overlay model, which was formalized by the ITU into its own architectural description for Automatically Switched Optical Networks (ASON). The overlay model separates the control of IP and transport networks, by treating the IP network as an overlay network that runs a completely

separate instance of the control plane - i.e. the IP/MPLS network runs the IP/MPLS control plane and the transport network runs the GMPLS control plane, and no information about the networks is shared across any boundary/interface between the networks in either direction. A request for services can be made across an interface (UNI) between the two networks (Fig.8).

V. SDN RELATED RESEARCHES AND FUTURE RESEARCH

The success of SDN requires improvements and developments at all the three layers, including the infrastructure layer, the control layer, and the application layer.

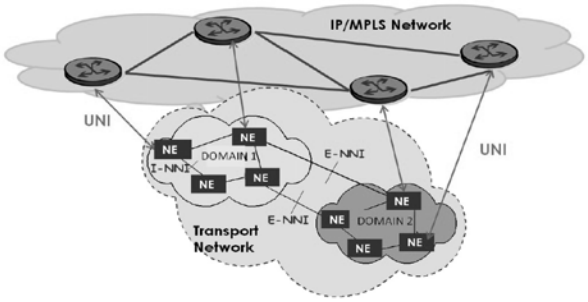


Figure 8. MPLS/GMPLS/ASON.

This chapter presents a literature survey of recent SDN researches in the infrastructure layer, the control layer, and the application layer, as well as future research of SDN. SDN related researches are summarized in Table I.

TABLE I. SDN RELATED RESEARCHES

Layer	*Issue*	*Approach /Reference*
Infrastructure	• Storage and Processing • Hardware Platform • Wireless Radio/Optical Fibers	• Combine different storage and processing hardware [44], [45] • Software/firmware implementation [46], [47], [48] • OpenRadio [49]/Unified control plane [50], [51], Split control plane [52], [53]
Control	• High Level Language • Network Status Collection and Synchronization • Increasing Processing Ability	• FML [54], Frenetic [55] • Open TM [56], HyperFlow [57], Onix [58] • Parallelism [59], [60]
Application	• Adaptive Routing • Network Virtualization • Green Networking • Cloud Computing	• Load balancing [61], cross-layer design [54], [62] • Flow Visor [63] • Elastic Tree [64] • Open vSwitch [65], VM Migration [66], [67], [68], [69]

Many challenges in SDN still need further research attention and many organizations have started research projects in various aspects of SDN as shown in Table II.

VI. CONLUSION

In this paper we discussed recent research activities related to the SDN/OpenFlow deployment for metro/backbone optical switching networks. We briefly presented the comparison between SDN/OpenFlow and other control plane techniques for optical networks.

In today's commercial IP/DWDM multilayer networks, different layers are separately operated without dynamic interaction, which leads to a low network efficiency, a high operational expense, and a long processing latency for path provisioning and restoration. With a decade of development and standardization efforts, traditional MPLS was extended to GMPLS to support multiple types of switching capabilities, especially the wavelength/fiber switching in the metro/backbone optical network, and accordingly, three different MPLS/GMPLS interconnection models (i.e., peer, overlay, and border peer) have been proposed for various multilayer network scenarios as a UCP.

However, to the best of our knowledge, there are still no commercial deployments of these models to date because the GMPLS UCP is too complicated. On the other hand, SDN and OpenFlow, which allow operators to control the network using software running on a network operating system (e.g., Nox) within an external controller, provide the maximum flexibility for the operator to control an optical network and match the carrier's preferences given its flexibility, simplicity, and manageability.

OpenFlow and SDN offer drastically reducing the complexity of the control plane, increasing programmability and upgrades with the gradual acceptance of the same. OpenFlow architecture, for a given scenario, shows better performance, in terms of less probability of blocking and the average time of establishing the light path, than GMPLS architecture in large multidomain networks. The challenges identified in this survey need further consideration in current and future research projects. It is essential to define optimal strategy transition from GMPLS control plane for optical multilayer networks to SDN control plane according to signaling migration.

TABLE II. CURRENT RESEARCH PROJECTS

Name	*Research Area*	*Organization*
ONRC (Open Networking Research Center)	• SDN Based Private Cloud; • Network Virtualization; • SDN Version Control; • Sergeant, a distributed operating system for software defined networks; • SDN for Dense Home Networks; • An SDN approach to MPLS Traffic Engineering and Virtual Private Networks; • P4: high-level language for programming protocol-independent packet processors; • Openradio and software	Stanford, Berkeley

	defined cellular wireless networking.	
BISMarc (The Broadband Internet Service Benchmark)	• BISmark is a research project and a colaboration effort between Georgia Tech, Princeton University and Google MLabs, created to develop an open platform for home broadband internet research.	Georgia Tech, Princeton University and Google Mlabs, University of Napoli Federico II
PANE (*Participatory Networking*)	• PANE is a prototype OpenFlow controller which implements Participatory Networking, an API for end-users, hosts and applications to take part in network management; • Participatory Networking is designed for networks with a single logical administrator, such as datacenters, enterprise or campus networks, and home networks.	Brown University
InCNTRE SDN Lab	• Developing testing tools, methodologies, and procedures; • Supporting OpenFlow events and showcases; • Contributing to Open Networking Foundation (ONF) working groups; • Providing educational opportunities.	Indiana University
Frenetic	• High-level language for programming distributed collections of network switches;	Princeton University, Cornell University
FP7 SDN related programs	• FEDERICA - Federated E-infrastructure Dedicated to European Researchers Innovating in Computing network Architectures; • FIBRE - Future Internet testbeds / experimentation between BRazil and Europe; • OFELIA - OpenFlow in Europe: Linking Infrastructure and Applications; • OFERTIE - OpenFlow Experiment in Real- time Internet Edutainment • SPARC - Split Architecture Carrier Grade Networks;	Mainly organizations in Europe
RouteFlow	• RouteFlow is an open source project to provide virtualized IP routing services over OpenFlow enabled hardware.	CPqD

REFERENCES

[1] T. Anderson, L. Peterson, S. Shenker, J. Turner, "Overcoming the internet impasse through virtualization," Computer 38 (4), 2005.

[2] HP, Realizing the Power of SDN with HP Virtual Application Networks, 2012.

[3] M. Grbavac, "Softverski definirane mreže," Mreža, Listopad, 2014.

[4] Open Networking Foundation (ONF). [Online]. Available: https://www.opennetworking.org/

[5] "Software-defined networking: The new norm for networks," Palo Alto, CA, USA, White Paper, Apr. 2012.

[6] J. Smith et al., "Switchware: Accelerating network evolution," CIS Dept., Univ. Pennsylvania, Philadelphia, PA, USA, White Paper, 1996.

[7] D. Alexander et al., "The SwitchWare active network architecture," IEEE Netw., vol. 12, no. 3, pp. 29–36, May/Jun. 1998.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] E. Kohler, R. Morris, B. Chen, J. Jannotti, and M. Kaashoek, "The click modular router," ACM Trans. Comput. Syst., vol. 18, no. 3, pp. 263–297, Aug. 2000.

[9] M. Handley, O. Hodson, and E. Kohler, "XORP: An open platform for network research," ACM SIGCOMM Comput. Commun. Rev., vol. 33, no. 1, pp. 53–57, Jan. 2003.

[10] Quagga Routing Software Suite. [Online]. Available: http://www.nongnu.org/quagga/

[11] The BIRD Internet Routing Daemon. [Online].

[12] N. Feamster, H. Balakrishnan, J. Rexford, A. Shaikh, and J. van de Merwe, "The case for separating routing from routers," in Proc. ACM SIGCOMM Workshop FDNA, 2004, pp. 5–12.

[13] L. Yang, R. Dantu, T. Anderson, and R. Gopal, "Forwarding and Control Element Separation (ForCES) Framework," Apr. 2004, RFC 3746.

[14] W. Wang et al., "Design and implementation of an open programmable router compliant to IETF ForCES specifications," in Proc. 6th ICN, 2007, pp. 1–6.

[15] A. Doria et al., "Forwarding and Control Element Separation (ForCES) Protocol Specification," Mar. 2010, RFC 5810.

[16] A. Greenberg et al., "A clean slate 4D approach to network control and management," SIGCOMM Comput. Commun. Rev., vol. 35, no. 5, pp. 41–54, Oct. 2005.

[17] H. Yan et al., "Tesseract: A 4D network control plane," in Proc. 4th USENIX Conf. NSDI, 2007, p. 27.

[18] A. Farrel, J.-P. Vasseur, and J. Ash, A Path Computation Element (PCE)-Based Architecture, Aug. 2006, RFC 4655.

[19] M. Casado et al., "SANE: A protection architecture for enterprise networks," in Proc. 15th Conf. USENIX-SS, Berkeley, CA, USA, 2006, vol. 15, p. 10.

[20] Q. Duan, Y.H. Yan, A.V. Vasilakos, A survey on service-oriented network virtualization toward convergence of networking and cloud computing, IEEE Trans. Netw. Serv. Manage. 9 (4) (2012) 373–392.

[21] Big Switch Networks, The Open SDN Architecture, 2012.

[22] M.K. Shin, Ki-Hyuk Nam, H.J. Kim, Software-defined networking (SDN): a reference architecture and open APIs, in: Proceedings, 2012 International Conference on ICT Convergence (ICTC), 15–17 October, 2012, pp. 360–361.

[23] IBM, Software-Defined Networking: A New Paradigm for Virtual, Dynamic, Flexible Networking, IBM Systems and Technology Thought Leadership White Paper, October 2012.

[24] M. Jammal, T.Singh, A. Shami, R. Asal, Y. Li, "Software defined networking: State of the art and research challenges", Computer Networks 72 (2014) 74–98

[25] N. McKeown, et al., "OpenFlow: Enabling Innovation in Campus Networks", SIGCOMM CCR, Vol. 38, Issue 2, March 2008.

[26] Saurav Das, Ali Reza Sharafat, Guru Par, et al., "MPLS with a Simple OPEN Control Plane," OFC2011, Los Angeles, USA, Mar.2011.

[27] S. Azodolmolky, R. Nejabati, E. Escalona, et al., "Integrated OpenFlow–GMPLS Control Plane: An Overlay Model for SoftwareDefined Packet over Optical Networks," ECOC 2011, Geneva, Switzerland, Sep.2011.

[28] L. Liu, T. Tsuritani, I. Morita, et al., "OpenFlow-based Wavelength Path Control in Transparent Optical Networks: a Proof-of-Concept Demonstration," ECOC 2011, Geneva, Switzerland, Sep.2011.

[29] L. Lei, C. Ramon, T. Takehiro, et al., "Interworking Between OpenFlow and PCE for Dynamic Wavelength Path Control in Multi-Domain WSON," OFC 2012, Los Angeles, USA, Mar.2012.

[30] S. Das, G. Parulkar, N. McKeown, "Why OpenFlow/SDN Can Succeed Where GMPLS Failed," ECOC2012, Amsterdam, Netherlands, Sep.2012.

[31] IBM, Software-Defined Networking: "A New Paradigm for Virtual, Dynamic, Flexible Networking," IBM Systems and Technology Thought Leadership White Paper, October 2012.

[32] Y. Zhao, J. Zhang, H. Yang, Y. Yu, X. Niu, X. Lin, "Which Is More Suitable for the Control over Large Scale Optical Networks, GMPLS or OpenFlow?," OFC/NFOEC Technical Digest, 2013

[33] B.N. Astuto, M. Mendonca, X.N. Nguyen, K. Obraczka, T. Turletti, "A Survey of Software-Defoned Networking: Past, Present, and Future of Programmable Networks," HAL, January 2014

[34] Open Network Foundation, OpenFlow Switch Specification v1.3.0, June 2012.

[35] N. Gude et al., "NOX: Towards an operating system for networks," SIGCOMMComput.Commun.Rev., vol. 38, no. 3, pp. 105–110, Jul. 2008.

[36] B. Lantz, B. Heller, and N. McKeown, "A network in a laptop: Rapid prototyping for software-defined networks," in Proc. 9th ACM SIGCOMM Workshop HotNets-IX, 2010, pp. 19:1–19:6.

[37] Mininet Overview [Online]. Available http://mininet.org/overview

[38] D. Mattos et al., "OMNI: OpenFlow management infrastructure," in Proc. Int. Conf. NOF, 2011, pp. 52–56.

[39] Trema. [Online]. Available: http://trema.github.io/trema/

[40] Ryu. [Online]. Available: http://osrg.github.io/ryu/

[41] Floodlight. [Online]. Available: http://www.projectfloodlight.org/

[42] NOX. [Online]. Available: http://www.noxrepo.org/

[43] OpenDaylight. [Online]. Available: http://www.opendaylight.org/

[44] O. Ferkouss et al., "A 100 Gig network processor platform for OpenFlow," in Proc. 7th Int. CNSM, 2011, pp. 1–4.

[45] V. Tanyingyong, M. Hidell, and P. Sjodin, "Using hardware classification to improve pc-based OpenFlow switching," in Proc. IEEE 12th Int. Conf. HPSR, 2011, pp. 215–221.

[46] E. Kohler, R. Morris, B. Chen, J. Jannotti, and M. Kaashoek, "The click modular router," ACM Trans. Comput. Syst., vol. 18, no. 3, pp. 263–297, Aug. 2000.

[47] M. Anwer, M. Motiwala, M. Tariq, and N. Feamster, "SwitchBlade: A platform for rapid deployment of network protocols on programmable hardware," ACM SIGCOMM Comput. Commun. Rev., vol. 40, no. 4, pp. 183–194, Oct. 2010.

[48] G. Lu et al., "Serverswitch: A programmable and high performance platform for data center networks," in Proc. NSDI, 2011, p. 2.

[49] M. Bansal, J. Mehlman, S. Katti, and P. Levis, "OpenRadio: A programmable wireless dataplane," in Proc. 1st Workshop HotSDN, 2012, pp. 109–114.

[50] S. Das, G. Parulkar, and N. McKeown, "Simple unified control for packet and circuit networks," in Proc. IEEE/LEOSST Meet., 2009, pp. 147–148.

[51] V. Gudla et al., "Experimental demonstration of OpenFlow control of packet and circuit switches," presented at the Optical Fiber Communication Conference, San Diego, CA, USA, 2010, Paper OTuG2.

[52] B. Raghavan et al., "Software-defined internet architecture: Decoupling architecture from infrastructure," in Proc. 11th ACM Workshop Hot Topics Netw., 2012, pp. 43–48.

[53] S. Azodolmolky et al., "Integrated OpenFlow-GMPLS control plane: An overlay model for software defined packet over optical networks," Opt. Exp., vol. 19, no. 26, pp. B421–B428, Dec. 2011.

[54] A. Ferguson, A. Guha, J. Place, R. Fonseca, and S. Krishnamurthi, "Participatory networking," in Proc. Hot-ICE, San Jose, CA, USA, 2012, p. 2.

[55] N. Foster et al., "Frenetic: A high-level language for OpenFlow networks," in Proc. Workshop PRESTO, 2010, pp. 6:1–6:6.

[56] A. Tootoonchian, M. Ghobadi, and Y. Ganjali, "OpenTM: Traffic matrix estimator for OpenFlow networks," in Passive and Active Measurement. Berlin, Germany: Springer-Verlag, 2010, pp. 201–210.

[57] A. Tootoonchian and Y. Ganjali, "HyperFlow: A distributed control plane for OpenFlow," in Proc. INM/WREN, 2010, p. 3.

[58] T. Koponen et al., "Onix: A distributed control platform for large-scale production networks," in Proc. 9th USENIX Conf. OSDI, 2010, pp. 1–6.

[59] Z. Cai, "Maestro: Achieving scalability and coordination in centralized network control plane," Ph.D. dissertation, Rice Univ., Houston, TX, USA, 2011.

[60] A. Tootoonchian, S. Gorbunov, Y. Ganjali, M. Casado, and R. Sherwood, "On controller performance in software-defined networks,"

[61] N. Handigol, S. Seetharaman, M. Flajslik, R. Johari, and N. McKeown, "Aster∗ x: Load-balancing as a network primitive," in Proc. 9th GENI Eng. Conf. (Plenary), 2010, pp. 1–2.

[62] G. Wang, T. E. Ng, and A. Shaikh, "Programming your network at runtime for big data applications," in Proc. 1st Workshop HotSDN, 2012, pp. 103–108.

[63] R. Sherwood et al., "Flowvisor: A network virtualization layer," OpenFlow Switch Consortium, OPENFLOW-TR-2009-1, 2009.

[64] B. Heller et al., "ElasticTree: Saving energy in data center networks," in Proc. 7th USENIX Conf. NSDI, 2010, p. 17.

[65] J. Pettit, J. Gross, B. Pfaff, M. Casado, and S. Crosby, "Virtual switching in an era of advanced edges," in Proc. 2nd Workshop DC-CAVES, 2010, pp. 1–7.

[66] P. Pisa et al., "OpenFlow and Xen-based virtual network migration," in Communications: Wireless in Developing Countries and Networks of the Future. Berlin, Germany: Springer-Verlag, 2010, pp. 170–181.

[67] V. Mann, A. Vishnoi, K. Kannan, and S. Kalyanaraman, "CrossRoads: Seamless VM mobility across data centers through software defined networking," in Proc. IEEE NOMS, 2012, pp. 88–96.

[68] Y. Pu, Y. Deng, and A. Nakao, "Cloud rack: Enhanced virtual topology migration approach with open vswitch," in Proc. ICOIN, 2011, pp. 160–164.

[69] E. Keller, S. Ghorbani, M. Caesar, and J. Rexford, "Live migration of an entire network (and its hosts)," in Proc. 11th ACM Workshop HotNets- XI, 2012, pp. 109–114.

About the Telco Cloud Management architectures

Ivan Nenadić, Davor Kobal, Domagoj Palata
Ericsson Nikola Tesla, Zagreb, Croatia
ivan.nenadic@ericsson.com

Abstract - In order to achieve the expected benefits provided by network function virtualization and software defined networks, the key role is placed on management and orchestration functionality with goal to automate services provider business processes. Network function virtualization and software defined networks establish new requirements upon network management systems which demand a change of existing management architectures and currently implemented management systems. The biggest challenge is to achieve unified multivendor monitoring and management for both virtualized and non-virtualized networks. To address these new requirements and enable further evolution and implementation of virtualized software networks, there is a clear need to define common framework and agree upon common information model between different currently available standards defined by many different standardization bodies.

The paper contains the overview of currently available important standards in the area of telecom cloud management as well as their short analysis with focus on network and services assurance systems – alarm monitoring and management, performance management, inventory management and service quality management. These are compared with today existing management architectures and systems in order to identify challenges and define requirements that will further drive the solutions in respect to new standardized framework. There are also some improvement guidelines for further considerations of management architectures.

I. INTRODUCTION

Network Function Virtualization (NFV), Virtualized Network Functions (VNF) and Software Defined Networks (SDN) provide new capabilities for telecom service providers with expected benefits in respect to better agility and increased operational efficiency as well as decreased capital and operative expenditures. Better agility is reflected in terms of increased flexibility, faster time to market for new services, scalability and elasticity of network architecture and network resources. Decreased costs are mainly seen in improved and simplified network build and maintenance.

However, in order to achieve these benefits the key role is to establish efficient management and orchestration through network and service management systems. Their function is to automate service providers' business processes with goal to achieve maximum operational efficiency. Telecom service providers were striving to that goal for last decades as industry was driven by mobile and wireline convergence, telecom and IT convergence, new digital services, new service value chains, etc. The introduction of NFV and SDN technology has greatly impacted the existing management architectures and harmonization is required across many standardization bodies that deal with telecommunication management. Currently Management and Orchestration architectural framework (MANO) for NFV defined by ETSI has been used as reference architecture for many different Telco Cloud management solutions. Nevertheless there is still some confusion in respect to interoperability with "legacy" non-virtualized (physical) networks, ability to clearly distinguish between resource and service management abstraction layers, as well as ability to realize resource and service assurance (fault and performance management). This paper examines the challenges of Telco Cloud management architectures as is structured as follows: Section II describes existing important standards related to management architectures, Section III discuss MANO architecture while Section IV provides analysis of MANO architecture in comparison with existing management architectures identifying challenges and requirements that will further drive the solutions in respect to new standardized framework and providing some considerations and guidelines for further improvement of management architectures.

II. RELATED WORK

The most adopted standardization role in this domain is led by TeleManagement Forum (TMF) through its Frameworx framework which is holistically combining other frameworks - Enhanced Telecom Operation Map (eTOM) as Business Process Framework, Shared Information and Data Model (SID) as information data model framework, Telecom Application Map as application functional architecture, Integration framework as well business metrics framework. These frameworks have greatly contributed to design and development of today's Operational Support Systems (OSS) and Business Support Systems (BSS) by taking the "top-down" modelling approach in architecture design starting from business process map (eTOM) that define and categorize business requirements, continuing across common information model of telecommunication business entities (SID), up to functional map and definition of applications/systems (TAM) supporting previously identified business processes and mutually integrated through a common integration framework based on service oriented architecture.

All TMF frameworks (eTOM, SID and TAM) have horizontal layers referencing the different management logical abstraction layers. The concept can be related to the original ITU-T TMN model that has identified element

management layer, network management layer, service management layer and business management layer. The idea behind such layered abstraction model is to lower the complexity, ensure the independent lifecycle of each layer, and ensure that higher abstraction layer is defining demands and requirements towards the lower abstraction layer that exposes its management capabilities.

The approach towards defining telecom network and service management is based on introduction of management model described through information model and standardized interfaces which are final goal of ensuring interoperability and enable management as such. Both ITU-T and 3GPP standardization approach is based on three mutually interrelated architectures: functional, information and physical having management business process requirements as a starting point. Functional architecture defines functions that need to be provided by implemented management system while information architecture defines information that need to be processed and exchanged in order to realize previously defined functions and business processes. The physical architecture represents the actual implementation of management functions and should ensure interoperability between different management systems. The management information model is object oriented and independent of implementation, technology and protocol. It represents the abstraction of management aspects (network resources and services with their characteristics) and defines the information that need to be exposed and exchanged through standardized way. Thus a functional architecture reference point to interact and exchange information between certain functional blocks is defined as subset of exposed information that needs to be specified for certain functional block.

However, there is no unique management information model today as different standardization organizations have defined their own information models depending of their subject of focus - 3GPP Network Resource Management (NRM) model for mobile networks, TMF Multi Technology Network Management model (MTNM) for transport networks, DMTF Common Information Model (CIM) for IT environment, etc. In relation to NFV and SDN it is worth emphasizing DMTF Cloud Management Initiative which resulted with Cloud Infrastructure Management Interface (CIMI) that models the management of services and the operations and attributes of the cloud service lifecycle. DMTF also defined Open Virtualization Format (OVF) standard [1] which is based on DMTF CIM and provides the industry with a standard packaging format for software solutions based on virtual systems. To enhance the portability of cloud applications and services, OASIS has defined Topology and Orchestration Specification for Cloud Applications (TOSCA) which is a standard language to describe cloud service components and their relationships using a service topology and for describing the management procedures that create or modify services using orchestration processes [2]. The adoption of TOSCA as a meta-model provides the ability to model and define VNF Forwarding Graphs (VNFFG) also known as Service Chains as well as other different entities such as Virtual Links (VL), NS (Network Service), VNFs. Thus TOSCA provides the abstract description of network function configuration such as virtual machine software (i.e., VM image provided by an ISV), types of interfaces, elasticity rules, metrics to be collected and monitored, network information, HA requirements, resources required to deploy the VNF, etc.

III. NFV Management And Orchestration Framework

Although many different standardization bodies, organizations and programs are dealing with NFV and SDN research and specifications, the most adopt work has been done by ETSI defining NFV Architectural Framework [3] and Management and Orchestration framework (MANO) [4]. In the contrast to TMF Frameworx approach, the MANO framework can be seen

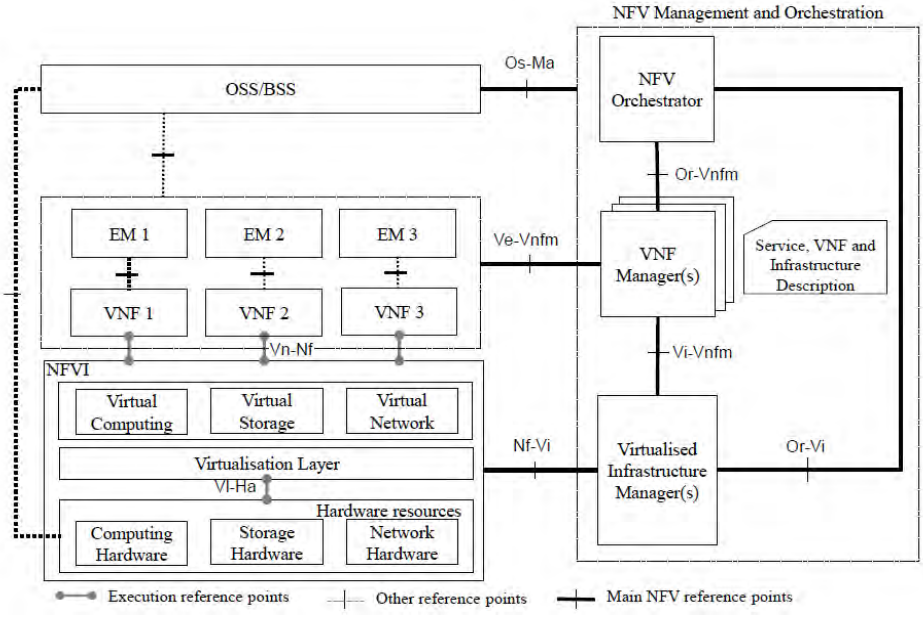


Figure 1. ETSI NFV reference architectural framework [3]

as a "bottom-up" approach to address management requirements as its interfaces are reflecting operations around its main information entities reflecting virtualization technology.

As seen on figure 1. the NFV reference architecture defines the VNFs domain that represent software implementation of network functions running over and NFV Infrastructure (NFVI) that is comprised of physical hardware resources virtualized through virtualization layer.

The MANO architecture represents the functional architecture and comprises of following functional blocks: Virtualized Infrastructure Manager (VIM) responsible for managing the NFVI, VNF Manager (VNFM) responsible to manage VNFs, NFV Orchestrator (NFVO) responsible to manage the Network Service (NS) and Service, VNF and Infrastructure description. Such architecture is further detailed in figure 2.

The NFVO function orchestrates the provisioning of NSs and VNFs based on their specifications defined in NS Catalog and VNF Catalog correspondingly. As such it manages the lifecycle of the NS including NFVI resource orchestration. The instantiated NFVs are persisted in the NFV Instances while the NFVI Resources repository manages the information about available/reserved/ allocated NFVI resources. The VNFM function manages the lifecycle of VNF instances while Virtual Infrastructure Manager (VIM) manages the NFVI compute, storage and network resources having interfaces towards hypervisors and Network Controllers (SDN Controllers).

The MANO architecture defines reference points between certain MANO function blocks, but also between existing domains outside MANO architecture thus trying to position MANO in the whole OSS/BSS environment. The Os-Nfvo is used to exchange the information between other OSS/BSS systems and NFVO. VeEn-Vnfm exchanges information between EMS responsible for certain VNF and/or more VNF types and VNFM while the VeNf-Vnfm exchanges information between VNFs and VNFM. The Nf-Vi reference point is used between VIM and NFVI.

MANO framework has also defined the information model thus describing certain entities (information elements). Based on defined functional architecture and information model, MANO framework defines interfaces that could be mapped to functional architecture reference points allowing certain interfaces to be applicable to many reference points. Such characteristic allows support for many different use cases allowing for innovation, but on the other hand it is not leading to concrete standard and best-practice decision agreed by the whole community eco-system.

It could be argued that existing MANO framework has not reached the sufficient mature level as there are still concerns that need to be addressed. To some extend this has also been noted by ETSI with harmonization effort being done across other standardization organizations. It is worth emphasizing the overview provided by [5] as part of T-NOVA project with focus to design and implement a management/orchestration platform for the automated provision, configuration, monitoring and optimization of Network Functions-as-a-Service (NFaaS) over virtualized Network/IT infrastructures. TMF has launched the Zero-touch Orchestration, Operations and Management (ZOOM) program to address the new management requirements imposed by NFV and SDN. This has yield with reports referencing the impact of NFV/SDN and needed changes and alignment s to be done in Frameworx [6][7][8][9]. For example [10] and [11] noted the need to

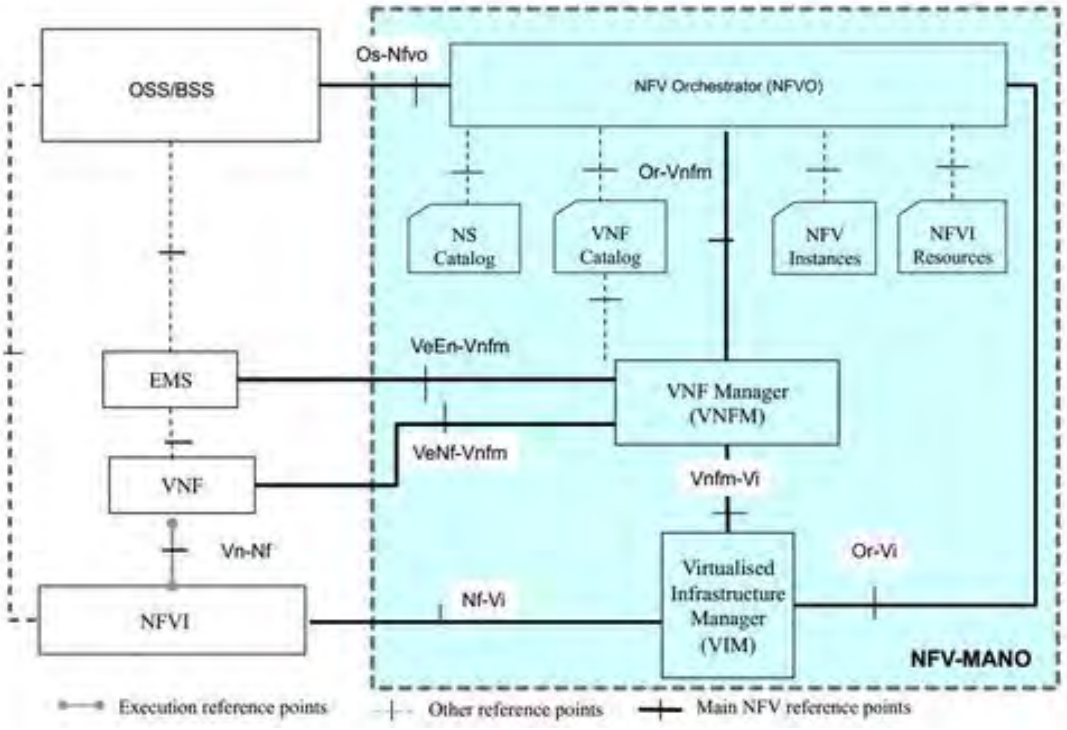


Figure 2. ETSI NFV-MANO architectural framework with reference points [4]

adjust the SID since entities such as Network Service, VNF Forwarding Graph and VNF are not part of SID today as its' concept is built upon Product-Service-Resource relationship and the concept of Customer Facing Services (CFS) and Resource Facing Service (RFS). Thus a need to standardized service template is essential to record service characteristics in a consistent form.

In order for ETSI MANO architecture to provide sufficient maturity level and enable further direction for physical architectures and their deployments allowing for modular multivendor SW architectures, there is a need to analyze it through its ability to interwork with existing OSS/BSS environments and enable smooth transition from existing management concepts. The analysis is focused on the assurance management aspects (including also inventory management) by taking into account Telco Cloud characteristics and management demands and interworking between existing management systems and their architectures.

IV. MANO Analysis and Comparison

A. *Telco Cloud characteristics and management demands*

In order to achieve the expected benefits, the characteristics of Cloud services, NFV and VNFs are imposing some new management demands. In comparison with todays "physical" networks we can see that assets and entities that are subject of management are virtual and potentially mobile (decoupled from physical HW location) with short life as VNFs are being created and deleted based on user demand for self-service, resource pooling and rapid elasticity. The NFV and VNF life-cycle is shorter across the whole management business processes phases as defined by eTOM. For example, in the Infrastructure and Product planning there is a requirement for quick development of new services using pre-built, re-usable and well defined components; the Fulfillment requires near real-time provisioning and activation on demand; Assurance requires automatic

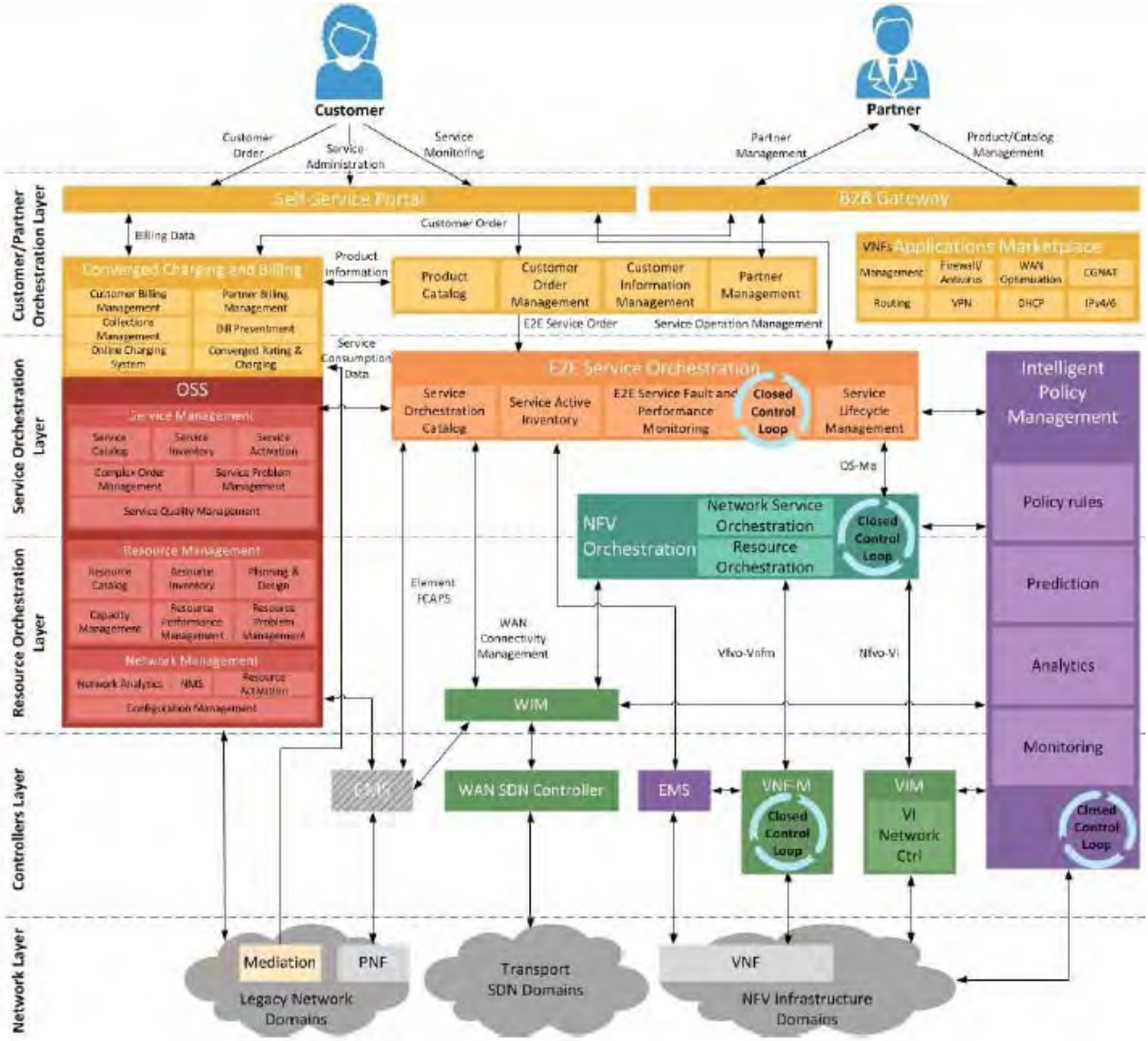


Figure 3. High level architecture for management of physical and virtual networks [8]

initiation of monitoring and measurement, etc. This is again indirectly impacting the whole OSS/BSS system architecture. For example, the integration process of new vendor "physical" network elements with the corresponding management system was ranging from couple of days in design phase through couple of hours in integration itself, depending on the available management interface support, and was primarily human driven work activity. On the other hand, the introduction of VNF during deployment phase requires "auto-integration" so that its management system is automatically configured and ready to monitor and manage its operational state reflecting both the VNF as well as its associated NFVI resources. Thus, the development of management capabilities for certain Network Service inside service providers MANO (OSS/BSS) environment has to be done in parallel with Network Service and VNF service components development themselves. This includes integration with different management components that form a Cloud Management System.

Unlike the IT Cloud where management is mainly oriented towards management of virtual machines (VM) the distributed resource allocation characteristic of Telco Cloud introduces additional complexity requiring control plane to be de-coupled from the devices and separated from the data plane (under management responsibility of SDN Controller). However, as noted also under [12], the current vendor approach towards virtualizing network functions is to encapsulate existing "physical" network functions as a big, stateful VM image due to lower cost of such approach and to preserve the control of management of the application. Such approach is based on the assumption that NFVI environment would ensure "carrier-grade" environment that never fails. The opposite approach needed by telco cloud is to expect VNF application to be highly tolerant on infrastructure failures so that VIM and virtualization layer orchestrate cloud infrastructure that supports stateless web applications that can be simply re-created in case of infrastructure failure.

Another important aspect is the ability to provide self-service management. The existing OSS/BSS systems were predominantly designed to provide management of certain service provider environment that is under his administrative domain. This is today highly being impacted by the multi-tenancy cloud characteristic that requires OSS/BSS system to separate management data between different tenants to ensure security when providing self-service management to different customers.

B. Interworking of MANO and existing OSS/BSS systems

By analyzing MANO framework, there are some overlaps between VNFM and EMS as well as between NFVO and OSS/BSS functional blocks. For example, the Catalog driven order management architectures that are following a principle of order decomposition and orchestration based on well-defined catalogued services and their relationships (CFS-RFS service model) as well as decomposition rules (policies) can be found in today's OSS/BSS environments and can be used to orchestrate the NFV services as well. On the other hand, according to ETSI, orchestration is provided in multiple functional blocks inside MANO architecture. Another example is VNFM that should enable management of VNFs which can also be managed by EMS systems providing FCAPS functionality. The goal of having standardized framework is to yield with architecture with de-duplicated functions and data stores and enable their efficient coordination.

Although SDN is not directly referenced in ETSI MANO architecture, it is important part of NFVI and as such it realized network policy control to enable separation of the control plane from the data plane. As depicted also in figure 3., the SDN (WAN SDN in this case) needs to be coordinated with orchestration functions with policies be used to define and control services and resources.

While ETSI MANO describes three functional layers, current results of these overlaps and gaps are NFV MANO implementations and Cloud management systems that don't follow the ETSI MANO functional block decomposition, but rather focus either on implementing both NFVO and VNFM in a single platform or implementing both VIM and VNFM functionalities in a single platform (see also [12] and [13]). This is potentially creating silo based management solutions unable to provide interoperability in a multivendor environment where VNFs from one equipment provider can be orchestrated by NFV MANO solution of another.

These challenges still need to be addressed in order to tackle the problem of co-existence with existing OSS/BSS systems and enable migration towards new hybrid network management.

C. Separation between resource and service management abstraction layers

ETSI MANO doesn't provide direct and clear distinction between resource and service management layers as defined by TMF (see figure 3). For example, NFVO is responsible for both global resource and service orchestration although orchestration is provided in multiple functional blocks inside MANO architecture (e.g. VIM). The distinction between resource and service layer in existing OSS/BSS architectures was a driver for independent development of management system as it lowered the complexity. Having separate service management layer would enable management of platform-independent entities (NS) while resource management layer would handle the management of platform-dependent entities on different VIMs and NFVIs [14]. Therefore, abstraction at each layer is important to enable flexibility and de-coupling of products from infrastructure allowing for independent spheres of management responsibility and avoid monolithic systems that reference all management processes.

D. Inventory management

The importance of inventory information has been addressed inside MANO architecture via NFV Instances repository and NFVI Resources repository components. All other management processes and systems (e.g. provisioning, fault management, performance management, service quality management, capacity planning, etc.) need reference data to provide their

functions. Thus, their accurate functioning depends on having in-sync consolidated inventory information about instantiated NS and its components, associated deployed VNFs and their internal structure, VNF associated virtualized resources (NFVI) as well as their physical resources.

Following the existing legacy OSS Inventory systems the information managed inside was mainly static with changes being reconciled mainly on weekly level for physical inventory) and daily level for logical and service inventory. The NFV inventory management requires the ability to have dynamic relationships and topology discovered and reconciled at much shorter periods managing the mobility characteristics of certain VNFs as noted before. Therefore, the inventory management systems should enable flexible and adaptive data modelling following the service models specified in Catalog components and ensuring efficient mechanisms for data accuracy and integrity across multivendor environment (orchestration process managing the instantiation of inventory objects, VNFM lifecycle process managing the VNF changes, discovery and reconciliation process towards VIM, etc.).

E. Fault management

The role of today's fault management systems is to capture and collect different events across entire multi-vendor network environment, consolidate them through event normalization and deduplication, enable their correlation and alarm definition and present them to operator in order to support Resource Performance and Trouble management processes providing efficient root cause and impact analysis. There are different monitoring methods ranging from agent-based to agentless monitoring, from network elements monitoring to monitoring interfaces via probes, from Deep Packet Inspection (DPI) to application performance monitoring (APM), etc. Implementing these methods in NFV environment would re-quire pre-build models associated with catalog services/resources to incorporate fault monitoring capabilities (e.g. SNMP agent) so that these could be incorporated during VNF instantiation (or NS instantiation) and "automatically" integrated with fault management function inside management system (e.g. VNFM or OSS) during service provisioning.

The fault management functionality in ETSI MANO is not explicitly defined, but is rather scattered among VNF, EMS, VNFM, VIM functional blocks. Additionally NFVO is set as being responsible for fault correlation on a NS level and further forwarding to OSS systems. Such gap is also noted by ETSI itself. Although ETSI deliberately doesn't specify the notifications mechanisms for fault management interfaces., TMF has proposed Resource Alarm Management (RAM) interface based on OSS/J to be used [15][16] for MANO fault management interfaces, but RAM has not been widely deployed in the industry. As previously discussed, the need to separate resource and service management layers could further simplify the event management as it would de-couple correlation of events related to resources and correlation of events impacting on service health.

With introduction of NFV the focus is being shifted from human operator surveillance and management towards auto-healing functionality (e.g. VNF healing) and automatic network optimization requiring constant automated closed loops to be created and triggered based on event correlation and polices (example of policy-based NFV Management and Orchestration is provided in [17]). Event correlation greatly depends on topology information requiring dynamic inventory to manage possible VNF mobility and understand the NS health state at any given time which is a function of different service and resource components (application, middleware, OS, compute, storage, LAN, WAN, etc.). This implies close integration between inventory management and fault management .

The automatic restoration and optimization requires policies (event correlation rules) to be based on sophisticated algorithms that would need to be driven by analytical insights and implemented by dynamically delivered Policy and Orchestration. This would again require Policy management in order to resolve policy conflicts and provide policy optimizations. Such architecture presents big shift for existing OSS/BSS systems as in its extreme scenario it would mean that their business logic is extracted from their functional components and managed in a separate Policy management system responsible for managing also policies across other management systems (that don't necessarily have to be provided from the same vendor). Although this would provide quite configurable and adaptive management architecture, this approach presents a challenge to build such architecture and still allow for multivendor interoperability at the same time, but this is out of scope for this paper.

F. Performance management

Performance management role is to collect performance measurements from variety of data sources, enable Key Performance Indicator (KPI) calculations and PI/KPI aggregation through different dimensions (time, managed objects, any other reference data, etc.). It also enables presentation and reporting towards operator in order to support Resource Performance management process providing performance monitoring and capacity planning. Although both fault and performance monitoring are focused towards supporting Assurance management processes, the SW architecture of their management systems is quite different due to different nature of their management information (distinct events usually generated and push towards management system vs. numerical counters referencing certain time period usually collected by management system).

Similarly as for Fault management, implementing performance management function in NFV environment requires pre-build models associated with catalog services/resources to incorporate performance monitoring capabilities. For example, having NS modelled in Catalog in a way so that its associated VNF from one vendor (later on instantiated VNF VM image) contains APM probe from another vendor allowing for seamless automatic integration with performance management function inside management system (e.g. VNFM or OSS) during service provisioning. Again, ETSI MANO does not explicitly

reference performance management functionality as it is again scattered among other functional blocks with no well-defined interfaces specified.

To be able to perform different PI/KPI aggregations, performance management system needs to have topology and inventory reference data thus requiring close integration with inventory management system. From the performance monitoring perspective the collected measurement values and calculated KPIs are transformed into events via performance management thresholds (again defined by policies). Such events are then subject of management on event (fault) management systems as they are further correlated to trigger closed feedback loop actions (already explained above).

G. *Service Quality assurance*

The end-to-end (E2E) QoS requires correlation of both fault and performance management data including inventory and topology information as referenced data to determine service quality. As such it does not reside only in the MANO domain, but is encompassed across non-virtualized networks collecting variety of data from multiple systems managing different network domains and technologies providing support for hybrid services and products delivered to end customers. It is also driven by need to establish customer experience assurance by correlating service impacting events to the individual customer experience in near real time and providing SLA data transparently to customer (measured service being one of the cloud characteristics).

The implementation of service assurance is based on definition of service models with KQIs and KPIs defined for its components. The traditional offline analytics approach now has to adapt to new paradigm ([18] and [19]). To enable NFV benefits, the service model in service assurance system has to model and monitor VNFs and Complex NS based on pre-build models associated with catalog services/resources and instantly instantiated during service provisioning. Again, these service models need to support the dynamic nature of virtualized services (for example, VNF's will move and the service model must be updated in real-time for any changes in the underlying composite services).

Although ETSI MANO defines service fault and performance management functionality to be part of NFVO functional block these are not directly referenced. Also without clear separation of service and resource management layers it will be difficult to address service quality assurance as such. Not having well-defined interfaces to collect required management data is again stumbling block in realizing E2E service assurance.

To address such complexity and challenges Big Data technology and Analytics are seen as possible solutions, but this is out of scope for this paper.

Additional aspects on management architecture changes due to introduction of NFV and SDN technology can be found in [20] and [21].

V. CONCLUSION

In order to achieve further progress in NFV management, the existing ETSI MANO framework has to be enhanced in respect to assurance management aspects and allow for more efficient interworking with existing OSS/BSS systems with minimal overlaps and gaps. It is also necessary to avoid creation of monolithic software physical architectures as final goal should be set towards achieving modular software architecture allowing for certain functional blocks to be implemented as separate physical components. Such architecture has to consider both telco cloud management demands and existing management architectures. The further work would be into assessing Policy management and Big data technology to address these issues.

Nevertheless, harmonization across existing management standards is necessary to ensure further progress. The need for a single unique information model has never been greater then today driven by the NFV and SDN management requirements. Having common information model has always been a practically unfeasible goal because of the need to meet and leverage the requirements of its generalization and simplicity on one hand and applicability and specialization in every domain on the other. The rapid technological advance requires fast update to the standardization model in parallel, but the standardization process has always been traditionally slower. By the time of writing this paper ETSI seeks harmonization of NFV information modelling across standards bodies and open source by gathering 3GPP, ATIS, Broadband Forum, DMTF, ETSI NFV, IETF, ITU-T SG 15, MEF, OASIS/TOSCA, Open Cloud Connect, ONF, OpenDaylight, OPNFV and TM-Forum to collaborate on a common workshop.

REFERENCES

[1] DMTF DSP0243, "Open Virtualization Format Specification", V2.1.1, http://www.dmtf.org, August 27, 2015.

[2] OASIS Standard, "Topology and Orchestration Specification for Cloud Applications Version 1.0", http://docs.oasis-open.org/tosca/TOSCA/v1.0/os/TOSCA-v1.0-os.html, November 25, 2013.

[3] ETSI Group Specification NFV 002, "Network Functions Virtualisation (NFV); Architectural Framework", http://www.etsi.org, V1.2.1 (2014-12)

[4] ETSI Group Specification NFV-MAN 001, "Network Functions Virtualisation (NFV); Management and Orchestration", http://www.etsi.org, V1.1.1 (2014-12)

[5] A. Gamelas, "Specification of the Infrastructure Virtualisation, Management and Orchestration - Interim", http://www.t-nova.eu. T-NOVA Deliverable D2.31, V1.0, work in progress, September 30, 2014.

[6] TM Forum TR228 Technical Report, "TM Forum Gap Analysis related to MANO Work", http://www.tmforum.org, Frameworx Release 14.5.1, V2.0.1, February 2015.

[7] TM Forum TR227 Exploratory Report, "TM Forum Specifications relevant for MANO Work", http://www.tmforum.org, Frameworx Release 14, V1.0.2, September 2014.

[8] TM Forum IG1130 Exploratory Report, "Future TAM Impacts by SDN/NFV", http://www.tmforum.org, Frameworx Release 15.5, V1.0.5, December 2015.

[9] TM Forum IG1120 Exploratory Report, "Virtualization Impact on SLA Management", http://www.tmforum.org, Frameworx Release 14.5.1, V14.0.12, April 2015.

[10] TM Forum TR178 Frameworx Best Practice, "Enabling End-to-End Cloud SLA Management", http://www.tmforum.org, Frameworx Release 14, V2.0.2, October 2014.

[11] TM Forum TR234 Technical Report, "ZOOM Information Model Snapshot", http://www.tmforum.org, Frameworx Release 14.5.1, V14.0.11, February 2015.

[12] C. Chappell, "NFV MANO: What's Wrong & How to Fix It", Heavy Reading, Vol. 13, No.2, February 2015.

[13] R. Mijumbi, J. Serrat, J. Gorricho, S. Latré, M. Charalambides, D. Lopez, "Management and Orchestration Challenges in Network Functions Virtualization", IEEE Communications Magazine, Vol. 54, Issue 1, pp. 98 - 105, 2016.

[14] Y. Chen, M. Lambe, Y. Qin, W. Chu, "Realizing Network Function Virtualization Management and Orchestration with Model Based Open Architecture", 11th International Conference on Network and Service Management (CNSM), pp. 410 - 418, 2015.

[15] TM Forum TR227-A Exploratory Report, "Mapping of TM Forum interface to ETSI NFV reference points", http://www.tmforum.org, Frameworx Release 14, V1.0.2, September 2014.

[16] TM Forum TR238 Exploratory Report, "TM Forum Interfaces: Fulfilling NFV MANO Interface Requirements", http://www.tmforum.org, Frameworx Release 14.5.1, V0.5.1, February 2015.

[17] C. Makaya, D. Freimuth, D. Wood, S. Calo, "Policy-based NFV Management and Orchestration", IEEE Conference on Network Function Virtualization and Software Defined Network (NFV-SDN), pp. 128 - 134, 2015.

[18] TM Forum IG1127 Exploratory Report, "End-to-end Virtualization Management: Impact on E2E Service Assurance and SLA Management for Hybrid Networks", http://www.tmforum.org, Frameworx Release 15, V1.0.0, May 2015.

[19] TM Forum GB963, "Cloud SLA Application Note", Approved Version 1.2, www.tmforum.org, October 2012.

[20] P. Kelly, G. Lenahan, "Managing the Telco Cloud: NFV and SDN role in the Existing OSS Domain", http://appledoreresearch.com, April 02, 2015.

[21] TM Forum IG1118 Exploratory Report, "OSS/BSS Futures - Architecture", http://www.tmforum.org, Frameworx Release 15.5, V2.0.2, November 2015.

CPE virtualization by unifying NFV, SDN and Cloud technologies

P. Cota, J. Šabec
* Ericsson Nikola Tesla d.d., Zagreb, Croatia
petar.cota@ericsson.com
jan.sabec@ericsson.com

Abstract - This paper will describe CPE virtualization solution for home and enterprise environments by unifying Network Function Virtualization (NFV), Software Defined Networking (SDN) and Cloud technologies in Internet providers' networks and data centers. The goal of this solution approach is to reduce operational and capital costs, increase network and service flexibility for providers and to broaden service offering and experience for the end-users. This is achieved by leveraging state of the art technologies that provide automation, flexibility and simplify operations. Possible applications, implementation and potentials will be studied throughout this paper.

I. INTRODUCTION

With commercialization of Internet access, telecom providers had to introduce new functionality in the demarcation point between end-user and internet service provider which is Customer Premises Equipment (CPE). Most often all end-user devices are connected to the same LAN behind one common IP address. Users hidden behind this IP address are provided connectivity through different types of broadband access technologies towards the Internet by the CPE. Depending on the end customer, residential or business, CPE contains multiple different functionalities, such as Firewall, NAT, Parental Control, etc. To fulfill different requirements and to distinguish themselves among competition, CPEs became more and more advanced, having a lot of features and functionalities from many different vendors on multiple different chipsets and hardware configurations. This makes managing CPEs lifecycle expensive and inflexible to introduce new services and functions. Other issue of this approach is that end-users are not visible inside telecom network which makes personalization of services per end-user impossible.

The limitations described above are trying to be solved with introduction of Virtual CPE Solutions (vCPE). There are several possible implementations of vCPE: on premises, in access and in cloud. In this paper the focus will be on the cloud implementation solution.

II. UNDERLYING TECHNOLOGIES

A. SDN

In the very beginning of networking, networks used to be quite simple. The applications of networking put very little requirements on the protocols and devices which also made designing and managing such networks easier.

With the development and commercialization of networking, where telecoms, enterprises and home users are putting a large demand on the network, network devices, network design and similar, operating and maintaining such complicated networks is becoming more and more difficult, time consuming and prone to human error.

Today's networks run a vast variety of devices such as routers, switches, firewalls, load balancers, intrusion detection/prevention systems etc. Those devices are developed by many different vendors, and although they conform to industry standards, each vendor has its own way of developing the product and different set of supported functions, proprietary closed operating system, and different command line interface for configuration of the device. This contributes to difficult, slow, inflexible and expensive managing of the network.

It was recognized very early, as far back as the mid-1990s. To overcome these obstacles, industry recognized the necessity to decouple the control from data plane and centralize the control plane. These two characteristics define what we call Software Defined Networking (SDN) today. The history of SDN can be divided into three stages [5]:

- active networks (from the mid-1990s to the early 2000s) - introduced programmable functions in the network to enable greater innovation;
- control and data plane separation (from around 2001 to 2007) - developed open interfaces between the control and data planes;
- OpenFlow API and network operating systems (from 2007) - represented the first instance of widespread adoption of an open interface and developed ways to make control-data plane separation scalable and practical.

Today there are several projects and foundations developing solutions for SDN, where two of the currently biggest contributors in telecom and IT segment are Open Networking Foundation (ONF) with OpenFlow and Open Networking Lab (ON.Lab) with ONOS. We will focus on ONF OpenFlow since it is used in the virtual CPE described in this paper.

OpenFlow is the first standard communications interface defined between control and forwarding layers of an SDN architecture. It allows direct access and manipulation of the forwarding plane of network devices such as switches and routers, both physical and virtual (hypervisor-based) [6]

By using SDN we are able to make networks[6]:

- Directly programmable - Network control is directly programmable because it is decoupled from forwarding functions.
- Agile - Abstracting control from forwarding lets administrators dynamically adjust network-wide traffic flow to meet changing needs.
- Centrally managed - Network intelligence is (logically) centralized in software-based SDN controllers that maintain a global view of the network, which appears to applications and policy engines as a single, logical switch.
- Programmatically configured - SDN lets network managers configure, manage, secure, and optimize network resources very quickly via dynamic, automated SDN programs, which they can write themselves because the programs do not depend on proprietary software.
- Open standards-based and vendor-neutral - When implemented through open standards, SDN simplifies network design and operation because instructions are provided by SDN controllers instead of multiple, vendor-specific devices and protocols.

This enables us to overcome many of the obstacles mentioned at the beginning of this chapter and to introduce the adaptability and dynamicity in networks required by today's applications, and significantly reduce operations and management complexity.

B. NFV

In SDN chapter we described how the networks started as very simple and during time which brought different applications and higher demands, they became complex and hard to manage. Network functions such as firewalls, network address translation (NAT), Dynamic Host Configuration Protocol (DHCP) and others, had a similar path. Functions above and those similar to them run on proprietary devices, developed by different vendors, where each device is based on different hardware, operating system, provides different functions and configuration interfaces. This makes them difficult to manage, upgrade, and expand. It became ever more prominent with lifecycle of hardware based appliances becoming shorter, resulting in little or no revenue benefit. Telecom operators have resolving these issues very high on the list of priorities, since it makes them less responsive to demands of the end user and market and, as the consequence, more expensive.

To resolve common issues, telecom providers came together in a group called European Telecommunications Standards Institute (ETSI). For the particular problem of network functions ETSI was chosen in November of 2012 to be Industry Specification Group (ISG) for Network Function Virtualization (NFV)[2].

Network Functions Virtualization aims to transform the way that network operators architect networks by evolving standard IT virtualization technology to consolidate many network equipment types onto industry standard high volume servers, switches and storage, which could be located in Datacentres, Network Nodes and in the end user premises. It involves the implementation of network functions in software that can run on a range of industry standard server hardware, and that can be moved to, or instantiated in, various locations in the network as required, without the need for installation of new equipment [3].

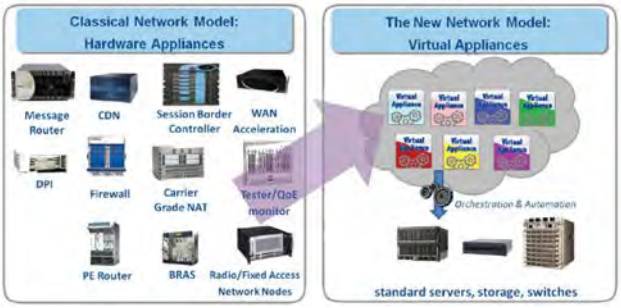


Figure II-1 Network Virtualization Approach

This provides several benefits[3]:

- Reduced operator CAPEX and OPEX through reduced equipment costs and reduced power consumption
- Reduced time-to-market to deploy new network services
- Improved return on investment from the new services
- Greater flexibility to scale up, scale down or evolve services
- Openness to the virtual appliance market and pure software entrants
- Opportunities to trial and deploy new innovative services at lower risk

This technology is highly complementary to SDN and by combining the two, impact on the flexibility and manageability of the network is even bigger. Having network functions on servers, makes the network more dynamic since the services can be easily moved across the servers. This adds to the demand on managing the network where SDN, which enables programmability of the network, can automate and simplify that task.

C. Cloud Computing

Data centers today host a wide variety of application and services which makes managing of those functions,

network connectivity and servers complex and expensive. Similar to SDN and NFV, this segment of the network was also identified that could benefit from automation and better manageability.

National Institute of Standards and Technology (NIST) defined Cloud Computing as following [4]:

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Essential characteristics of the cloud computing are:

- On-demand self-service. The consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service provider.
- Broad network access. Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, tablets, laptops, and workstations).
- Resource pooling. The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location independence that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, and network bandwidth.
- Rapid elasticity. Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.
- Measured service. Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

The combination of SDN, NFV and Cloud proved to be very synergistic and complementary to each other. With the introduction of NFV, there is still an issue of managing each individual instance and its lifecycle. Cloud computing enables process automation in the data center. On the other hand, automation of the lifecycle and possibility to autoscale network functions puts additional demand on networking. SDN enables automation of the network. By orchestrating all three technologies, they make excellent foundation of introducing solutions such as virtual CPE. The following chapter describes application of this solution its use and elaborates how it orchestrates these three technologies and provides the end-user and telecom with scalable, manageable and cost effective CPE solution.

III. VCPE SOLUTION

Virtual CPE solution is an end to end architecture with products that follow the open-source framework of SDN, NFV and Cloud and is flexible to accommodate different vendor building blocks. These building blocks might be networking related or just IT enterprise related. vCPE solution can be applied both for residential (home) users or business (enterprise) users.

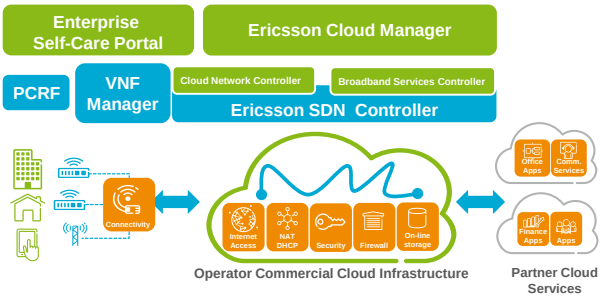


Figure III-1 vCPE architecture

In the vCPE solution the enterprise or home LAN is backhauled to operator domain. The outcome of this is:

- Basic services provided by the traditional CPE today (NAT, DHCP, FW) can be offered and provided by the operator domain.
- Additional optional services such as content filtering, video optimization, network attached storage, etc. can be offered by the operator to the end-users as a service.
- Possibility of Service Chaining for individual users. Service Chaining is the forwarding of data flows through a set of Service Functions (SF) which are selected according to the data flow characteristics. Extending the enterprise/home LAN into the operator's domain enables the operator to identify each individual end-user device in this enterprise/home. Hence, service chains can be configured and selected per end-user device.
- Support of overlapping IP addresses. IP addresses which are assigned to end-user devices are unique per enterprise/home but they may be overlapping between enterprises/homes. Hence, traffic between an end-user and the operator's domain must be encapsulated in a way that in the

operator's domain it can be associated with a dedicated enterprise/home. This encapsulation is done in the CPE and the can be achieved with different technologies such as VXLAN, GRE, L2TP, etc. In case of VXLAN, all traffic from and to one enterprise is identified by the same Virtual Network Identifier (VNI). The VNI identifies the enterprise/home and the tuple {VNI, IP} uniquely identifies an end-user device. In that way, VXLAN is used as tunneling technology with the VNI encoding not only transport but also service related information.

On the figure III-1 it can be seen that vCPE solution is built from several different components situated in different domains. On the high level we can divide the solution into these main domains:

- Access network
- Cloud
- SDN
- Management and orchestration
- Applications (Virtual Network Functions)
- Self-care portal

A. Access network domain

Most southbound part of the vCPE solution is the traditional physical CPE itself which is installed on the customer location. For this solution, where the idea is to identify each individual end-user device, CPE must be configured in the L2 bridge mode. There are two different options in this case when CPE is configured in L2 mode which can be used to provide vCPE solution to the customer:

- CPE supports L2 tunneling towards the SDN domain. As mentioned several L2 tunnels can be considered VXLAN, GRE, L2TP,
- CPE in L2 bridge mode using only native VLANs towards the BNG. Then the tunneling is done from the BNG towards the SDN domain.

In some scenario's, especially for enterprises, it might be still required to have L3 routing functions available in the CPE, for example when redundant network connectivity to the Internet is needed. This is also possible to achieve with some vCPE solutions.

B. Cloud domain

Big benefit of the vCPE solution is that all Service Functions can be hosted as virtualized instances in an SDN enabled datacenter. To conform to openness of used technologies, OpenStack as a Cloud IaaS is preferred, but other cloud solutions can be used too. The Data Center Gateways (DC-GW) provides the L3 routing and connectivity of the DC, while the Top of Rack (ToR) switch provides L2 connectivity between the servers. Physical NICs of the vSwitches on the compute servers are interconnected through these switches and they belong to the infrastructure (SDN domain). In a similar way the compute and storage resources allocated to Virtual Machines (VMs) running on compute nodes belongs to the Cloud. The Cloud infrastructure is controlled by an OpenStack controller.

Opposed to plain vanilla OpenStack deployments, Ericsson IaaS solution is called Cloud Execution Environment (CEE). In that way the vSwitches hosted on the compute blades are not controlled by the OpenStack Controller but by a Cloud SDN Controller (CSC). The CSC controls the vSwitches called Cloud SDN Switch (CSS), since the requirements on them slightly differ from the Open vSwitch (OVS) requirements in OpenStack in the following aspect: the CSS supports complex packet header matching and packet forwarding mechanisms to fulfill the requirements of Service Chaining in the vCPE solution without use of metadata. The CSS is controlled by the CSC with OpenFlow.

C. SDN domain

The vCPE SDN network includes the following logical building blocks:

- Service Chaining: by which the operator can create granular services by chaining multiple service functions. The service chaining can be aware of the enterprise/home user, destination, application, or network performance.
- Extended LAN (E-LAN): by which the operator can host L2 and L3 services in the cloud and they will behave as if they were sitting in the customer premises.
- VXLAN or similar tunneling protocol: a tunneling mechanism to support site identification, and per-enterprise network slicing. The traffic from a branch is mapped to an internal tunnel, so that the traffic from that branch is placed in a network slice, completely isolated from other tenants.
- Load balancing and failover functionality: allows the operator to have scalable service functions with high availability.

The vCPE user plane is going through these components in the following way. The user plane traffic is encapsulated in the vCPE solution provided CPE, into a tunnel that terminates in one vSwitch (Cloud SDN Switch) within the datacenter. In the vSwitch the tunnel header is removed and traffic is forwarded according to SDN rules (i.e. source and destination IP address in combination with TCP/UDP ports) to SF within the service chain that matches the end user device MAC address. Service chain is assigned to the MAC address through self-care portal. The last SF in the service chain is the CGNAT SF, and then traffic will be forwarded by the DC-GW to the internet peering gateway and final destination IP on the internet (example, google.com). The principle of SDN in vCPE solution is shown on the Figure III-2.

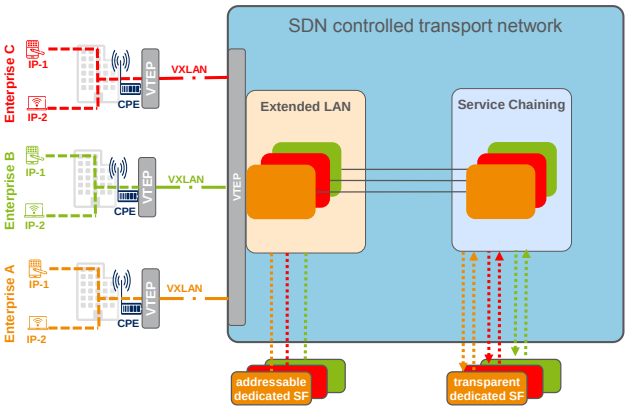


Figure III-2 Service Functions and Enterprise modeling

As mentioned in the previous chapter, two main components of the vCPE SDN part are vSwitch and SDN controller. In Ericsson vCPE solutions CSS is used as vSwitch, while SDN controller has an addition of Cloud SDN Controller (CSC) and an additional application Broadband Service Controller (BBSC).

The Cloud SDN Switch (CSS) is the network element that implements the data plane aspects of the SDN. It accepts configuration in the form of OpenFlow commands from the SDN Controller and implements the corresponding data path functionality. The CSS also collects data plane counters that can be queried by the SDN Controller and allows the SDN Controller to directly inject or receive data plane packets. The CSS is a virtualized OpenFlow switch that resides in the host hypervisor (typically KVM).

The SDN Controller is a network element in software-defined networks that centrally controls a network of Virtual Switches (the CSS) through a south-bound control plane interface, typically OpenFlow and OVSDB.

The Broadband Service Controller (BBSC) is responsible to configure an SDN network through an SDN Controller in such a way that certain traffic flows are forwarded along predefined Service Chains. With this kind of solution there is a possibility to enable the Operator to deliver per device service chains to their customers.

D. Management and orchestration

The vCPE solution orchestration is done at several levels depending on what needs to be orchestrated. The SDN domain is orchestrated via BBSC while typically some kind of Cloud Manager is the cloud orchestrator. For the vCPE solution there is a need to combine these two orchestrators. There is also a need for a business logic application in combination with the Portals. This is realized via flexible System Integration interface of Cloud Manager that is developed for this kind of tasks.

Service and Resource catalogue and component workflows are realized in the Cloud Manager. At this layer, the "subscriber services" as such are decomposed into "network service chain" services. The Cloud and Service Manager will provision new service function instances, scaling and instantiating as requested on demand. The Cloud Manager will communicate with the SDN layer to send notifications about the deployed service functions and the network service chains that have been provisioned. This layer will provide assurance functions, monitoring infrastructure and service functions (or interfacing with the service function manager) triggering scaling actions when required, following defined Operational Level Agreements (OLAs) and mapping infrastructure impacts to service function impacts.

Ericsson solution for Cloud Manager is called simply Ericsson Cloud Manager (ECM) and it is part of the complete vCPE solution.

E. Applications (Virtual Network Functions)

In general, SFs (Service Functions) process traversing or terminating traffic to add value to the operator, the end-user or both. Examples of service functions are:

- Spam Filters,
- Content Filtering systems,
- Security Gateways,
- Traffic monitoring and analysis systems,
- Network Attached Storage (NAS).

As seen in Figure III-2 each enterprise or home can have its own dedicated service functions (in other case it can be shared between more) and they can be addressable SF's, meaning they can be directly addressed by the end users or transparent for the end user. Addressable SFs belong to the same L2 broadcast domain as the end-user devices within an enterprise or home. Addressable SFs terminate traffic from and to the end-user devices and they are addressed by the end-user devices with their L2 (MAC) address. The vDHCP is an example of an addressable SF.

In contrast to addressable SFs, transparent SFs do not terminate traffic; they are not explicitly addressed by the traffic in any way. Transparent SFs process the (payload) data of the traffic but they do not modify the five tuple information in the packet header; since these fields are used to classify the packet and to identify the data flow each time they enter the SDN domain. A modification in this case would result in wrong forwarding decisions. Transparent SFs are subject to Service Chaining: Service Chaining is the forwarding of data flows through a set of service functions which are selected according to the data flows' characteristics. In the SDN controlled transport network, one Service Chaining instance is provided per enterprise or home. Each user in the enterprise or home can have its own Service Chain which is defined through Self-Care portal and BBSC. When a packet enters the Service Chaining instance, it is classified according to information in its packet header. Depending on the desired forwarding behavior this may be e.g. an IP address to identify a particular end-user device, a TCP or UDP port to identify a traffic protocol or again another IP address to identify the traffic destination. After the forwarding

decision has been taken, the packet is forwarded to the next SF. An example of SF aimed for service chaining would be a CF (Content Filter) and/or FW SF's.

The idea of the vCPE solution is that any Virtual Network Function can be part of this solution. Ericsson also supports this idea and in its vCPE solution is using both its own VNFs and the ones from other vendors. Since solution is not using metadata for Service Chaining it is open to integrate any service.

F. Self-care portal

The Self-care portal (part of the presentation layer) provide the main access to the enterprise/home administrators and users. These elements provide a graphical interface and execute the different flows for the service request and configuration. The Self-care portal is actually divided in two different but similar portals, Operator Portal and Enterprise Portal.

The Operator Portal is the main interface for the operator to orchestrate the vCPE once the initial setup and integration have been performed. It's a web based GUI with wizards to simplify the creation of enterprises and its associated vSFs and service chains. The Operator Portal uses REST interface towards Cloud Manager to transfer all the information about what should be executed towards the CEE. For the SDN service chaining the Portal uses the REST interface to BBSC.

Through the interface of Enterprise portal, the enterprise/home administrator will have access to where he/she will be defining the users, different profiles and association between both, different services (exposed as part of the catalog, specific for that enterprise/home or based on the generic portfolio) that will be available and the way to combine them for the enterprise/home users that will consume them. Additionally, it enables the enterprise/home user to monitor and act on the services available (as delegated from the administrator).

IV. USE CASES

As already mentioned in previous chapters, vCPE solution can be used both for enterprises and homes. In this chapter some use cases for both scenarios will be described.

Typically some small enterprises don't have their own IT sector or people who know have to handle small LANs. In that way they will have to use some 3rd party services to install their LAN network. With vCPE solution they can deploy only simple L2 bridge CPE and all the services and administration is done by their network operator. The other advantage is that they have full flexibility of using some network functions. For example the director of the company can have different network functions deployed in his Service Chain than the other employees. Or one department can have its Service Chain and the other department a different one. Also, if the enterprise wants to add new network functions to their portfolio it can easily be added with vCPE solution.

For the homes and residential users, vCPE solution can also be useful. For example, if the parents want to add Parental control (Content filtering) only for the devices which their children are using, this can also be done easily with vCPE solution. It is only "one click" in the Operator portal. There are many other possibilities where service chains and traffic can be managed differently depending on time of day, traffic behavior, network status and similar.

V. CONCLUSION

Automation, dynamism, flexibility and manageability are very important aspects of modern telecom's networks. In order to have competitive advantage, they must be able to assure innovation, time to market and efficiency in their service offering. Smart use of underlying technologies providing characteristics necessary to fulfill those requirements, is essential in today's market.

Virtual CPE's use of NFV, SDN and Cloud is a good example of this approach. Through automation of services, network and application it is possible to introduce new service to end-users quickly and efficiently, provide tailored service for each individual user and device, as well to reduce the cost of operating and maintaining the services.

By introducing ability to address individual users and their needs in a dynamic and automated manner, it is possible to broad the services and offering while still keeping the costs to a manageable amount. It is possible to cost effectively and quickly introduce different applications and services and to identify successful ones and continue their development, while eliminating unsuccessful ones.

We are still at an early stage of this approach and the full potential is yet to be revealed. But the current market interest and possibilities are very promising and show great potential for further development.

REFERENCES

[1] Ericsson, "Virtual CPE and Software Defined Networking" – White paper

[2] ETSI, "Network Functions Virtualization"

[3] Margaret Chiosi, Don Clarke, Peter Willis, Andy Reid and others, "Network Functions Virtualisation – Introductory White Paper", October 22-24, 2012 at the "SDN and OpenFlow World Congress", Darmstadt-Germany

[4] Peter Mell, Timothy Grance, "The NIST Definition of Cloud Computing", Special Publication 800-145

[5] Nick Feamster, Jennifer Rexford, Ellen Zegura, "The Road to SDN: An Intellectual History of Programmable Networks"

[6] ONF, "Software-Defined Networking: The New Norm for Networks" April 13, 2012

Soft Sensors in Wireless Networking as Enablers for SDN Based Management of Content Delivery

M. Tosic*, O. Ikovic * and D. Boskovic*
* Foundation VizLore Labs, Novi Sad, Serbia
milenko.tosic@vizlore.com, ognjen.ikovic@vizlore.com, dragan.boscovic@vizlore.com

Abstract - In this paper we present an approach for establishing soft/virtual sensors across wireless network components (access points, connected devices and routers) and a cloud based software defined networking (SDN) controller. These soft sensors provide the contextual insight necessary for efficient management of available storage and networking resources, all the while ensuring necessary QoS levels in content delivery services (i.e. video delivery). The approach presented converts wireless network infrastructure into an edge computing platform which provides real time decision making based on results received from established soft sensors. This provides the level of reactivity necessary for fast changing conditions in IoT era wireless networks. The cloud based SDN controller oversees the content delivery system and performs strategic/proactive decision making and guides edge computing processes and decisions in line with defined policies. In this paper we show the impact of SDN based resource management, supported by soft sensors and edge computing across wireless networking infrastructure, on resource utilization efficiency and QoS in content delivery.

I. INTRODUCTION

Content delivery constitutes the majority of all internet traffic, and has done so for quite some time. Video content delivery, both real time and on demand, is always seen as the biggest challenge for legacy and modern internet infrastructures and communication networks. Therefore, research efforts mainly focus on solving challenges related to video content delivery. Recent trends in communication network evolution indicate that the next major step (colloquially called 5G [1]) will require drastic changes in networking paradigms. The focus is shifting from improving performance of access technologies to providing new services (in addition to existing ones like user data transfer, voice and traditional content delivery) like machine-to-machine communication, energy efficient content delivery, communication with millisecond latency, augmented reality etc. In all these services, content delivery plays a major role. Networking systems should be able to perform context informed resource allocation, integrate content storing and delivery processes and make different access technologies complement each other (not replace one another). To this end, software defined networking (SDN), content caching, adaptive content delivery and context awareness are seen as enablers for true 5G evolution. By combining these techniques, the networking system can improve efficiency and quality of the most demanding and pervasive service - content delivery. Content delivery comprises delivery of entertainment content, video surveillance streams, application/framework/configuration updates, augmented reality overlays, batches of sensory/contextual data, deployment of containers (Docker, Unikernel) etc.

SDN is not a technology in the strict meaning of the word, but a mindset which brings the concept of abstracting complex systems (like communication networks and connected devices) into a form which can easily be used by sophisticated decision making algorithms (e.g. abstracting a communication network with a network graph composed of different weights assigned to nodes and edges). There are different enablers for SDN based management of communication networks and for the services provisioned through them. The most widely accepted enabler is the OpenFlow protocol [2] for managing forwarding tables in network switches. There are other more mission specific enablers and best practices which are used for specific scenarios and challenges. What is missing is one general enabler: the enabler which will ensure context awareness of decision making processes. Because we rely on connected devices and internet enabled applications and services in day to day decision making, it is evident that communication networks are deeply integrated with non internet systems and services. Therefore, proper management of resources and service provision processes requires pervasive context awareness not just related to the status of network links and nodes, but also to systems and services which exist in the same temporal and spatial context as internet access services.

In this paper we present the concept of soft sensors transforming network infrastructure into a distributed computing environment, capable of collecting and analyzing contextual data and acting on the results of their analyses. This soft sensor approach is the key enabler for proper resource and service provision management based on SDN. We showcase the soft sensor approach in two scenarios: 1. Context aware reactive decision making applied to network assisted rate adaptation for MPEG DASH video streaming technique; 2. Context aware proactive decision making applied to content caching based on request distribution and device profile.

Within the remainder of this paper we will first analyze the existing approaches for content delivery management; next we will provide more details about our soft sensor approach as applied to content delivery; and finally we will describe proof of concept experiments and analyze the results.

II. Related Work

Solutions for improving efficiency of content delivery services and corresponding QoS and QoE are well studied in previous and current research. We do not propose an alternative for these solutions, but provide an enabler for their proper implementation into an overall SDN system for content delivery management.

Content delivery research usually focuses on video delivery, since it is the most dominant and resource hungry service on the internet. Even in tough competition with proprietary solutions (such as Apple HLS, Microsoft Smooth Streaming, and Adobe HTTP Dynamic Streaming) MPEG DASH has managed to establish itself as the standard adaptive video streaming solution. Its popularity and widespread acceptance is the result of its ability to use the well established HTTP framework without the need to make any changes to it. This is why DASH received significant attention from the research community focused on improving efficiency of video delivery (offloading servers, backhaul links, use of multicast, provisioning of context aware load balancing), QoS and QoE.

Although lower (access technologies) and upper layers (applications and services) of network stack are constantly evolving, the intermediate part (networking and transportation) still relies on protocols put in place several decades ago. Changing networking and transport protocols internet wide is deemed impossible nowadays. However, they can be organized and utilized in a context aware and efficient manner which can boost their performance. This is where SDN shows its full potential.

The importance of SDN in forthcoming 5G evolution is analyzed in [1]. Also, a majority of research work conducted nowadays in the domain of content delivery tries to integrate content delivery protocols and methods with SDN control logic. [3], [4] and [5] provide analysis of the potential for using SDN logic to improve performance, efficiency, QoS and QoE of adaptive video content delivery. [6] analyzes content caching as a service by integrating SDN and CDN (Content Delivery Networks) logics.

All these solutions rely on the SDN approach to integrate networking and decision making so that the network is not just a transportation highway anymore, but an active part of service provision and adaptation. Context awareness is the main enabler that can make SDN the right tool for taking content delivery to the next level of efficiency and performance. In this paper, we describe how our soft sensor approach enables SDN controllers to turn networking elements into sensors for collecting and analyzing contextual data thus ensuring context informed reactive and proactive decision making.

III. Soft Sensors Approach Applied to Wireless Access Networks

Soft sensors can be deployed on networking elements and connected devices enabling collection, formatting and analysis of contextual data available from these devices/elements. Every networking element consists of a number of sensor primitives providing data on: status of interfaces (queue levels, SNR, Tx/Rx bit rates, packet drops etc.), system status (CPU/RAM usage, local storage usage, routing/forwarding table, power consumption) and status of its environment (connected devices, status of links, surrounding WiFi networks). By combining different sensor primitives, soft sensors are capable of deriving contextual insight without post processing performed on dedicated servers and SDN controllers. This enables reactive decision making on networking elements based on more insightful contextual input.

In general, our approach requires the deployment of a software agent on a networking device such as a WiFi AP, distribution switch or network router. The software agent bridges sensor primitive handlers of a network element's operating system (such as the OpenWRT) with data analysis model of the soft sensor and reactive decision making algorithm (see Fig. 1). It provides an abstraction of operating system controlled resources that data analysis processes can utilize without further formatting and preparation. This speeds up data analysis and enables different software sensors (different data analysis models) to be deployed on different networking equipment. The main components of a software agent are: 1) Software hooks for integration with procedures of the underlying operating system (like OpenWRT) which expose status and measurements on networking interfaces and other networking node resources (CPU, storage); 2) buffer for sensory and contextual data necessary for proper formatting; 3) data formatting procedure which exposes sensory and contextual data, collected from the underlying operating system, to data analytics procedures of 1st layer soft sensors; 4) action handler module of software agent transforms results of decision making into configuration commands for networking node; 5) query handler provides API elements which enable soft sensors to obtain formatted sensory and contextual data from software agent's buffer.

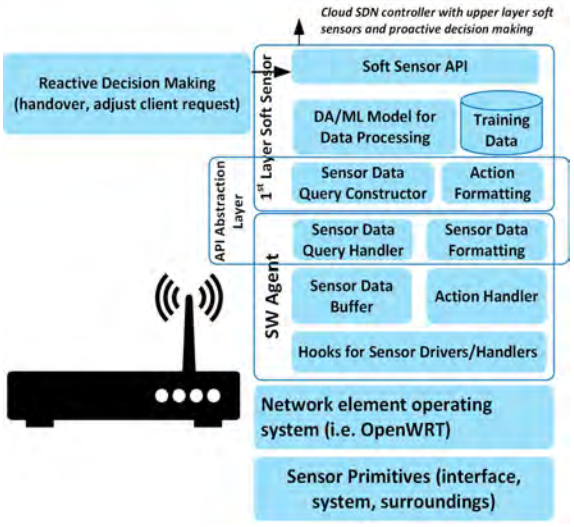


Figure 1. 1st layer soft sensor deployed on a networking element

In SDN logic, a software agent is implemented either as a southbound API (if 1st layer soft sensors are implemented on the controller or networking device) or

northbound API (if 1st layer soft sensors are deployed on cloud or SDN controller with higher hierarchy).

There are different layers of soft sensors. 1st layer soft sensors are usually located on networking elements and they are used for combined analysis of measurements and inputs from sensor primitives (part of network element's operating system). Upper layer soft sensors are deployed on the virtual network controller hosted on a cloud platform (we utilize Google Compute Engine for deploying our virtual network controller). These upper layer soft sensors combine results provided by lower layer sensors, sensor primitives and other sources of contextual data (e.g. web portals and open databases). They provide deeper analytical insight necessary for proactive resource and service provision planning. Fig. 2 shows how different sources of contextual data and sensory measurements can be combined to create soft sensors on 1st, 2nd and 3rd layer. Each layer provides deeper analytical insight into network status, trends, state transitions and correlations between network events. Our soft sensor approach enables dynamic creation of upper layer soft sensors by combining lower layer sensors and other sources of contextual data through well defined REST API (the API abstraction layer). This allows deployment of soft sensors and reactive/proactive decision making modules as they are needed (depending on services and application on a particular communication network). Therefore, the SDN control logic can dynamically adapt to the current network and service context (type of services being provisioned, profile of users and currently connected devices, situational context etc.).

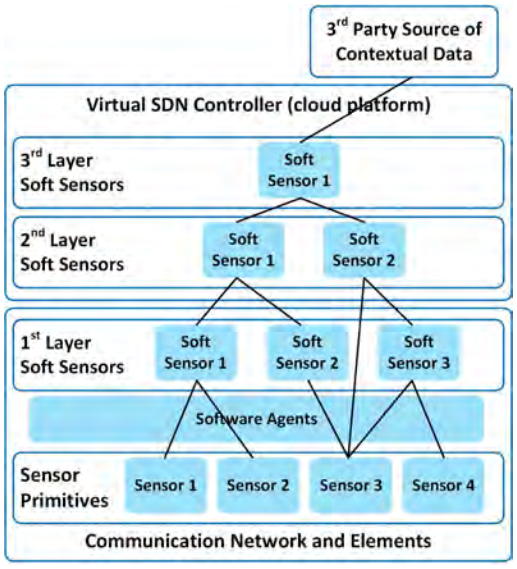


Figure 2. Soft sensor creation logic

We have developed, implemented and successfully validated the following first and upper layer soft sensors for reactive and proactive decision making in content delivery:

- 1st layer soft sensor for detecting connected devices and their classification (appliance, smart TV, gaming console, smartphone, tablet, desktop PC) and profiling (screen resolution, OS, buffer size, browser, battery capacity etc.).
- 1st layer soft sensor for classification of uplink and downlink capacity (detects congested status and estimates available capacity as a function of number of content delivery streams which can be established at the same time).
- 1st layer soft sensor for measuring and classifying cause of packet drop rate (due to low SNR, due to congested uplink/downlink, due to loss of connectivity).
- Upper layer soft sensor for deriving spatial and temporal distribution of content requests throughout the network.
- Upper layer soft sensor for user profiling with respect to requested content, used devices, viewing patterns and mobility.
- Upper layer soft sensor for identifying handover opportunities between access points and interfaces.
- Upper layer soft sensor for identifying opportunities for improved load balancing.
- Upper layer soft sensor for identifying suitable networking nodes for proactive content caching.

We employ these soft sensors for two proof of concept scenarios as described below.

IV. EVALUATION

The evaluation of the soft sensor approach for context aware decision making in content distribution is conducted through two proof of concept scenarios.

A. Proof of concept scenario 1- Context aware reactive decision making applied to network assisted rate adaptation for MPEG DASH video streaming technique

In this scenario we showcase how the 1st layer soft sensors can be used by the virtual SDN controller to provide reactive decision making necessary for a video streaming service. MPEG DASH technique is a standard for adaptive video streaming over the internet and TCP/IP networks. It relies on content with different quality representations (quality layers) which is divided into segments so that the player on the client side can request delivery of different quality levels for each segment based on estimated capacity of a link between client and server. Significant research is conducted with the main goal of improving the way in which MPEG DASH client estimates link quality and adapts streaming rate. The main conclusion behind that research is that the network needs to assist in estimating the quality of the link. Network elements usually can measure status of links more precisely and SDN controllers have much better understanding of overall network status. Our soft sensor

approach enables the SDN controller to reactively modify client requests for a video content segment and adapt that request for content representation (quality level) in line with the network status and context derived by the 1st layer soft sensors and the controller's overall knowledge of the network.

For the purpose of this proof of concept scenario we have deployed a MPEG DASH video streaming testbed as shown in Fig. 3. We have deployed the Wowza streaming server [7] (for real time and on demand video streaming) and used its MPEG DASH streaming engine to deliver scalable coded video content (Big Bunny test video). 1st layer soft sensor has been deployed on the WiFi AP powered by the OpenWRT to estimate uplink and downlink capacity. Software agent is developed in Python and deployed on the WiFi AP. The software agent exposes OpenWRT system resources (interface drivers) to the 1st layer soft sensor in format and through API calls required by the sensor. The soft sensor combines data from interface queues, SNR on WiFi interface, number of connected devices and packet drop rate (all provided by the software agent) to estimate link capacity between the client and the streaming server. Network conditions are varied between experiments by introducing capacity limitations and by increasing channel SNR.

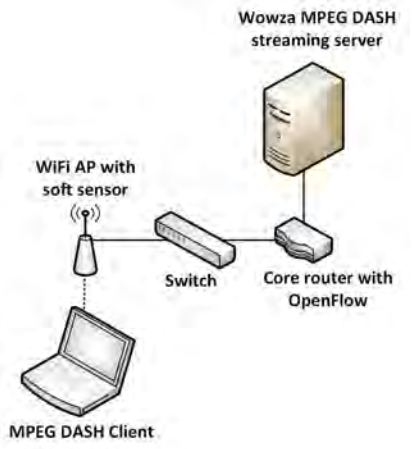


Figure 3. Scenario 1 testbed configuration

We have conducted experiments with 4 different streaming bitrates: 450, 750, 1100 and 1500Kbit/sec (quality representation 0 - base layer with lowest bit rate, layer 1 corresponding to 750kbit/sec, layer 2 corresponding to 1100kbit/sec and layer 3 corresponding to 1500kbit/sec streaming bit rate). We have used queues on the WiFi AP and core router to induce throughput limitations on the link between the client and the streaming server. Throughput limitation is varied between 2Mbit/sec and 10Mbit/sec in 1Mbit/sec increments. This emulated congestions on access and backhaul links on the path between the client and the server. WiFi channel width (20/40MHz) and Tx power are also varied in order to emulate changing interference levels in a wireless access part of the network. In one set of experiments we have let the DASH player on the client side use its built in adaptation algorithm to select bit rate for each requested segment. The other set of experiments are conducted with 1st layer soft sensor deployed on the WiFi AP in combination with a reactive decision making algorithm which intercepts and modifies the client's requests for a video segment in line with estimated link capacity (both from AP to client and between AP and server). Experimental results show that the built in adaptation algorithm of DASH JS player [8] tends to shift between base layer (representation 0 - 450kbit/sec - total of 29 segments with the lowest representation are requested) and the highest representation 3 (1500kbit/sec - also 29 segments with the highest representation are requested for the first three minutes of the test video content streaming) completely avoiding intermediate representations (see Fig. 4 for the case where we have introduced 4Mbit/sec capacity cap on the path between client and server). This caused significant fluctuations in perceived QoE since quality of representation was shifting often and not gradually. With the soft sensor approach we have seen that we can maintain the packet drop rate (percentage of dropped packets was around 6% for both test cases which didn't affect QoE) as in the case of client controlled streaming adaptation while minimizing fluctuations between different representations and maximizing the number of segments successfully delivered with the highest streaming rate (see Fig. 4). The soft sensor intercepted client's requests for representation quality change and reactively modified them in line with the derived content delivery link/path capacity. Although the client player requested the same segment representations as in the previous test case, only 3 packets are delivered with the lowest quality representation 0 (including the fast start buffer fill) and 50 (out of 60 in total for the first three minutes of the streamed content) of the highest quality representation 3. This shows that the network element (WiFi AP in this case) can measure and estimate link conditions and capacity much more precisely than the end client if it is properly equipped with the soft sensor which combines all available measurements necessary for proper estimation of link status and capacity.

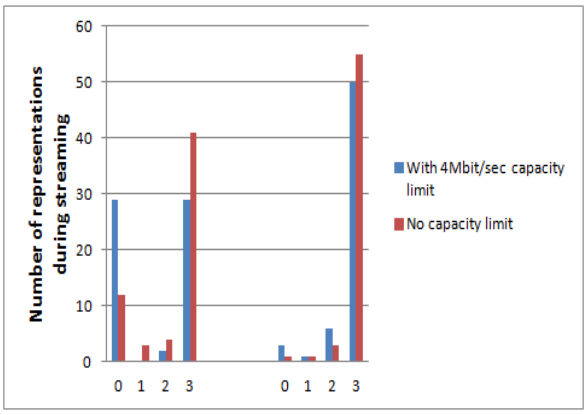
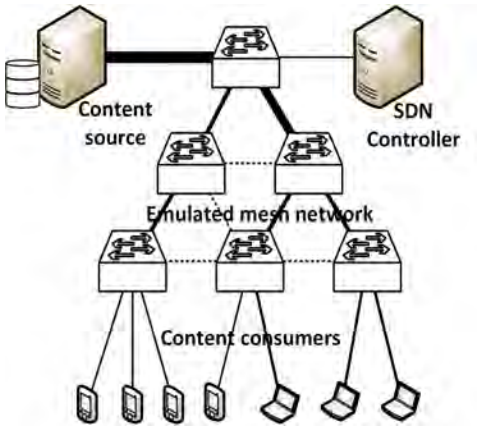


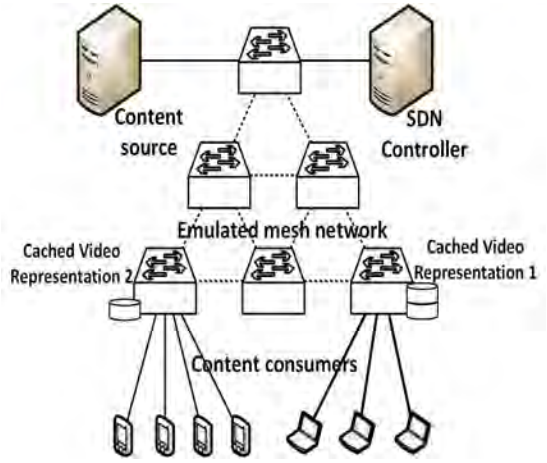
Figure 4. The number of different quality representations in two cases for the first three minutes of the test video streaming. Left set: where the client controls rate adaptation and Right set: where the soft sensor on the WiFi AP provides input for decision making on rate adaptation. Number of dropped packets in the first case = 1303, number of dropped packets in the second case = 1417. There were no buffer underflows in both cases.

B. *Proof of concept scenario 2 - Context aware proactive decision making applied to content caching based on request distribution and device profile*

In our previous work ([9], [10], [11]) we have developed, implemented and tested a MILP (mixed integer linear program) model for deriving optimal content placement on mesh routers in a wireless mesh network. Later on, we generalized the model so that it solves optimal wireless mesh gateway placement in line with client distribution with the goal of maximizing total streaming capacity of the mesh network. Now, we have implemented the model in Python and we use it as a proactive decision making algorithm for our SDN controller. For this proof of concept scenario we use the Mininet SDN emulator [12] and its Python API through which our SDN controller interfaces with the Mininet environment (see Fig. 5a).



a)



b)

Figure 5 - Emulated testbed for the scenario 2: a) Initial testbed setup where content is located only on the content server (dotted lines are auxiliary backhaul links which are not currently in use and line thickness represents link load); b) Status of the testbed after proactive content placement and context aware client handover between Mininet switches emulating mesh APs (we can see that none of the backhaul links between switches are utilized for content delivery)

The OpenFlow switches from the Mininet emulate the wireless mesh network. Links between switches are defined so as to represent backhaul wireless links of a typical wireless mesh network (802.11a protocol). Mininet Python API handlers are used as the first component of the software agent providing contextual data to soft sensors and transforming results of data analysis to the MILP model on the SDN controller for decision making. The MILP model utilizes data from upper layer soft sensors which provide spatial and temporal distribution of content requests and profiles of connected devices. This knowledge is used for deriving optimal content placement on wireless mesh nodes, optimization goal being maximizing total streaming capacity (serving maximal number of clients before one of the link in the network becomes congested) and minimizing average delivery tree length (number of wireless hops between client and content source which directly translates to streaming delay and jitter). The OpenFlow is used for modifying forwarding tables of the Mininet switches. Therefore, the OpenFlow southbound API is the second component of software agent which translates outputs of the decision making algorithm into rules for updating forwarding tables of Mininet switches. If a Mininet switch (it represents wireless mesh router in the MILP model) is selected for content placement, then forwarding tables are modified so that requests for the content server are redirected to that switch (in line with the optimization function which tries to minimize delivery tree length for requesting clients and maximize total streaming capacity).

Results obtained through this emulation correspond to the main findings in our previous work [11] where the MILP model was implemented using MatLab and Wolfram Mathematica software. These findings include:

1. Average delivery tree length (with respect to the observed content) is reduced by 1.2 wireless hops (backhaul wireless links in the wireless mesh network) when at least 30% of wireless mesh routers store requested content. This can depend on topology of the mesh network. Our analysis corresponds to mesh network topologies where every mesh router has at least two neighboring mesh routers. Since wireless hops introduce the biggest delay in content delivery and jitter, reducing the number of wireless hops on content delivery path can significantly improve QoS and QoE;
2. The number of users served with requested content from the in-network storage enters saturation after content is stored on at least 70% of mesh routers (can depend on network topology) so there is no need to store content replica on every mesh router to completely offload content server. This indicates that for a content which is very popular among network clients, placing the content on wireless mesh routers can significantly reduce burden on content servers and also reduce load on shared backhaul links (both wired and wireless) thus increasing overall streaming capacity of the network. The fact that replica of the content does

not need to be placed on every wireless mesh router to completely eliminate backhaul wired link load and content server load caused by content streaming, offers opportunity for optimizing utilization of limited storage on mesh nodes while maintaining acceptable level of load on content origin server.

Furthermore, based on data obtained from upper layer soft sensors, the SDN controller is able to select which SVC (scalable video coding) representation of video content needs to be stored on which wireless mesh router in order to maximize QoE and storage resource utilization efficiency (see Fig. 5b). The upper layer soft sensors provide the SDN controller with a profile of connected client devices (their screen resolution and processing capabilities). This enables the SDN controller to select proper video content representation (bitrate) to be stored on a particular mesh node. In coordination with soft sensors deployed on client devices, the SDN controller can even instruct context aware device handover within overlapping coverage region of wireless APs so that all low resolution client devices are shifted to one wireless mesh router while devices with larger screen resolution are handover to another network node. This will be the main focus of our future work.

V. CONCLUSION

In this paper we have presented our soft sensor approach which turns network elements (first of all WiFi APs) into sources of contextual data necessary for context informed decision making. We have presented multi layer soft sensors and their role in enabling reactive and proactive decision making when it comes to content delivery service. We are not providing an alternative for research results described in the related work section. We are providing a key enabler for SDN-based management of network systems and services. The two proof of concept scenarios showcase the potential of the soft sensor approach for reactive decision making as applied to adaptive video delivery and proactive decision making as applied to context aware content pre-placement on network nodes.

REFERENCES

[1] J. Hansen, D. E. Lucani, J. Krigslund, M. Medard, and F. H. P. Fitzek, "Network coded software defined networking: Enabling 5G Transmission and Storage Networks," Communications Magazine, IEEE, vol. 53, pp. 100-107, September 2015.

[2] OpenFlow (OpenFlow Switch Specification Version 1.5.0. [Online]. Available: https://www.opennetworking.org/).

[3] F. Bronzino, D. Stojadinovic, C. Westphal, and D. Raychaudhuri, "Exploiting network awareness to enhance DASH over wireless," arXiv: 1501.04328. January 2015.

[4] A. Advani, C. Joseph, D. Parwani, S. Bhol, and D. Reed, "Optimizing video delivery using OpenFlow," Capstone Final Paper, April 2015.

[5] P. Georgopoulos, Y. Elkhatib, M. Broadbent, M. Mu, and N. Race, "Towards network-wide QoE fairness using OpenFlow-assisted adaptive video streaming," Proceedings of the ACM SIGCOMM workshop on Future human-centric multimedia networking, pp 15-20, 2013.

[6] P. Georgopoulos, M. Broadbent, B. Plattner, and N. Race, "Cache as a Service: leveraging SDN to efficiently and transparently support video-on-demand on the last mile," 23rd International Conference on Computer Communication and Networks (ICCCN), IEEE, pp. 1 - 9, 2014.

[7] Wowza (Wowza - Media streaming systems. [Online]. Available: https://www.wowza.com/).

[8] Dash.js (Open source library for building DASH players. [Online]. Available: https://github.com/Dash-Industry-Forum/dash.js/wiki)

[9] D. Boscovic, F. Vakil, S. Dautovic, and M. Tosic, "Greening of video streaming to mobile devices by pervasive wireless CDN," Journal of green engineering, River publishers, Vol. 2, pp. 1-27, 2011.

[10] D. Boscovic, F. Vakil, S. Dautovic, and M. Tosic, "Pervasive wireless CDN for greening video streaming to mobile devices," Proceedings of the 34th International Convention MiPRO, IEEE Xplore, pp. 629-636, 2011.

[11] M. Tošić, M. Ćirilović, O. Iković, D. Kesler, S. Dautović, and D. Boscovic, "Impact of different content placement and delivery strategies on streaming capacity of wireless mesh networks," Proceedings of the 11th International Conference, ADHOC-NOW, Springer, pp. 302-315. 2012 Belgrade, Serbia.

[12] Mininet (Mininet: An Instant Virtual Network on your Laptop. [Online]. Available: http://mininet.org/).

A *FIRM* Approach for Software-Defined Service Composition

Pradeeban Kathiravelu
INESC-ID Lisboa
Instituto Superior Técnico,
Universidade de Lisboa
Lisbon, Portugal
pradeeban.kathiravelu@tecnico.ulisboa.pt

Tihana Galinac Grbac
Faculty of Engineering
University of Rijeka
Rijeka, Croatia
tihana.galinac@riteh.hr

Luís Veiga
INESC-ID Lisboa
Instituto Superior Técnico,
Universidade de Lisboa
Lisbon, Portugal
luis.veiga@inesc-id.pt

Abstract

Service composition is an aggregate of services often leveraged to automate the enterprise business processes. While Service Oriented Architecture (SOA) has been a forefront of service composition, services can be realized as efficient distributed and parallel constructs such as MapReduce, which are not typically exploited in service composition. With the advent of Software-Defined Networking (SDN), global view and control of the entire network is made available to the networking controller, which can further be leveraged in the application level. This paper presents *FIRM*, an approach for Software-Defined Service Composition by leveraging SDN and MapReduce. *FIRM* comprises Find, Invoke, Return, and Manage, as the core procedures in achieving a QoS-Aware Service Composition.

I. Introduction

Service composition [1] allows complex web services to be designed by composing simpler web services, and aggregating them to offer a complex execution of a business process or an enterprise requirement. As service composition involves execution of interdependent web services, multiple services can be composed such that they can be alternatives in offering the same service composition.

SDN separates the control plane that controls the network from the data plane that consists of the switches that actually forward the network traffic [2]. Software-Defined Cloud Networking (SDCN) enables effective configuration of cloud deployments, extending the SDN paradigm to cloud-scale, with multiple heterogeneous physical entities and logical components, such as data centers, storage, and middleboxes [3]. Context-aware service composition has been proposed by leveraging SDN for the deployment of service composition [4].

MapReduce [5] is a programming model that can be executed in a parallel and distributed manner. While web services engines such as Apache Axis2 [6] and Apache CXF [7] are traditionally used for creating and hosting web services, MapReduce frameworks such as Apache Hadoop [8] can provide an effective distributed and scalable alternative for web services in service compositions, by realizing and offering the service implementations as MapReduce applications.

FIRM is an approach for large-scale QoS-aware service composition, leveraging SDN and approaches and paradigms such as MapReduce and dynamic programming. It proposes four procedures, named Find, Invoke, Return, and Manage for an application-aware service composition. Find procedure finds the appropriate service installations as the core services in the composition. Invoke invokes the chosen service deployments in an efficient and distributed environment. Return returns the results of the service compositions back to the user. Manage manages and orchestrates the service composition through a web service registry as well as the SDN controller.

In the upcoming sections, we will further analyze the proposed FIRM approach. Section II will address background information and related work on SDN and service composition. Section III discusses the *FIRM* approach for Software-Defined Service Composition and the design of the solution architecture implementing *FIRM*. Preliminary evaluations on *FIRM* are discussed in Section IV. Finally, Section V concludes the paper with the current state of the research and future work.

II. Background and Related Work

As the scale and complexity of the networks and systems is growing larger and larger with time, programmability of clouds and networked systems is researched intensively [9]. SDN facilitates effective management of large networked systems of cloud scale, increasing the reusability of the architecture and configurations, by providing a logically centralized control plane separated from the data plane that forwards the data [10].

SDN controller manages the routing and forwarding rules, and updates the data plane which actually carries out the forwarding rules decided by the control plane [11]. Data plane consisting of multiple instances can be centrally managed [12]. OpenFlow [13] protocol is considered a driving force behind SDN. OpenFlow enabled switches can compose a software-defined network, along with one of the OpenFlow SDN controller implementations. Controllers execute, orchestrate, and manage the SDN algorithms and architectures in a physical network consisting of switches and hosts, or a network emulated by a network emulator such as Mininet [14].

Among many industrial and research SDN controllers, OpenDaylight [15] assumes a prominent place, as founded

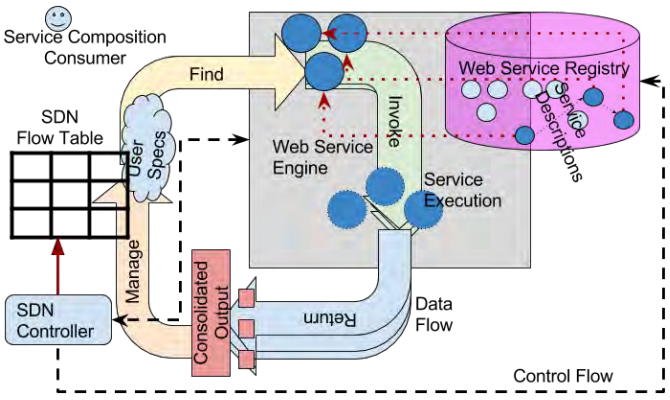


Fig. 1. *FIRM* Procedures

by large vendors under the collaboration of Linux Foundation as an open source project. In research, OpenDaylight has been proposed for global-scale very large deployments. CHIEF [16] leverages an orchestrated deployment of OpenDaylight for community network clouds. Further, OpenDaylight model-driven service abstraction layer (MD-SAL) can be leveraged to store contextual information, for decision making in application level [17].

The demand for offering more configurability to service composition has been on the rise. The Next Generation Service Overlay Network (NGSON) is a specification offering context-aware service compositions [18], leveraging SDN and Network Function Virtualization (NFV) [19] for the orchestration of services management and composition. The standardization effort of NGSON has been the motivation for many related works, focusing on efficient resource utilization and achieving pervasive services [20]. However, the existing implementations so far have not exploited the programmable networks offered by SDN effectively, to offer a Software-Defined Service Composition.

III. Solution Architecture

FIRM is an architecture and framework for Software-Defined Service Composition. The architecture is separated from the deployment, offering a loose coupling between the logic and implementation.

A. FIRM *Approach*

FIRM defines service compositions loosely, as a compound of multiple execution of services, which can be traditional web services or individual execution of parallel execution frameworks such as Hadoop or Spark [21]. Hence, throughput or efficiency of large batch jobs can be improved by decomposing them as smaller interdependent services or jobs, that can execute independently, and later be composed to provide the final result. *FIRM* effectively delegates the service registry and management functionalities to the SDN controller extensions, foreseeing a customizable and adaptive web service composition.

FIRM defines Find, Invoke, Return, and Manage as the core procedures, as shown by Figure 1. The web service engine hosts the web services. Web service registry functions as a registry that stores the descriptions of web service deployments including the web service end points. Web service engines often dynamically store information such as the requests on the fly, requests completed, requests failed, for each of the services. Web service registry is connected to the web service engine to dynamically update the registry based on the changes in the service deployment and health status.

Upon receiving the service composition requests from the client, Find procedure identifies the services to be invoked by analyzing the alternative service end points from the web service registry. Invoke procedure consists of invoking each of the services identified for the service composition. Invoke procedure is distributed and executed in parallel at service granularity. Return procedure composes and returns the final output of the service composition request. Find procedure interacts mostly with the service registry, where Invoke and Return procedures interact with the web service engine.

SDN controller, along with its *FIRM* service composition extensions, leverages the network topology and flow table information readily available to it, and also exploits the health information from web service engine. Manage procedure checks the health and congestion of the service deployments and reorganizes the services registry and controller extensions by prioritizing the service implementations dynamically. Hence *FIRM* offers a preferential list based on QoS parameters at network level as well as services or application level.

B. FIRM *Procedures*

The overall execution of *FIRM* is defined by Algorithm 1 that calls the 4 major *FIRM* procedures.

Algorithm 1 Execute *FIRM*

```
1:  procedure EXECUTE()
2:      Manage()
3:      repeat
4:          for (sc in serviceCompositions) do
5:              endPoints ← Find(sc)
6:              for (ep in endPoints) do
7:                  Invoke(ep)
8:              end for
9:              Return(sc)
10:         end for
11:     until (aborted)
12: end procedure
```

Manage procedure is invoked at the beginning of the execution at once, as shown by line 2, which is executed as an independent thread. Find, Invoke, and Return procedures are executed from the instance as a loop in the order, until *FIRM* is aborted.

Service composition requests are processed in parallel by the framework, for each of the service composition as indicated in the lines 4 - 11. Each service composition is given as an input, which is further decomposed by the Find procedure to get the series of actual service end points, as

in line 5. For each service end point that has been identified (line 6), Invoke procedure is invoked (line 7). Hence, Invoke procedure is invoked in parallel for multiple service end points. Finally, Return procedure consolidates the outputs of the service invocations and returns the final outcome of the service composition (line 9).

Manage procedure is invoked once to initialize the system during the start up of *FIRM*, and waits to be triggered for updates after initialization execution in the beginning. This enables a learning behavior, as reportedly congested service end points and longer execution paths are avoided or 'blacklisted' temporarily, as they are encountered and reported.

1) ***Manage****: Manage* procedure is executed in parallel, independent of the service composition workflows which majorly consist of Find, Invoke, and Return procedure calls. *Manage* consists of the initiating and maintenance flows interacting between the controller and service elements. Algorithm 2 depicts the Manage procedure.

Algorithm 2 Managing the *FIRM* Framework

```
1:  procedure MANAGE()
2:      initController()
3:      initRegistry()
4:      for (serviceEngine in serviceEngines) do
5:          init(serviceEngine)
6:          updateRegistry(serviceEngine)
7:      end for
8:      execPromoterThread(frequency)
9:      repeat
10:         if (triggerUpdate(sc, serviceProperties)) then
11:             for (service in serviceProperties) do
12:                 updateFlowTable(
                                service, serviceProperties)
13:             end for
14:         else
15:             sleep(frequency)
16:         end if
17:     until (aborted)
18: end procedure
```

At the beginning of the Manage procedure, the SDN controller is initialized with the network. *initController*() in line 2 initializes the controller and the extensions, with the flow table entries in the controller. *initRegistry* invoked in line 3 initializes the web services registry with the information of the web service engines to be configured with *FIRM*.

Each of the web service engine defined in the web service registry is initialized with the services deployed in them as shown by line 5. Initially the registry does not have any services information. As the web service engines are initialized, the registry is updated with the service information for each of the web service engines, by *updateRegistry*() in line 6.

Manage procedure executes from a master instance, which is configured with the controller as an extension. Client threads from the service instances update the web services registry by invoking the master, when the preference order of the end point changes for the next invocation of web service composition. Manage procedure waits for the updates to be triggered while idling otherwise.

Once the controller, registry, and the web service engines are initialized, *Manage* procedure waits to be triggered for updates. An update is triggered when the service engine notices a delay in completing the web service invocation or when there are an increasing number of web service requests on the fly for a certain web service engine deployment. A reference to the triggering or offending service composition (sc), as well as the exact service properties (serviceProperties) consisting of the service, sorted list of preferred installation end points, and description are received along with the triggerUpdate method, as shown by line 10.

For each service that is included in the serviceProperties as the offending services, the SDN flow table is updated to reorder the preferred node to be used among the identical service deployments. updateFlowTable() invoked in line 12 updates the flow tables in the switches accordingly to route to the specific deployment of the web service implementation among the identical implementations.

Periodically Promoting the Demoted Deployments: As updateFlowTable() usually operates in demoting the service hosts from the routing tables, the services that are removed from the list must be added back periodically as they may have recovered from the overload, congestion, or the adverse status observed previously. execPromoterThread() invoked in line 8 ensures that the service deployments that have been demoted are periodically promoted back to serve the service invocations. This thread functions as a shuffling operation to ensure that no service is underused.

Algorithm 3 presents the promoter thread which executes independently in a timely manner to promote the 'blacklisted' service deployments back to serve the web service requests. While this can be done by analyzing the health statistics of completed web service requests previously on the fly after 'blacklisting' the service deployment, in simplistic approach it is often just decided based on a random event such as flipping a coin at random intervals, as shown by line 5 by invoking flipACoin() which returns a boolean value in a fair random manner. A frequency value is given as an input to the method to indicate how frequently the flipACoin() should be invoked.

To avoid invoking the promote() too frequently, the promote flag is initialized to false (line 3), and the method sleeps for a given time for each iteration. At the specified frequency, the promote flag is reset by flipping a coin (line 5). If promote.flag value is set to true by the random event flipping coin, a random service deployment that was previously disabled is set to receive further web service requests in the future, till it is blacklisted again due to congestion or overload. The service to promoted is chosen by reading the list of black listed deployments at a random line. As each line represents a service, the service that is randomly chosen is promoted back to receive the

Algorithm 3 Promoter Thread

```
1: procedure ExecPromoterThread(frequency)
2:     repeat
3:         promote.flag ← false
4:         sleep(frequency)
5:         promote.flag ← flipACoin(frequency)
6:         if (promote.flag) then
7:             promote.serviceID ← readRandomLine(
                       listOfBlacklistedDeployments)
8:             promote(promote.serviceID)
9:         end if
10:    until (noSerivcesToPromote)
11: end procedure
```

web service invocations, as shown by line 8. As with the invocation frequency of promote(), choosing the service to be promoted can also be executed adhering to more intuitive algorithms than finding a random service.

2) ***Find****: Find* is the first procedure in the *FIRM* workflow, which finds the relevant service deployments to invoke for the service composition. *Find* is depicted by Algorithm 4.

Algorithm 4 Find Services

```
1: procedure Find(sc)
2:     sc.serviceMap < services, properties > ← parse(sc)
3:     for (service in serviceMap.getKeys()) do
4:         endPoints ← getAvailableImpls(service)
5:         ep.properties ← properties.get(service.getID())
6:         ep.service ← getEndPoint(ep.properties)
7:         sortedEndPoints.add(ep)
8:     end for
9:     Return (sortedEndPoints)
10: end procedure
```

Each service composition *sc* is parsed into a map of services and properties, as in line 2. For each of the services in the service properties map, the list of implementations is derived from the registry (line 4). Service properties are retrieved from the service, as shown by line 5, retrieving information crucial for finding the service deployment best-fit for the composition.

One of the service deployment end points, among the multiple potential service deployments is chosen, using the properties of the service as the parameter for the method getEndPoint() as in line 6. The end point is often a symbolic reference to the list of deployments of the same implementation. The exact deployment end point URI is chosen by the SDN extension, from the routing table as maintained by Manage procedure. Upon delegated to the network level, finding the exact host to invoke the service deployment is orthogonal to the Find procedure, as handled effectively by the controller.

The service end point as well as its properties are added to the sortedEndPoints variable, as shown in line 7. The service properties include service composition operators, which define how the service is related to the other services in the composition, whether the service execution can be distributed, should it block the subsequent service calls to the next service invocation, or can it be executed in parallel. Finally the list of sortedEndPoints is returned for the input of *sc*.

Minimizing Communication Overheads*:* While the service deployment can be chosen following a naïve approach such as choosing the first in the list of available service end points, *FIRM* enables finding the service deployments that are in close proximity to each other, for each invocation of service composition. This minimizes the overhead caused by the inter-rack communication across the service deployments of the service composition, exploiting the network topology readily available to the controller.

3) ***Invoke****: Invoke* is the second procedure in the *FIRM* workflow, which invokes one deployment for each of the service in the service composition. A service implementation is chosen by the Find procedure, and the controller is responsible for picking one of the exact service deployment end points for the execution. Manage procedure will be invoked if the status of a service deployment changes due to the service invocations. *Invoke* procedure is depicted by Algorithm 5.

Algorithm 5 Invoke Services

```
1: procedure Invoke(ep)
2:     for (ep′ in ep.properties.dependsOn()) do
3:         repeat
4:             sleep()
5:         until (ep'.out ≠ ∅)
6:         ep.params.add(ep′.out)
7:     end for
8:     ep.out ← call(ep)
9: end procedure
```

Dynamic programming has been leveraged to effectively reuse the previously computed service execution results in the latter service executions or service compositions that depend on the previous. The service invocations that the current service depends on, are included into the service properties, such that they can be retrieved for the execution of the current service. The service invocation waits till the previous service executions that are the dependencies for the current execution are completed.

Once the previous service execution results are available, they are added to the current service invocation parameters, as in line 6. Finally the current service call is initialized, and the result is stored as shown by line 8, for the future service executions that depend on the current.

4) ***Return****:* Return procedure consolidates and returns the final result of the service composition execution back to the user. The web service engine updates its status on completed and pending web service requests.

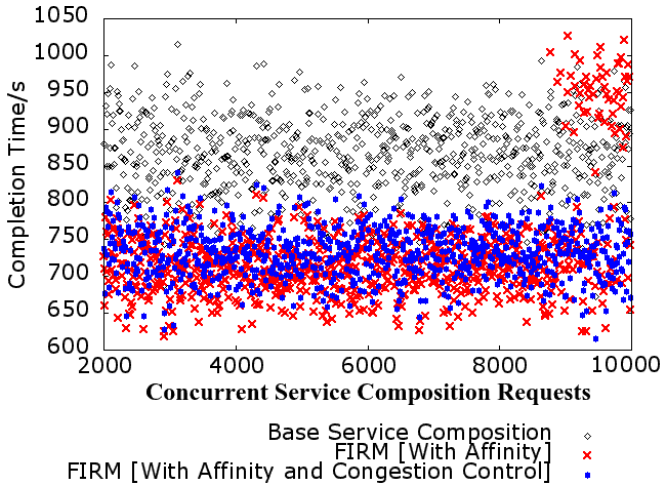


Fig. 2. *FIRM* Approach With and Without SDN

TABLE I. Enhancements to Service Composition

Feature	Existing Approaches	*FIRM*
Scalability	High	Higher
Bandwidth Overhead	Low	Low
Congestion Awareness	Low	High
Performance	High	Higher
Distributed Execution	High	Higher

C. Implementation

A prototype of *FIRM* was implemented to examine the feasibility of the proposed architecture with heterogeneous web service engines, Apache Hadoop 2.7.1, and OpenDaylight Beryllium SDN controller. Apache Axis2 1.6.3 and Apache CXF 3.1.3 were used as the web service engines in the prototype developments. Service Registry was custom developed in order to accommodate descriptions of services deployed in heterogeneous web service engines, including the use of MapReduce frameworks such as Hadoop in the service composition workflow, replacing a traditional web service engine.

FIRM prototype was implemented incrementally, first offering adaptive service composition with affinity while being network agnostic. With affinity, service requests are aware of the initial service instance that served them, and hence the subsequent requests are routed to the same instance, maintaining the state. This minimizes the frequent state transfers across multiple instances. Finally, leveraging SDN, congestion control was implemented. Hence *FIRM* offers congestion-aware software-defined service composition, while still offering the affinity guarantees.

IV. Evaluation

FIRM was evaluated for its performance and scalability in a combination of real and emulated service composition networks. Due to the limited accessibility to a 1024 host large scale multi-rack data center network, emulated networks were developed with *FIRM*, Mininet, and OpenDaylight, with web service frameworks and Hadoop. Performance of *FIRM* and Software-Defined Service Composition approach was benchmarked against the regular service compositions, for a complex service composition of weather prediction.

A. Performance and Scalability

Three major application scenarios were evaluated for the execution of multiple service composition requests: (i) Base service composition as offered by the web service engines and MapReduce frameworks; (ii) Service invocation with flow affinity improvements as proposed by *FIRM*, while not leveraging SDN and being completely agnostic to the network; (iii) With both affinity and congestion control offered by *FIRM*, leveraging SDN. As *FIRM* development and deployment followed an incremental approach, (ii) offered an intermediate state without deployment on SDN, while (iii) offered a complete implementation of *FIRM* and deployment on SDN. Approach (ii) finds the service deployments entirely at the service level using the service status information available to the web service engines.

Figure 2 depicts the time taken for all these three scenarios. Base approach took up to 1000 seconds to complete the service compositions. The solution scaled well, efficiently across the large cluster of 1024 nodes. However, there was a considerable data transfer across the nodes and racks even for a single service composition, which increased the bandwidth consumption.

In the approach (ii), the network was observed to be adaptively scaling based on the service composition requests. However, since the network statistics and load was not monitored or considered, when the load of the web service requests went beyond a certain value, a few of the services faced overload and congestion. As a result, though the system scaled for reasonably large service compositions, it started to take much more time when the requests reached 9,000.

In the final congestion-aware deployment of *FIRM*, the nodes that were already serving larger number of services were avoided dynamically. While this initially seemed to consume more time than the second approach using the statistics available for the web service engine, this approach scaled uniformly even for much larger systems.

B. Load Balancing and Congestion Awareness

Figure 3 depicts the deviation in completion time across multiple service composition executions, in order to find the imbalance in service distribution. The base approach handles the load balancing reasonably well, as Hadoop and web service engines are already optimized to distribute service requests across multiple instances. When attempting to distribute the load across multiple instances respecting flow affinity without leveraging SDN, there was a considerable overload in a few services.

By leveraging SDN to dynamically route the traffic to the instances that are under-loaded, *FIRM* offers congestion awareness and balances the load even for higher concurrency levels effectively. Table I compares the existing network agnostic approaches in service composition with SDN-based *FIRM* approach. While the related service composition frameworks offer a high scalability and performance, *FIRM* offers an even higher scalability and performance by leveraging SDN.

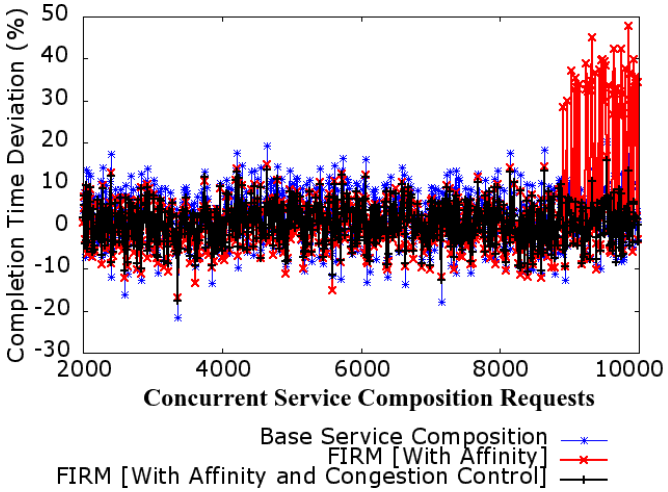


Fig. 3. Deviation (%) in the Completion Time

Service executions should be executed with less communication overheads by avoiding frequent transfer of data and state across multiple servers, to minimize the bandwidth consumption. Both a network-agnostic and *FIRM* approaches produce low bandwidth overhead by offering flow affinity. While not compromising flow affinity, *FIRM* offers a more distributed approach to service composition with congestion awareness, which is lacking in the network agnostic service composition approaches.

V. Conclusion and Future Work

FIRM proposes Software-Defined Service Composition by leveraging SDN for a QoS-aware service composition. Axis2 and CXF web service engines and Hadoop MapReduce framework are used for constructing the services for the compositions. Evaluation on the prototype implementation indicated that by utilizing the status information of the network available from SDN, service composition frameworks can be made scalable and efficient.

FIRM has been designed loosely coupled to specific technology, and avoids depending on web services as it can be implemented for any execution that can be made into a series of executions. As a future work, *FIRM* should be implemented for more distributed execution frameworks such as Dryad [22] and Apache Spark [21] and evaluated for more service composition use case scenarios in real-world physical deployment environments.

Acknowledgements: The work presented in this paper is supported by COST action 1304 Autonomous Control for a Reliable Internet of Services (ACROSS).

References

[1] J. Rao and X. Su, "A survey of automated web service composition methods," in *Semantic Web Services and Web Process Composition*. Springer, 2005, pp. 43–54.

[2] B. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, T. Turletti *et al.*, "A survey of software-defined networking: Past, present, and future of programmable networks," *Communications Surveys & Tutorials, IEEE*, vol. 16, no. 3, pp. 1617–1634, 2014.

[3] R. Jain and S. Paul, "Network virtualization and software defined networking for cloud computing: a survey," *Communications Magazine, IEEE*, vol. 51, no. 11, pp. 24–31, 2013.

[4] F. Paganelli, M. Ulema, and B. Martini, "Context-aware service composition and delivery in ngsons over sdn," *Communications Magazine, IEEE*, vol. 52, no. 8, pp. 97–105, 2014.

[5] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[6] S. Perera, C. Herath, J. Ekanayake, E. Chinthaka, A. Ranabahu, D. Jayasinghe, S. Weerawarana, and G. Daniels, "Axis2, middleware for next generation web services," in *Web Services, 2006. ICWS'06. International Conference on*. IEEE, 2006, pp. 833–840.

[7] N. Balani and R. Hathi, *Apache Cxf web service development: Develop and deploy SOAP and RESTful web services*. Packt Publishing Ltd, 2009.

[8] T. White, *Hadoop: The definitive guide*. O'Reilly Media, Inc., 2012.

[9] A. T. Campbell, H. G. De Meer, M. E. Kounavis, K. Miki, J. B. Vicente, and D. Villela, "A survey of programmable networks," *ACM SIGCOMM Computer Communication Review*, vol. 29, no. 2, pp. 7–23, 1999.

[10] N. McKeown, "Software-defined networking," *INFOCOM keynote talk*, vol. 17, no. 2, pp. 30–32, 2009.

[11] S. H. Yeganeh, A. Tootoonchian, and Y. Ganjali, "On scalability of software-defined networking," *Communications Magazine, IEEE*, vol. 51, no. 2, pp. 136–141, 2013.

[12] H. Kim and N. Feamster, "Improving network management with software defined networking," *Communications Magazine, IEEE*, vol. 51, no. 2, pp. 114–119, 2013.

[13] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "Openflow: enabling innovation in campus networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, 2008.

[14] B. Lantz, B. Heller, and N. McKeown, "A network in a laptop: rapid prototyping for software-defined networks," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*. ACM, 2010, p. 19.

[15] J. Medved, R. Varga, A. Tkacik, and K. Gray, "Opendaylight: Towards a model-driven sdn controller architecture," in *2014 IEEE 15th International Symposium on*. IEEE, 2014, pp. 1–6.

[16] P. Kathiravelu and L. Veiga, "Chief: Controller farm for software-defined community clouds," in *Cloud Engineering (IC2E), 2016 IEEE International Conference on*, April 2016, p. To Appear.

[17] P. Kathiravelu, L. Sharifi, and L. Veiga, "Cassowary: Middleware platform for context-aware smart buildings with software-defined sensor networks," in *Proceedings of the 2nd Workshop on Middleware for Context-Aware Applications in the IoT*. ACM, 2015, pp. 1–6.

[18] W. John, K. Pentikousis, G. Agapiou, E. Jacob, M. Kind, A. Manzalini, F. Risso, D. Staessens, R. Steinert, and C. Meirosu, "Research directions in network service chaining," in *Future Networks and Services (SDN4FNS), 2013 IEEE SDN for*. IEEE, 2013, pp. 1–7.

[19] J. Batalle, J. Ferrer Riera, E. Escalona, and J. A. Garcia-Espin, "On the implementation of nfv over an openflow infrastructure: Routing function virtualization," in *Future Networks and Services (SDN4FNS), 2013 IEEE SDN for*. IEEE, 2013, pp. 1–6.

[20] J. Liao, J. Wang, B. Wu, and W. Wu, "Toward a multiplane framework of ngson: A required guideline to achieve pervasive services and efficient resource utilization," *Communications Magazine, IEEE*, vol. 50, no. 1, pp. 90–97, 2012.

[21] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: cluster computing with working sets," in *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, vol. 10, 2010, p. 10.

[22] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, "Dryad: distributed data-parallel programs from sequential building blocks," in *ACM SIGOPS Operating Systems Review*, vol. 41, no. 3. ACM, 2007, pp. 59–72.

Test Environment & Application as a Service

Tomislav Žitnik, Mladen Galin, Goran Pauković, Ivan Dević, Renato Čižmar, Zoran Bosić
Ericsson Nikola Tesla d.d, Zagreb, Croatia
tomislav.zitnik@ericsson.com, mladen.galin@ericsson.com, goran.paukovic@ericsson.com, ivan.devic@ericsson.com, renato.cizmar@ericsson.com, zoran.bosic@ericsson.com

Abstract: In the highly competitive global information technology (IT) industry it is crucial to apply the latest scientific know-how and service lifecycle models straight to daily business. Test Environment and Application as a Service (TEAaaS) is an OpenStack based Cloud solution developed in Ericsson Nikola Tesla, a Croatian provider of modern information-communications products, solutions, software and services, to transform the way IT applications, tools and environments are delivered to users. Information Technology Infrastructure Library v3 (ITILv3) Service Lifecycle Model has been applied in all development phases.

The solution comprises applications and tools made available to users as a customized service portfolio through a Web Cloud Portal, ready to be deployed as an automated development and test environment model. Multiple controller nodes in a high availability cluster provide effective reliability on the OpenStack level, while the custom adapted FreeNX based Load Balancer improves the overall application capacity. The OpenStack Ceilometer component provides real-time monitoring of the used resources and their utilization, so TEAaaS delivers a transparent and flexible pay-per-use service model.

I. INTRODUCTION

Test Environment and Application as a Service (TEAaaS) is an OpenStack based Cloud solution developed in Ericsson Nikola Tesla's Information Technology and Test Environment (ITTE) unit to bring the provision of applications, tools and environments up to speed on the latest developments. TEAaaS supports a new paradigm in providing information technology and test environment products and services to the business. In traditional approach the IT environment was delivered over a physical machine or a terminal server dedicated to users (Figure 1). With development of virtualization technology IT was able to speed-up environment delivery, but delivery and maintenance of virtual machines (VMs) was not completely automated as updates and administration regularly take significant effort. By the introduction of TEAaaS all services and tools are automatically provided through a Cloud service, as opposed to the traditional approach. All services are created and deployed conforming to the users' demand.

Users can choose among customized applications, tools and environments presented on the Cloud web portal and readily available for development and test purposes. After a user has selected the desired application or environment, a request is processed and dispatched by the Load Balancer (LB) component. The solution provides a real-time monitoring of the resources used and their utilization, which is controlled by the OpenStack Ceilometer component. In this way a transparent and flexible pay-per-use service is delivered to the user.

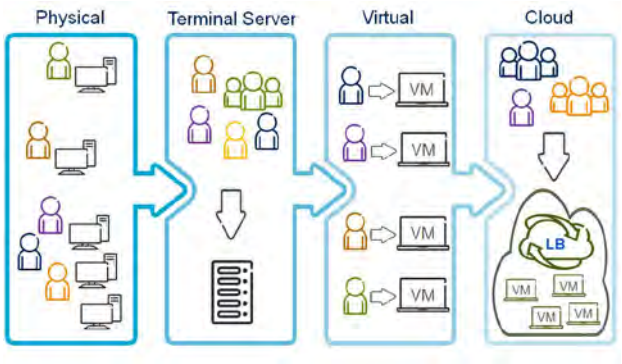


Figure 1. IT Environment delivery - technology evolution

Solution design requirements have been determined by improved time to market, full automation support, high portability and platform independence. Mirantis OpenStack has been chosen in alignment with the global Ericsson Cloud Strategy. The strongest advocate of this choice is a global Ericsson product, Cloud Execution Environment (CEE), which also utilizes Mirantis OpenStack as it has been thoroughly tested and made available by fully supported distribution through the latest Cloud technologies.

The cost analysis of TEAaaS solution showed a very low capital expenditure (CapEx) and operational expenditure (OpEx). TEAaaS can generate additional value based on adaptation of the existing open source. The Return on Investment (ROI) is positive due to high reusability of the existing hardware, the free-of-charge OpenStack and the FreeNX software, as well as the use of the standard open source development tools. Whenever any free software is considered to be included in an Ericsson product, an initial assessment of the license terms has to be conducted with the purpose to verify that the free software can be considered as Free Open Source Software (FOSS) [1]. The use of open source software in an Ericsson product is defined by the Ericsson Group directive "Free Open Source Software" [2].

II. CLOUD DESCRIPTION

The TEAaaS has been developed based on Information Technology Infrastructure Library v3 [7-11] service lifecycle model. ITIL Service Strategy deems it necessary to define a Service Catalogue. The Service Catalogue is a database or a structured document with information about live TEAaaS services, including those available for deployment. The Service Catalogue is an important Service Strategy tool because it represents the service provider's actual and present capabilities [7].

The TEAaaS Service Catalogue examples are Application and Environment on demand where a user can create his own fully configured, patched and secured "sandbox" VM with any application and/or test environment needed for daily business. For example, users can choose the Android development service which is already predefined in the Android environment. In this case TEAaaS will deliver a VM with the dedicated vCPU, vRAM, vHD, Eclipse and Android Studio.

Future development plans include Docker containers utilization to provide high level of independency and portability. Integration with Ericsson CEE and ESXi/VmWare will enable compliance with Ericsson global virtualization standards and compatibility with other virtualization platforms. Complete support for Bring Your Own Device (BYOD) model will enable users' mobility and hardware infrastructure independency.

Also, the TEAaaS Service Strategy considers it crucial to understand and anticipate the demanded service, as well as to counsel the customer on it. In discussion with business and service users an efficient capacity management needs to be established in order to prevent either over capacity, which could lead to extra cost without any benefit, or under capacity, which could lead to the loss of the business. In order to verify that TEAaaS solution meets business and users' requirements, the Proof of Concept (PoC) test was performed within the process of service validation and testing, as recommended by the ITIL Service Transition best practices [9]. During the PoC phase the use of the delivered VM environment and the use of applications were monitored in order to define daily business activities and how different types of users influenced the demand for service.

III. SYSTEM ARCHITECTURE

The TEAaaS system comprises the following key architectural components (Figure 2):

- Cloud Web Portal
- OpenStack with modified FreeNX Load balancer
- Production Test Environment

A. Cloud Web Portal

A Cloud Web Portal (CWP) is intended to serve as users' primary access point to TEAaaS environment. Users

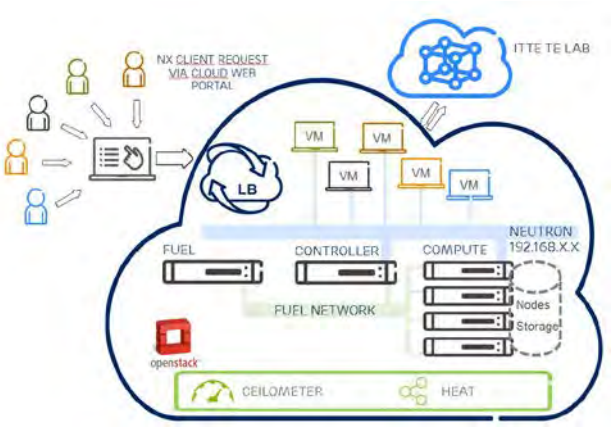


Figure 2. TEAaaS system architecture overview

have to order access to Cloud service, which needs to be approved by the Cloud service owner. When users access the Cloud Web Portal and are authenticated on the first level, they are presented with the list of TEAaaS applications and environments that they are entitled to according to predefined User Profiles. Authentication and a User Profile are defined by Lightweight Directory Access Protocol (LDAP) service to meet security requirements. TEAaaS Service Catalogue contains the list of all available tools, applications and environments which users can order. After users establish the connection to the CWP, they can access enabled services, currently used resources (for e.g. Central Processing Unit - CPU, memory) and information about previously created virtual machines which are still in use. The Cloud Web Portal provides users with the possibility to reconnect to the existing virtual machines. When users chose a new service, the required environment will be delivered through a newly created virtual machine.

The Cloud Web Portal runs on a dedicated web server and its code is written in PHP and Java script. The information related to customers, such as User Profile, sessions, enabled services, are collected in the background and stored in a MySQL database. The purpose of the MySql database is to speed up and improve CWP responsiveness. The information related to the consumption of resources is collected from the Ceilometer service.

B. OpenStack with modified FreeNX Load balancer

In TEAaaS solution, VMs are allocated per user and on demand. This is achieved by using Free NX server, which is configured as a Load Balancer running on a separate VM. For the purposes of this solution an open source FreeNX version is used. The Free NX source code modification enables communication with the OpenStack in order to request automated creation of new VMs or termination of the currently running VMs. Once users decide to terminate the session, the Load Balancer will do the cleanup of their VMs and thus release resources. This mechanism contributes to savings for commercial applications as floating licenses are delivered to VMs only during their actual use. A Load Balancer serves as an entry point to the test environment. It provides the second level of user authentication and keeps track of users' sessions and VMs dedicated to them. The existing two levels of user

authorization can be enhanced by the introduction of Single Sign-on (SSO).

TEAaaS utilizes the OpenStack Fuel node as a cloud deployment and management tool. Fuel provides an intuitive, Graphical User Interface (GUI) driven experience for the deployment and management of the OpenStack, related projects and plug-ins. [3], [4].

A Compute node is an Open Stack node which runs a Kernel-based Virtual Machine (KVM) hypervisor used for creating virtual machines and provides resources for users' workload.

The Software Defined Networking (SDN) functionality is enabled by the use of the standard Open Stack Neutron service implementation. Neutron with Virtual LAN (VLAN) segmentation has been used. There is a new tagged VLAN created for each network. Neutron was configured to use Modular Layer 2 (ML2) plugin with VLAN driver. Neutron L3 agent has been used to provide routing and floating IP functionalities. Where needed, it is possible to establish the Internet connection through a proxy server which is a part of standard Ericsson environment.

A Ceilometer is an Open Stack telemetry service. It collects data on the usage of different sort of resources, such a CPU cycles for virtual machines, the amount of storage space used, etc. The collected telemetry data are viewed as an input for the defined pay per usage charging model.

A Controller node is an Open Stack node on which processes acting as a Representational State Transfer Application Programming Interface (REST API) endpoint are running. It can be deployed in a multiple controller mode as a high availability cluster. The OpenStack Nova-scheduler service improves the capacity and overall reliability of applications by distributing workload across multiple compute nodes. A compute uses the nova-scheduler service from controller node to determine how to dispatch compute and volume requests [3].

C. *Production Test Environment*

A Production Test Environment is the setting that provides conditions for testers and developers to perform their daily tasks. In our setup it is devised as either a physical System Test Plant (STP), or a Simulated Test Environment (STE), or it is a combination of both. STP consists of physical telecommunication nodes such as MSC, HLR, IMS, MGW and other. STE is comprised of simulated telecommunication nodes, networks and any other simulated tools. TEAaaS provides internal resources (VMs) for partial or complete STE implementation and supports connection to STP. Application specific VMs in TEAaaS are a virtual set of nodes and instruments running in a lab, or a virtual set of VMs running specific applications or services like a web or a database server. The connection to a physical STP environment is established via the standard OpenStack Neutron's SDN implementation. The SDN functionality connects physical STP nodes and TEAaaS either through the OSI layer 2 or OSI layer 3 [13].

A Heat is an OpenStack orchestration service used for provisioning larger set of dependent VM resources. Configurations defined through Heat templates can consist of multiple VMs, and networks which enable users to perform the required activities. In the case of TEAaaS virtual test environment, a FreeNX Load Balancer is used to deploy complex configurations through predefined Heat templates. The entire multi-node application with all its services, dependent packages, operating system, software configuration and underlying infrastructure in terms of compute VMs, virtual networking, and storage connection is deployed on request through the simple start of a predefined template.

IV. PROOF OF CONCEPT

The TEAaaS Proof of Concept test has been successfully concluded after being conducted by a selected group of Ericsson users. The purpose of the PoC was to verify TEAaaS architecture as shown in Figure 2 and described in the System Architecture.

The Proof of Concept test was divided in two phases. The emphasis in the initial phase was put on the Free NX load balancer environment testing. This includes adding and removing users and User Profiles, network connection setup, the definition of ports and session types, the initial and automated creation and termination of VMs.

During the test, users experienced issues related to network (IPv6) configuration. Modern test environment for Ericsson products requires the IPv6 connectivity. Successful IPv6 connectivity was achieved by additional modification of configuration files and packages on the VM template.

The second phase covered the test of tools and applications running on Linux and Windows virtual environment. Some test tools and development applications included in tests were ClearCase, git, Eclipse and client applications for traffic generators. During this phase there were no major problems and the environment was stable.

In addition, PoC confirmed one of the main TEAaaS advantages over the traditional IT environment delivery approach: users and TEAaaS administrators can use a mobile smart phone or a tablet PC and achieve mobility in



Figure 3. TEAaaS over mobile connection

daily work (Figure 3). Users with Windows OS tablet PC needed to enable Wi-Fi connection and connect to ETK Intranet Network. Users installed free-of-charge NX client software on their tablets and they were able to request a VM from TEAaaS. After establishing a connection between TEAaaS and a user's tablet PC, a VM was delivered to the user and he/she managed to start the daily work on the dedicated VM. In this way users managed to work remotely from any location covered with ETK Wi-Fi intranet network, having the option to remain mobile during the entire work.

TEAaaS administrators could also enjoy the benefits of mobility as they were able to connect their mobile phone or tablet PC to the ETK Wi-Fi network.

After establishing a connection, an administrator has OpenStack Dashboard [3] available via the mobile or tablet's web browser. Hence, administrators can remotely control OpenStack functions available in the Dashboard.

V. CONCLUSION

The TEAaaS is currently under development for the production environment. The original TEAaaS portfolio will be expanded in this process since, based on a very positive feedback from PoC team, there is a demand for further TEAaaS development and new services. The integration with Ericsson CEE and ESXi/VmWare will enable compliance with Ericsson global virtualization standards and compatibility with other virtualization platforms.

There are several key aspects of TEAaaS solution. The environment for each user is created from the same sets of virtual machine images, thus ensuring identical environments for all uses. VM images can be customized according to the project needs with the required set of tools and applications. In order to ensure data availability in the case of unplanned VM termination, all user data are kept on a remote storage. In case there are severe problems, a simple termination and a restart of the VM ensures an easy and fresh start of users' environment. The configuration and security updates of an environment are introduced by installing updated VM images and by the updates of Heat templates. Users need to restart their environment in order to work on the updated version.

Some limitations, however, are imposed by the used technology: OpenStack lacks High Availability mode support for VMs. The use of VLAN segmentation in SDN implementation imposes a limit to max 4096 VLANs. FreeNX Load Balancer is a single point of failure, so there is a need for a Load Balancer Compute Cluster to make it easily available. The implementation of the Load Balancer Cluster is planned for a future TEAaaS development.

The commercial aspect of TEAaaS solution is built upon a very low Total Cost of Ownership (TCO), which has been achieved with the use of the free-of-charge OpenStack, the FreeNX software and standard open source development tools. The added value is brought by the free choice of access platforms and the support for the Bring Your Own Device (BYOD) concept. The future development aims at utilizing Docker containers to provide a higher level of independency and portability.

REFERENCE

[1] Free and Open-source software, Available at: https://en.wikipedia.org/wiki/Free_and_open-source_software#CITEREFFeller2005, [Accessed: 08-Feb-2016]

[2] G. Nilsson: Open Source Software Within Ericsson Group, Ericsson internal document, Stockholm:2013.

[3] OpenStack Foundation, "OpenStack Configuration Reference", 2016, Available at: http://docs.openstack.org/liberty/config-reference/config-reference-liberty.pdf, [Accessed: 08-Feb-2016]

[4] Mirantis, Inc., "OpenStack Deployment Fuel", Available at: https://www.mirantis.com/products/mirantis-openstack-software/openstack-deployment-fuel, [Accessed: 08-Feb-2016]

[5] Akemi Yagi, HowTos/FreeNX, 2015, Available at: https://wiki.centos.org/HowTos/FreeNX, [Accessed: 08-Feb-2016]

[6] Kevin Tatroe, Peter MacIntyre, Rasmus Lerdorf, Programming PHP, 3rd Edition, O'Reilly Media, 2013

[7] Office of Government Commerce (OGC), ITIL - Service Strategy, LAG, 2007, ISBN 9780113310456.

[8] Office of Government Commerce (OGC), ITIL - Service Design, TSO, 2007, ISBN 9780113310470.

[9] Office of Government Commerce (OGC), ITIL - Service Transition, LAG, 2007, 9780113310487.

[10] Office of Government Commerce (OGC), ITIL - Service Operation, LAG, 2007, 9780113310463.

[11] Office of Government Commerce (OGC), ITIL – Continual Service Improvement, LAG, 2007, 9780113310494.

[12] Ian Miell and Aidan Hobson Sayers; Docker in Practice, Manning Publications, 2015, ISBN 978-1-6172-9272-9.

[13] ITU-T X.200: OSI - Model and Notation, International Telecommunication Union, 07/1994, Available at: http://www.itu.int/rec/T-REC-X.200-199407-I, [Accessed: 08-Feb-2016]

Workaround solutions used during PSTN migration of customer to IMS network

Nenad Štokić
T-HT/ IMS Shared Service Center Hub Department (T6),
Zagreb, Croatia
nenad.stokic@t.ht.hr

Abstract - PSTN migration of customers in Croatian Telecom from legacy TDM network to the new IMS network brought many challenges. Not all of the new products on IMS platform were ready to be applied for customer migration, therefore alternative solutions as technical workarounds were needed. Main issue was how to switch customers, without many problems, on the new system with retention of the existing QoS until final solutions were available. We had to provide quick response on various challenges, because migration of customers to the new technology required different solutions. Paper describes the alternative solutions that have been applied to successfully resolve challenges in this large and complex project of PSTN migration.

I. INTRODUCTION

There are three 3GPP / ETSI TISPAN standardisation tracks describing approaches for migration of PSTN/ISDN telephony services using IP-based call/session control.

The softswitch (SSW) based PSTN/ISDN Emulation Subsystem (PES), is defined by ETSI TISPAN (ES 282 002) where the Access Gateway (AGW) (an access node in an operator controlled environment) or a Residential Gateway (RGW) (a CPE by the subscriber/in a non-secure environment) is controlled from the Access Gateway Control Function (AGCF) in a Telephony Softswitch by an H.248 signaling.

The 2nd ETSI TISPAN approach (descibed in TS 182 012 and TS 183 043) was IMS based PSTN Emulation (IMS PES). This approach implies that an IMS system can control an AGW via a stand-alone AGCF over H.248 protocol. The stand alone AGCF converts the SIP messages to and from the CSCF in the SIP server to the H.248 AGW profile in the access node (AGW/RGW). The IMS based PES standardisation was never finished due to lack of market interest.

Finally, the 3rd approach, the PSTN Simulation Services (PSS), are defined by 3GPP/ETSI TISPAN (TS 22 273, TS 181 002) as when the CSCF (Call Session Control Function) in an IMS backbone system, controls the SIP client in a native SIP phone or in Voice over IP Gateway. The VGW (Voice Gateway) which can be AGW or RGW plays the role of a UE (User Equipment) with regard to the P-CSCF (Proxy Call Session Control Function). The IMS PSS track is also called IMS Multimedia Telephony as compared to emulation it also supports the usage of other media than voice like video, chat or messaging. While SSW and IMS based PSTN/ISDN emulation approaches were intended for near-exact replica of the legacy PSTN/ISDN services, simulated services (IMS PSS) on the other hand, reproduce the functionality of traditional services (Call Forwarding, Call Waiting, etc), in a slightly different manner. From an end user perspective, some simulated services can be very like the traditional service while others are significantly different – and some services are not be supported at all.

T-HT selected IMS PSS Simulation as main aproach for migration of PSTN/ISDN network to IP considering it more future proof compared to emulation tracks. The main reasons: IMS PSS is designed for native SIP terminals over Broadband Access. It was promising easier introduction of new and enriched services as well as reusage for later migration of mobile telephony to Voice over LTE.

As intermediate step, SSW based PES approach was used for special types of users that could not be initially migrated to IMS due to different reasons. SSW as IMS MGCF node was also reused for several workarround solutions described in next chapters.

However, due to selected IMS PSS approach, T-HT PSTN migration process required different types of workaround solutions to reproduce supplementary and regulatory services that were not defined nor supported on IMS PSS platform and also to achieve smooth migration retaining the existing service quality and end-user experience.

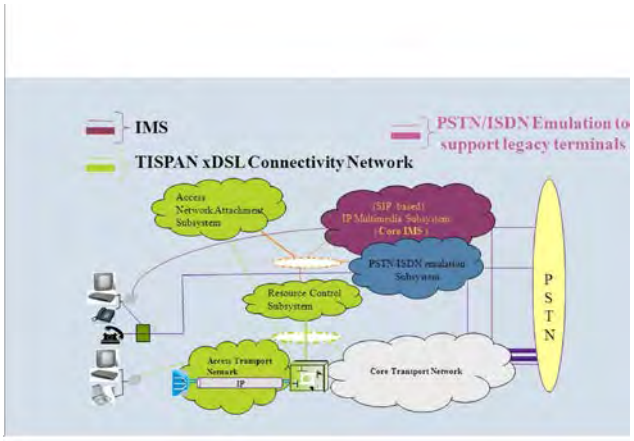


Figure 1. TISPAN NGN arhitecture

II. LEGACY PSTN NETWORK

Traditional telecommunications systems are characterized by an exceptional diversity in the philosophy of building networks. The major system components are based on the traditional circuit switching network, which is characterized by powerful and complex switching nodes, with complex application systems and a variety of features, created during the so many years of Development of public telecommunications networks (PSTN / ISDN - Public Switched Telephone Network / Integrated Services Digital Network) .

The T-HT legacy PSTN network consisted of:

- 68 Local Exchange
- 10 Transit Exchange/SSP
- 2 International Exchange
- 2 SCP nodes

The following subscriber types were available:

- POTS,
- ISDN – BA/PRA
- V5.1

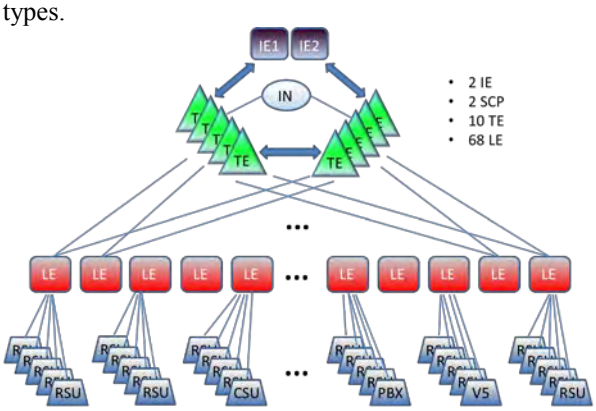
The T-HT legacy PSTN network vertical approach had lots of disadvantages in the organization of the telecommunications system:

- Each type of traffic requires its own network infrastructure, which entails higher capital investment.
- PSTN network has its own monitoring system and an associated organization, and that means higher operating costs.
- Complex (and often impossible) procedures of introducing new, unique services for many types of users.
- Inefficient utilization of a transmission system,
- Duplication of resources (servers, and nodes) in the application layer.
- Cumbersome, often non-dynamic organization, which is difficult to adapt to market demands.

Benefit of PSTN migration to horizontal all IP network for T-HT would have to make a reduction in operating costs, long-term reduction in capital costs, more efficient organization and a greater dynamism in the market (shorter process of implementation of new services, faster reaction to customer complaints and service offerings, etc.).

III. MIGRATION STRATEGY

Fixed network market is entering a transition from narrowband/voice centric towards broadband/multimedia centric future. The biggest driving factor for fixed network modernization is coming from broadband rollout and capabilities broadband provides. T-HT's vision is that the core architecture for this network modernization will be IMS, which provides the possibility of multiple services to a multitude of devices over multiple access types.



Fugure 2. T-HT legacy layered PSTN network

PSTN/ISDN telephony is steadily being substituted by VoIP telephony.

The options for securing current revenue streams and reducing Total Cost of Ownership include:

- Wireline narrowband modernization to replace ageing circuit-switched networks, using Telephony Softswitch Solution (SSW) and IMS for narrowband access.
- Rolling out wireline broadband VoIP to provide single-medium VoIP telephony service over broadband accesses using IMS solution.

SSW provides telecom grade characteristics, full legacy telephony and support of market specific functionality, i.e. it offers full compatibility with legacy PSTN/ISDN telephony services including all legacy supplementary services, residential and business i.e. fax, pulse metering, pay phones, regulatory services, Centrex, VPN, IN and existing market adaptations.

The SSW network includes a Telephony Server (TeS) which primary task is to manage call control and to provide legacy telephony services. The TeS uses telephony software which enables the solution to fully emulate all existing PSTN and ISDN services of most countries worldwide. The TeS also controls interworking with other networks using ISUP, BICC, SIP-I/T, SIP or H.323.

Further, SSW network has Media Gateways (MGW) and Access Gateways that transform TDM traffic into packet format and serve as ingress and egress points for the bearer connection in the packet backbone as well as termination of H.248 based signaling protocols. To ensure manageability and serviceability of the network OSS is included. SSW is also supported by a full set of Customer Services.

Telephony Softswitch is a commercially proven migration path from TDM to packet, providing telecom-grade telephony services over an IP-based access and backbone. It is the ideal, low-risk starting point for the cost effective evolution of your telephony network. This

pre-verified, end-to-end engineered solution provides high-quality voice capabilities over IP and includes a full set of PSTN/ISDN services.

SSW supports local network as well as transit and international applications. SSW also serves as a high-capacity voice gateway between TDM, IMS and other IP-networks. SSW can simultaneously support all above applications and it is possible to change the share of each application as the network evolves.

As the large scale network modernisation is mainly driven from broadband rollout and the capabilities broadband offers. Therefore the T-HT long term architecture is IMS, which supports multiple services, to multiple devices over multiple access technologies. Due to this IMS, it is promoted as the preferred offering for network modernisation on markets where broadband is the main driver.

Nevertheless as soon as the IMS domains start to grow the capacity need to follow the same trend. Capacity is needed for interworking, for handling what remains of PSTN, for connecting with other operator's networks and mobile networks. At the same time IP interconnected networks will need more and more international transit switching as well as to trigger legacy IN nodes.

These functionalities are excellent applications for a telephony softswitch. SSW is not only a powerful server, it may serve as a high-capacity gateway between IMS and other TDM and IP-networks, as an international gateway, as a transit node, as trigger towards legacy IN node. SSW can be seen both as a part of the IMS architecture and a provider of PSTN/ISDN-services to subscribers on legacy access nodes and access gateway nodes.

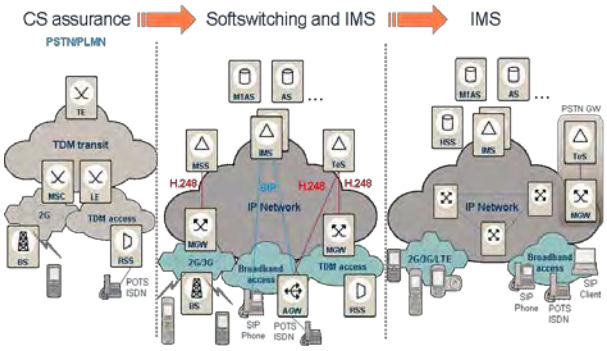


Figure 3. Migration strategy

IV. WORKAROUND SOLUTIONS

The most of workaround solutions designed during the T-HT PSTN/ISDN migration may be grouped to two main groups:

- Solutions for reproduction of existing PSTN regulatory or supplementary services that were not natively supported on IMS PSS platform as are not standardized by 3GPP / ETSI.
- Solutions related to access network or customer premises devices.

The chapters below describe typical or most important migration workarounds.

Anti-dialer

Service description: Anti-dialer is the legacy regulatory service which prohibits direct dialing toward expensive numbers announced by regulatory agency. Purpose of this service is to protect PSTN/ISDN users from fraudulent programs (dialers) which use dial-up modems on personal computers to initiate calls towards Premium Rate, satellite phones and other expensive numbers (usually located abroad). Although most of telephony users now use fast-internet service over BBA (broad Band Access), the service had to be reproduced on IMS platform as it is still mandatory regulatory service.

Workaround solution: As expensive numbers announced by regulatory agency are located abroad out of T-HT IMS domain, any call towards those numbers over telephony line of an IMS user (whether initiated by malicious dialer programs or normally by user) is normally as IMS break-out call routed by IMS/BGCF to SSW having MGCF function. SSW/MGCF nodes are provisioned with "black lists" of suspicious numbers or number ranges announced by regulator. Whenever dialed B-number ranges matches the "black-list" SSW/MGCF routes the call to network IVR that inform the caller via short voice prompt that the call can only be completed with additional authorization by the user in the form of a random 4-digit PIN. If the call is initiated by malicious dialer program, it cannot understand the voice announcement from IVR and cannot dial the requested PIN. Consequently such fraudulent call will be rejected by network.

If the caller is real user he/she will hear the random PIN, re-enter the same digit sequence using DTMF and the call will be authorized and further routed by SSW/MGCF via international exchange towards dialed destination. In case of successful authorization of the call, the IVR application will store the calling party number and will not request PIN for calls initiated from the same line for next 10 minutes in the case of voice call or for next 20 minutes if it was authorized data/fax call.

Billing Amount Limit

Service description: The Bill Amount Limit is regulatory service that informs the telephony user that his/her monthly billing amount limit has been reached. By interaction with the service IVR, user can choose either to proceed with the call or to block outgoing calls, except emergency calls and calls to T-HT customer care service. In case the user chooses to proceed with call, he/she must enter the personal PIN and after the PIN is validated by the IVR, the user is prompted to repeat the call attempt. In case the user chooses to proceed with the calls five successive times, the billing amount limit for

current month will be discarded. At the beginning of each new month, monthly limit is restarted.

Work-around solution: BAL is regulatory service that is not standardized in IMS PSS. The full-fledge implementation would require integration of IMS Telephony Application Server (TAS) with some prepaid or online charging system that would in real time "measure" user's current accumulated spending for current month and inform the user upon the monthly limit reach. However, such solution would be too expensive taking into consideration the cost of prepaid systems for small number of users that actually use BAL service. Instead, the workaround solution was implemented based on integration of T-HT billing system (ordinary offline charging) and "Virtual Carrier Preselection" service on IMS TAS.

T-HT billing system continuously receives charging records generated by IMS TAS for all calls generated by all IMS telephony users. Billing system usually process all collected charging records at the end of month to produce the monthly bill for fixed telephony user. However, for BAL users, billing system is provisioned to constantly calculate current spending in "near real time manner". Upon reaching monthly spending limit for particular BAL user, billing system initiates provisioning of temporal "Virtual Carrier Preselection" service for that user on IMS TAS. The logic of "Virtual Carrier Preselection" service is that it inserts virtual Carrier Identification as parameter Request- Uri of any initial SIP INVITE message generated by further BAL user calls. According to IETF RFC 4694, Carrier Identification Code (CIC) can be used to carry the pre-subscribed information identifying the current service provider where call should be routed. In BAL workaround solution, upon reaching monthly spending limit, BAL users' calls are temporary provisioned with virtual CIC that is not identifying any real carrier (operator). Instead, this virtual CIC is used to mark all calls of BAL users that reached spending limit and to route them to service IVR that inform the user about reaching the spending limit and offer the possibility to continue with the call or to block the call. Based on virtual CIC parameter in Request-URI of SIP INVITE message, the call is routed by IMS core (CSCF/BGCF) to SSW/MGCF. Additionally, CSCF/BGCF node copies the CIC parameter in front of dialed B-number. Such CIC prefix will be then used by SSW/MGCF to route the call to service IVR and also by by IVR to offer BAL voice prompt to A-user. Only if user dials the personal PIN number, the call is allowed to proceed. The temporary virtual "Virtual Carrier Preselection" service is automatically de-provisioned to all BAL users at the beginning of each month enabling them again to establish their calls without routing them to service IVR.

Public Payphones

Service description: As dominant fixed operator, T-HT must maintain Public Payphones enabling users to place calls by using prepaid cards for call charging. In T-HT PSTN network, Public Payphones were connected to analog POTS line (Z-interface) of remote subscriber stage (RSS) controlled by Local Exchange (LE). Based on dialed B-number, LE calculates the call tariff that is signaled to Payphone device by using 16 kHz pulses over POTS line. Depending on tariff, LE generates particular number of initial pulses at the call establishment and additional periodical pulses during the call. Payphone device converts pulses to charge units to be deducted from users' prepaid cards.

Work-around solution : In order to maintain existing Public Payphones when all PSTN lines are migrated to IMS and PSTN LE nodes are switched-off, T-HT designed the workaround solution based on MMTel Advice of Charge (AoC) supplementary service available on IMS TAS as standardized ISDN simulation service (3GPP TS 32.280; 3GPP TS 24.647).

Important part of solution is Online Charging Server (OCS) module that was provisioned with the same tariff setting data that was previously been provisioned in legacy PSTN local exchanges. If AoC service is provisioned to particular IMS user, during user's originating call setup, MTAS will query OCS with Credit Control Request over Ro interface and process Credit Control Answer into AoC information included in the SIP response. The tariff information received from the OCS specifies the current tariff and may optionally include the next tariff together with an indication of when that tariff becomes applicable (i.e. cheaper tariff during night, weekend days or on holidays). At call establishment, when transferring tariff information to the IMS terminal (to AGW in T-HT case), MTAS uses the tariff applicable at that particular time. AoC data is sent to the User Agent in the body of SIP messages in the form of an AoC XML document. AoC XML instance documents are transported as a SIP MIME body. Although MMTel AoC service is primarily intended to native SIP terminals (i.e. IP Phones), in T-HT payphones solution it was reused to generate and transport tariff data via SIP signaling towards MSAN/AGW access nodes where SIP signaling is terminated and converted to POTS signaling. MSAN/AGW POTS card extracts the tariff from AoC xml body found in SIP response and converts it to particular number of initial 16kHz pulses at call establishment as well as to subsequent periodical number of pulses during the call. Payphone devices count number of received pulses and convert them to charge units to be deducted from user's prepaid card. This way, all existing Public Payphones in Croatia were reused even all PSTN Local Exchanges were switched-off and POTS lines were migrated to IMS.

Nomadism Restriction (related to emergency calls)

Service description: Most of PSTN users migrated to IMS are BBA users using IAD devices (combined xDSL modem and RGW). BBA users may try to disconnect their BBA terminals (i.e. xDSL modems) from their BBA

line to some other BBA line that can be even located in another city or county. In that case their internet connection might work correctly as they will be using their own credential to connect to their internet service provider network (i.e. T-HT). However, in case BBA terminal is also IMS terminal (IAD or RGW) it will automatically try to register to IMS network. This situation is called nomadism. Similarly, user may try to register to IMS via public internet. Although this possibility might be beneficial for T-HT as it would enable users to generate fixed calls even when they are not at home location, there is issue with regulatory requirement related to routing of emergency calls. If for example IMS user from Split (Area Code = 021) connects its IAD to xDSL line in Rijeka, and if its' registration to IMS is allowed, this user may initiate emergency call by dialing non-geographic number of some emergency service (i.e. 193 in case of fire). In case user's physical location is not known, emergency calls of fixed telephony users are routed to nearest emergency service based on caller A-number (Area Code particularly). In above case, user located in Rijeka initiates emergency call that might be wrongly routed to fire-department located in Split which is against regulatory requirements. Accordingly, T-HT had to implement Nomadism Restriction solution that would prevent registrations if IMS terminal is connected to some BBA line outside user's county.

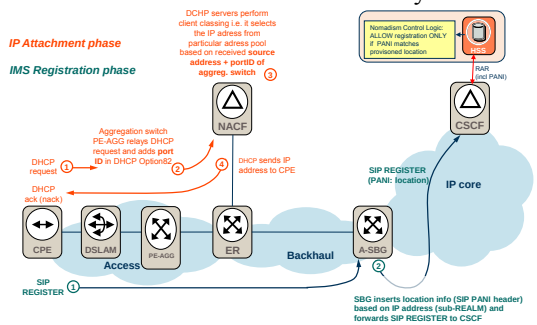


Figure 4. Nomadism restriction

Work-around solution: T-HT users are associated with number of IP ranges (configured in DHCP server) which correspond to ports Ethernet Aggregation switches (PE-AGG) that in turn are associated with county emergency centers. During the attachment phase (assignment of IP address to IMS CPE), based on PE-AGG port id value provided by PE-AGG using DHCP Option82, DHCP server will assign IP address from sub-range belonging to that particular port. Based on this mechanism, VoIP users belonging to one county will always be assigned IP addresses from sub-range belonging to that county. For each IP address range (county) a separated Realm was defined on the A-SBC access interface. By using Realm information SBC inserts PANI SIP header including user's current access location (county). Current access location from SIP PANI header is forwarded by CSCF to HSS by mapping SIP Register message to Diameter RAR message. HSS then checks received current location against provisioned reference access location and reject the registration case of mismatch.

Migration of residential ISDN-BA lines

Service description: Prior the massive deployment of BBA technologies (xDSL, GPON), ISDN-BA was attractive service for residential and SOHO subscribers requiring faster internet access or possibility to have two simultaneous calls over the same physical subscriber line and use multiple subscriber numbers. ISDN specific services (i.e. User-to-User Signaling, Sub-addressing, Closed user Group, Telephony 7 kHz) were not massively deployed world-wide and it was one of reasons on why they were not included in set of MMTel services (TS 22.273) by TISPAN and 3GPP. Accordingly, though there was significant number of residential ISDB-BA users in its PSTN network, T-HT decided not to reproduce all ISDN services. Instead, T-HT decided to implement pragmatic workaround solution using the standard set of MMTel supplementary services and same IMS RGW terminals (IADs) as for POTS users while still reproducing most used ISDN-BA features (two simultaneous calls over single physical pair, MSN, CW, CLIP, CLIR, …).

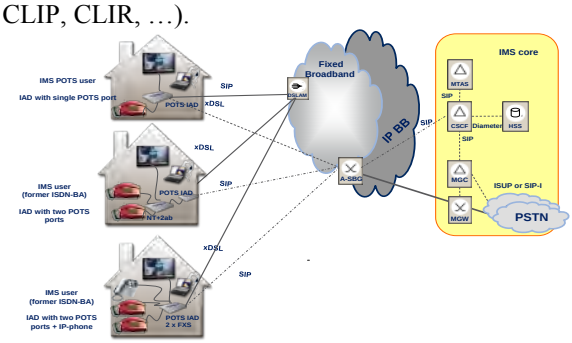


Figure 5. Residential ISDN-BA lines

Workaround solution: The migration solution for residential ISDN-BA user to IMS is indicated in the figure above where it can be seen that former ISDN-BA lines (two B-channels; two MSN numbers) were simulated by IMS IADs (RGWs) having activated two POTS ports (two numbers registered in IMS). For users with 3 MSN numbers, additional ethernet port on RGW was used to connect 3rd telephone set (IP-phone). Such solutions require bandwidth reservation of up to 400 kbit/s for voce service on BBA line (3 simultaneous calls x 120 kb/s per call for G.711 codec), however, with xDSL2+, VDSL and FTTx technologies it was not an issue.

Migration of business ISDN lines

Service description: As business users that have ISDN-PBX connected over ISDN line(s) usually wants to keep the ISDN PBX (and possibly ISDN terminals behind PBX), their migration to IMS required introduction of digital VGWs (IADs with ISDN ports) as indicated in figure 6.

In this case digital VGW is implicitely registering in IMS all users behind ISDN-PBX so there is no need for customers to upgrade their TDM PBXs to IP-PBXs.

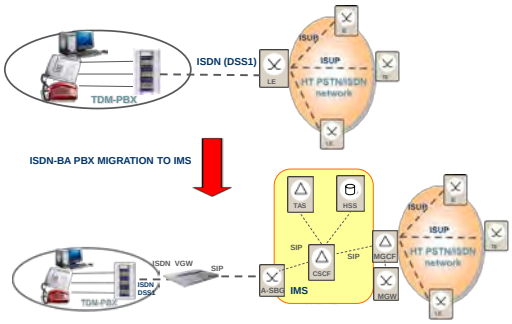


Figure 6. Business ISDN lines

Workaround solution: Since some companies (i.e. shops) were using ISDN-BA line also for data transfer (using Bearer Capability: 64 kbit/s unrestricted digital information), digital VGW has to be designed to convert this ISDN Bearer Capability to CLEARMODE pseudo-codec to indicate 64 kbit/s unrestricted sessions over SIP. In case data call is initiated by ISDN-PBX registered in IMS towards user in PSTN network, this CLEARMODE codec signals to media gateways that linear voice processing mechanisms (Echo Cancellation, Silence Suppression, Comfort Noise Generation) must be suppressed for these data calls. Without support of this mechanism, data would be corrupted during media conversion (IP->TDM).

The CLEARMODE codec is considered as intermediate solution until data traffic is moved to IP only transport (it is slow process as requires changes on customer devices and/or customer data applications).

Migration of analog PBXs (APBX)

Service description: Number of SoHo users still uses small APBXs that were connected to PSTN Local Exchanges over 2 or more analog POTS lines. APBXs usually serve 4-10 end-users (direct-dial-in numbers). In most cases PSTN LE exchange was providing the some kind of line hunting for APBX, i.e. in case called party is main PBX number, LE performs serial or paralel ringing of all PBX lines until the call is answered or is timed-out.

Workaround solution: As such functionality is not standardized in IMS, it was realized by reconnecting APBXs from POTS cards in PSTN local exchanges to POTS cards of IMS controlled MSANs (AGWs). In T-HT solution, MSAN registers all users behind PBX to IMS. The main PBX number is provisioned in IMS TAS with Line Hunting (serial or paralel ringing) depending on customer requirements. Incoming calls towards main PBX number are then distribuded by IMS TAS in serial or paralel way until it is not answered. In case all APBX lines are busy, not answering or not reachable /indicating some issue, IMS/TAS may be configured to alert also some backup number that is not connected onto company APBX . This setup is especialy appliable for small call centers or emergency serivces that are migrated to IMS but still want to use APBX devices.

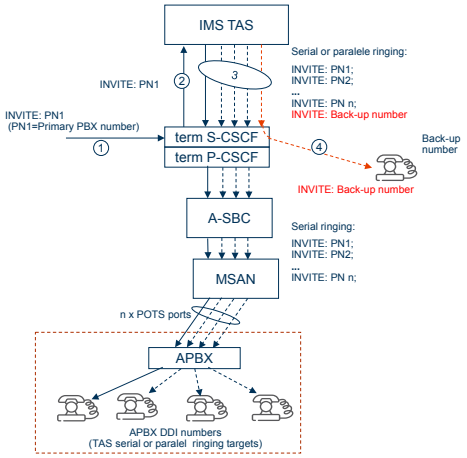


Figure 7. Analog PBX

V. CONCLUSION

Since the value of each network is proportional to the square of the number of users (Metcalfe's Law, 1946), it is essential to enter the transformation of the network in the right way at the right time in order to achieve the set of activities that includes the modernization and harmonization of all processes with the smooth running of services and optimum costs. To achieve the conditions with which we can come to the realization, it is essential that all aspects are clearly defined, as well as the way of resolving them in such a major project. The project will inevitably have a clear standard solutions, even when it contain temporary solutions while maintaining quality to the network transformation phase from a vertical to a horizontal structure where all users were clearly structured in the technological-technical terms.

It is not always easy to completely transform the network in a way, where there are no quick and easy solutions, but certainly it must be flush-mounted in the process of modernization of the network, which also represents the modernization of the user. One of the methods to achieve a continuity of networks transformations are workaround solutions, which under the condition of a comprehensive preparation can solve the most complex technical problems that are an integral part of all complex technical interventions in the telecommunications infrastructure. Using workaround solutions, customizations on IMS are avoided and thus additional costs which typically exceed the initial investment based on a standard solution. Therefore, combining IMS with older SW technology proved to be the right move for optimum costs and reaching the final goal, the All - IP network.

It can be concluded that the unavoidable evolution of telecommunications services to the media, where the IMS as a global standard certainly has a key role in next-generation networks - NGN, where the IMS taken precedence.

LITERATURE

[1] http://www.etsi.org/tispan (TS 22 273, TS 181 002, TS 182 012, TS 183 043, ES 282 002)

[2] http://www.3gpp.org/ (TS 22.273, TS 32.280; TS 24.647)

[3] Internal T-HT documents

Development of the generic OFDM based transceiver in the LabView software environment

Damir Hamidović*, Nermin Suljanović**
*BH Telecom d.d. Sarajevo, Tuzla, Bosnia and Herzegovina
**University of Tuzla, Faculty of electrical engineering, Tuzla, Bosnia and Herzegovina
damir.hamidovic@bhtelecom.ba, nermin.suljanovic@untz.ba

Abstract - OFDM technique is the most adopted multiplexing technique offering many advantages such as high spectral efficiency, noise, inter-symbol interference and multipath robustness. These have made OFDM a first-choice technique in the majority of today's wireless and wired communication systems. In this work, physical layer of the generic transceiver, based on OFDM, is implemented. Focus is kept on the time and phase synchronization mechanisms as a key functionality of the receiver part. Complete implementation and simulation is conducted in the LabView software environment. BER results are obtained for different system parameters and presented in this paper. The transceiver implemented serves as a model for building enhanced, hardware device which could then be used for testing real, physical channels. Possibilities for the future work of using this device for testing real communications systems, are briefly described in this work. For this purpose, National Instruments 5640R platform, with FPGA core, is presented as an optimal hardware solution. Final section of this paper highlights possibilities of using this hardware solution for in depth analysis, testing and improving of high-demand communication systems.

Keywords: **OFDM, LabView, transceiver, synchronization, implementation, physical layer, development**

I. INTRODUCTION

OFDM technique is the most adopted multiplexing technique, offering many advantages such as high spectral efficiency, noise, inter-symbol interference and multipath robustness [1]. These have made OFDM a first-choice technique in majority of today's wireless and wired communication systems such as: DSL (*Digital Subscriber Line*) and DVB (*Digital Video Broadcasting*) systems, PLC (*Power Line Communication*), WLAN (*Wireless LAN*), WiMAX (*Worldwide Interoperability for Microwave Access*), LTE (*Long Term Evolution*) and many other demanding systems [1–3]. In this work, complete implementation and simulation is conducted in the LabView software environment. Implemented system is the extended version of the OFDM transceiver implemented in [4], with added features such as channel coder/decoder, model of Gaussian noise channel, time and phase synchronization mechanisms (a key functionality and the most demanding task in implementation of the receiver part). Various research groups [5–7] presented different synchronization techniques used as a theoretical basis for implementing fine and coarse time synchronization in this work. Model presented in this paper is a prototype for developing complete configurable SDR[1] (*Softwere Defined Radio*) simulation platform, for testing and developing new algorithms, or optimizing and enhancing existing, and for implementation of the extended OFDM transceiver. Furthermore, this simulation model can be used as a basis for transferring methods implemented in software to the FPGA (*Field Programmable Gate Array*) based hardware platform. This would enable testing and analyzing signals on real, physical channels, as reported by [7–9].

II. ARCHITECTURE OF THE OFDM TRANSCEIVER

Architecture of a fully functional OFDM transceiver, implemented as a testing prototype in the LabView, is presented in Fig. 2. As can be seen, architecture consists of three stages operating on the: bit level, complex-symbol level and OFDM symbols. Basis for this architecture is inherited from the work presented in [4], however enhancement has been made by implementing channel encoder/decoder, interleaver/deinterleaver, synchronization mechanisms, bits sequences generator, as well as by adding/removing test symbols to/from selected sub-channels, modeling AWGN channel and calculating BER (*Bit Error Rate*).

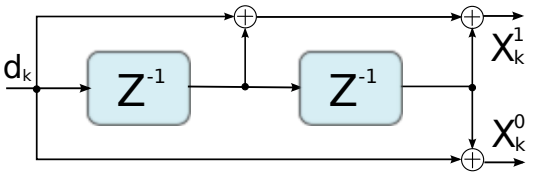


Figure 1: Schematic of the convolutional encoder, k/n=1/2

[1]Authors acknowledge Faculty of Electrical Engineering of the University of Ljubljana, for providing hardware and software resources needed for this work

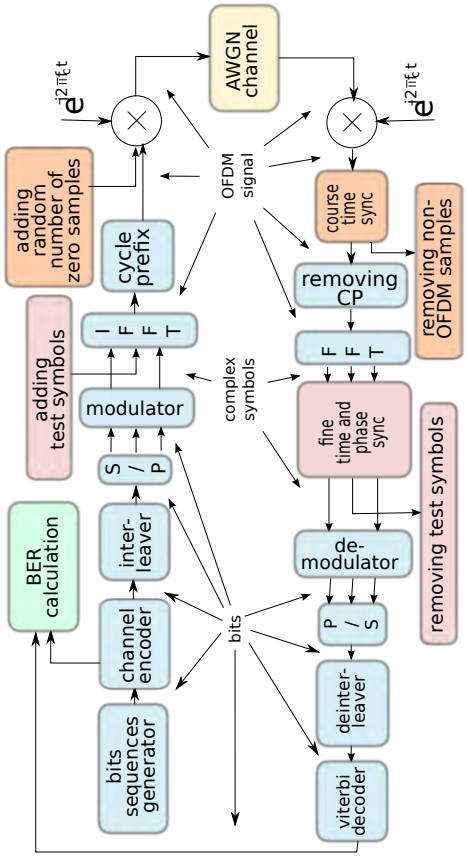


Figure 2: Architecture of a fully functional OFDM transceiver

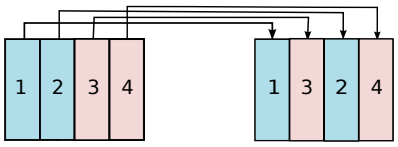


Figure 3: Work principal of the interleaver. Operations are carried out on a two-bit level

Channel encoder implements convolution encoding, as shown in Fig.1, with parameter *k/n*, where k is number of information bits and n is total number of coded bits. Decoding is carried out using *Viterbi* algorithm [10]. Additionally, work principal of the implemented interleaver is shown in Fig. 3 [10].

The purpose of adding channel encoder and an interleaver, in any communication system, is to make the system more robust to the noise intercepted in the channel [11]. In order to increase system scalability, all constructing parts of the system are implemented as separate blocks. If needed, these blocks can then be easily replaced with the similar, upgraded blocks, or replaced with the completely new parts. Additionally, benefits of implementing separate blocks reflect in their compatibility with other processing blocks, which creates numerous possibilities to easily enhance existing system. Furthermore, this implies that the system is highly suitable for testing and optimizing new algorithms. As mentioned previously, architecture also includes a separate block for generating random bits sequences on the transmitter side, as well as blocks for the model of the

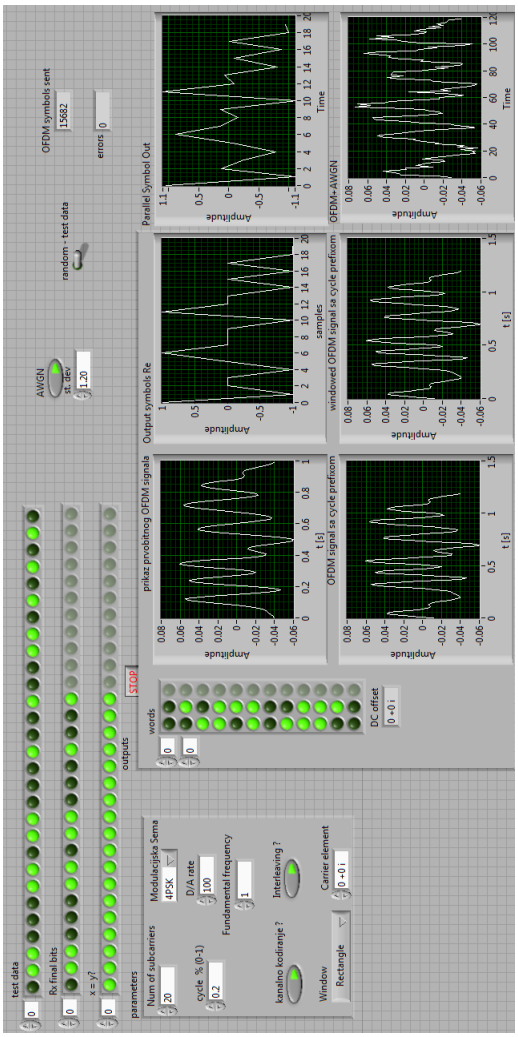


Figure 4: Labview user interface corresponding to the simulation model shown in Fig.*2*

Gaussian channel and synchronization mechanisms on the receiver side. For testing time synchronization mechanisms, simulation model for transmitting OFDM symbols, in random periods of time, is made by adding random number of zero samples between OFDM symbols on the transmitter side.

In the section III, synchronization mechanisms are described in detail. In [4], in depth LabView implementation for some parts of the basic model is presented, while complete implementation of the extended model can be found in [12]. User interface for presented simulation model, previously shown in Fig.2, is shown in Fig. 4.

III. Synchronization Mechanisms in OFDM

Being a multi-carrier multiplexing technique, OFDM experiences much higher sensitivity to the phase, frequency and time shifting, in comparison with the single-carrier techniques [5]. Furthermore, time synchronization demands increase when increasing the signal waveform complexity. Several different methods of synchronization are presented in [5], but in summary, all forms of synchronization in OFDM consist of the coarse and fine synchronization

phases, where coarse phase is followed by the fine phase. In coarse synchronization, mechanism is applied on the waveform coming out from the channel,

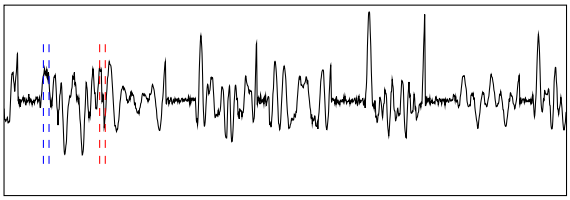


Figure 5: Waveform received from the channel (not plotted on scale)

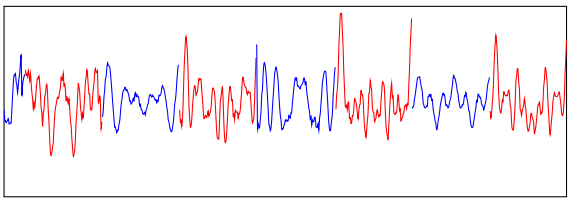


Figure 6: Coarse synchronized OFDM symbols (not plotted on scale)

while in the fine synchronization mechanism is applied on the complex symbols that are results of the FFT (*Fast Fourier Transformation*) algorithm on the OFDM waveform. Furthermore, fine synchronization aims to eliminate the errors resulted after applying coarse synchronization mechanism. To visualize time synchronization problem that needs to be solved, samples of the received waveform from the channel are shown in Fig. 5, and samples that need to be extracted from received signal, for continuing demodulating process, are shown in Fig. 6. In this work, coarse time synchronization utilizes feature of correlating a cycle prefix with the part of the OFDM symbol, which, in ideal case (without any disturbance), is exact replica of the cycle prefix. Theoretical background of this method is described in [5–7]. Principle of the coarse time synchronization is based on comparing two sliding windows (highlighted with dashed, blue and red, lines in Fig. 5). The lengths of the sliding windows are equal to the cycle prefix and the windows are mutually shifted by the number of samples in one OFDM symbol, *N*. In explanation, sliding windows are shifted by one sample along observed samples of the received signal, until the maximum number of samples (*M*) has been reached. Sum is then calculated for every position of the windows according to (1). In order to simplify calculations, sum is being calculated recursively as shown in (2).

$$\sum_{1}^{M} |X_n - X_{n+N}| \tag{1}$$

$$\begin{aligned}\sum^{M} |X_n - X_{n+N}|_k = \sum^{M} & |X_n - X_{n+N}|_{k-1} \\ & - |X_1 - X_N|_{k-1} \\ & + |X_{M+1} - X_{M+1+N}|_{k-1}\end{aligned} \tag{2}$$

In order to enable calculations of the starting point of the OFDM signal, package detection is performed by calculating the sum according to (3), and verifying if the obtained value is above the threshold value. Additionally, threshold value is obtained empirically and differs for different transmission parameters.

$$\sum_{1}^{M} |X_n|_k = \sum^{M} |X_n|_{k-1} - |X_1|_{k-1} + |X_{M+1}|_{k-1} \tag{3}$$

In the last equation, X_n is the value of the nth sample, M being the length of the cycle prefix and N being the length of one of the OFDM symbols.

Because of the noise component, the resulting sum in (1) will not have the value zero even in the case where one observed window is the cycle prefix of the other window. However, it is most likely that the minimum value of the sum will be obtained in that particular case. Only when minimum value of the sum is obtained and packet detection threshold is exceeded, starting point of any OFDM symbol can be estimated to within several samples. An error in estimating the starting point can occur as the result of: noise component, waveform of the OFDM symbol or similarity between adjacent symbols [5]. To eliminate possibly occurred error, fine synchronization mechanism needs to be employed. An example of calculating the sum from (1), in order to detect starting point, is shown in Fig. 7. Obtained results correspond to those obtained by the research group [7], due to the similar working principal used for obtaining the starting point.

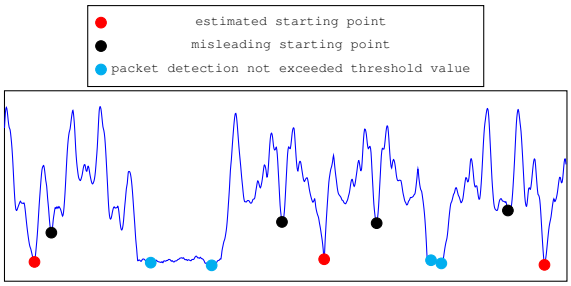


Figure 7: Calculating the correlation sum in (1) and detecting the starting point of the OFDM symbol

After the coarse time synchronization mechanism is applied, demodulation process, up to FFT stage, is applied on the samples of one of the OFDM symbols. FFT stage results with the complex symbols, potentially phase-rotated, where phase rotation is directly proportional to the number of samples for which the error is occurred during the coarse synchronization stage. This fact is carried out directly from the basic Fourier transform property, shown in (4) [13].

$$\begin{aligned} IF &: \mathscr{F}\{x_n\} = X_k \\ THEN &: \mathscr{F}\{x_{n-m}\} = X_k e^{-\frac{2\pi i}{N} km} \end{aligned} \tag{4}$$

The fine sync mechanism is implemented by adding several test subcarriers (complex test symbols), besides information symbols, and calculating relative phase rotation

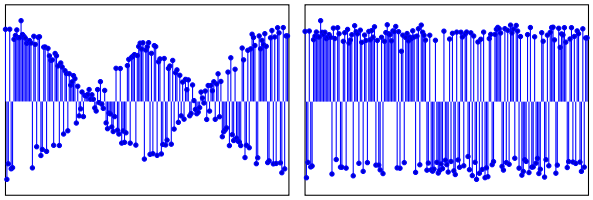
offset between test subcarriers. In theory, this could be calculated using just two test subcarriers, but in this work, higher number of test subcarriers is used. This will enable averaging the relative phase rotation which leads to suppressing the random phase rotation, caused by the noise component and added by the channel. Average value of the phase rotation (obtained from the adjacent test subcarriers), number of samples for which the error during the coarse synchronization phase occurred by, and rounded phase-rotation value, have to be considered for every received subcarrier while performing error correction calculations. Equations (5), (6) and (7) give the mathematical expressions for described calculations.

$$\tilde{m} = \frac{\sum_{i=2}^{t} \angle((S_{test})_i - (S_{test})_{i-1})}{f_{max}} \tag{5}$$

$$l = round\left(\frac{\tilde{m} \cdot N}{2\pi}\right) \tag{6}$$

$$m = \frac{l \cdot 2\pi}{N} k, \quad -\frac{N}{2} \leq k \leq \frac{N}{2} \tag{7}$$

In (5), t is number of test subcarriers and f_{max} is frequency of the last test subcarrier. Fig. 8 presents cases before and after applying the fine synchronization mechanism.



(a) (b)

Figure 8: Subchannels: *a)* before applying the fine time synchronization mechanism and *b)* after applying the fine time synchronization mechanism. The case is considered for an error of the one sample in the coarse synchronization phase. (Not plotted on scale)

In Fig. 8 *a)* and *b)*, real components of the complex symbols of the subcarriers are shown. In this case, OFDM symbol consists of *210* subchannels (*200* information and *10* test subcarriers), carrying one bit of information on each subcarrier. Waveform of the OFDM symbol contains *211* samples in time domain.

Besides time synchronization, this work considers phase synchronization mechanism as well. Although, phase synchronization is not used in LabView simulation model, it was still implemented separately and employed for receiving OFDM symbols via real, physical channel, with hardware components. Phase error on the receiver side can be easily determined by calculating the phase difference between one transmitted and received symbol on the same subcarrier. Theoretically, this can be done with just one subcarrier used, but, for the same reason as for the fine time synchronization, phase synchronization mechanism is implemented with multiple test subcarriers.

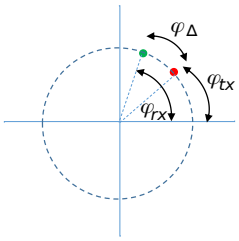


Figure 9: Working principle of the phase synchronization, where ϕ_{tx} is phase of transmitted symbol, ϕ_{rx} phase of received symbol and $\phi_\Delta = \phi_{rx} - \phi_{tx}$

$$IF : \mathscr{F}\{x_n\} = X_k \tag{8}$$
$$THEN : \mathscr{F}\{x_n e^{j\phi}\} = e^{j\phi} X_k$$

Moreover, the same subcarriers used for fine time synchronization have been used for this synchronization. As the phase shift is added in the channel on the signal waveform, the phase synchronization mechanism can be applied on the symbol values, obtained after applying FFT algorithm, due to the *Fourier* property presented in (8). Working principal of the phase synchronization mechanism is shown in Fig. 9 [13].

IV. Results and Discussion

Fig.10 presents BER results for different system parameters (block diagram of the system is shown in Fig.2). For obtaining these results, ideal synchronization between transmitter and receiver was assumed. Additionally, number of samples of one OFDM symbol, in time domain, and length of the cycle prefix, were kept fixed on *1000*, and *20%* of the length of one OFDM symbol, respectively, while obtaining these results. By decreasing the number of bits per symbol, or using less subchannels for the same number of samples in time domain, the system robustness to AWGN noise increases, which means that the probability of the error is lower for the same SNR. Fig.10 *a)* and *b)* present these results graphically. Moreover, system can be made even more robust, with performing channel encoding and/or interleaving, as it can be seen in Fig.10 *c)*. Moreover, system robustness can be furthermore increased, even beyond the results previously presented on Fig.10 *a)* and *b)*. The results obtained, and shown in Fig.10 *c)* show that while manipulating input bits by preforming channel encoding and/or interleaving, the system robustness additionally increases. In summary, results presented of Fig.10 demonstrate that the system robustness highly depends on the number of bits per symbol, number of subchannels, and channel encoding/interleaving.

Fig.11 graphically presents the influence of the time synchronization mechanisms for the different OFDM modulation parameters. Fixed parameters for the simulations presented in Fig.11 are: *1000* information subchannels, the cycle prefix of *2%*, and the number of samples of one OFDM signal of *1051*.

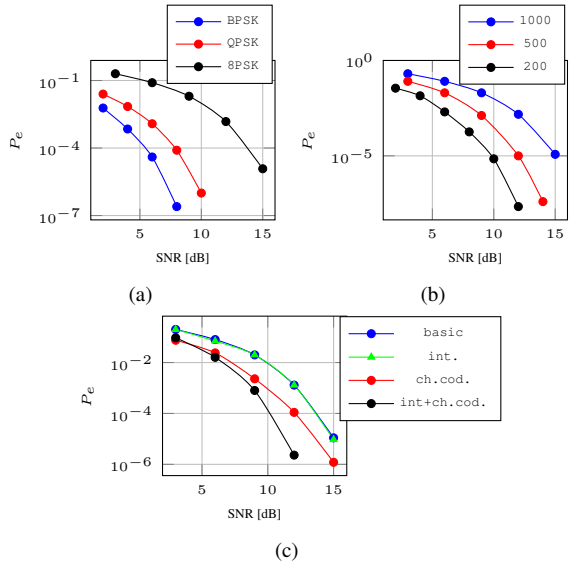


Figure 10: Simulation results for: *a)* different modulation schemes (the number of subchannels of *1000*); *b)* modifying the number of subcarriers for fixed modulation scheme (*8PSK*-3 bits per subchannel); *c)* channel coding and interleaving for fixed modulation scheme (*8PSK*) and fixed number of subchannels (*1000*)

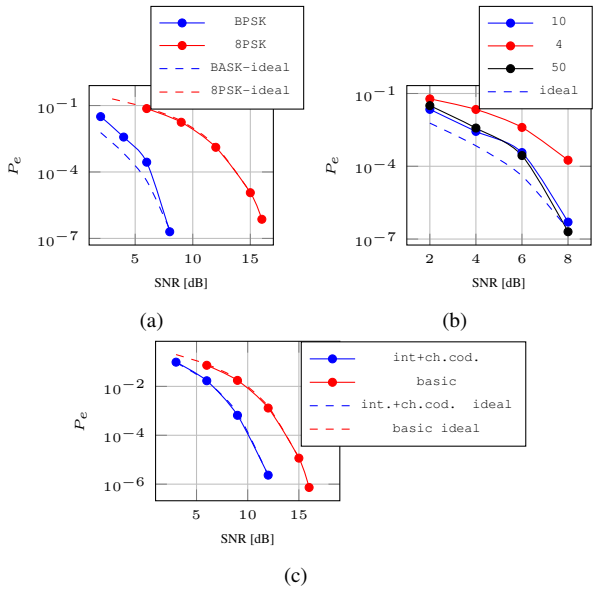


Figure 11: Probability error simulation results for: *a)* different modulation schemes and fixed number of test subchannels (*50*); *b)* different number of test subchannels and fixed *BASK* modulation scheme; *c)* with and without channel encoding and interleaving for fixed *8PSK* modulation scheme. Number of bits transferred in each simulations is 10^8.

Ideal synchronization was simulated in order to verify obtained results for when synchronization mechanisms are applied on the system. The performance of the synchronization mechanisms can then be evaluated by measuring the error probability, for different SNR ratio, and comparing the obtained values with the ideal case. Fig.11*a)* presents error probability curves for different modulation schemes. Results for both *8PSK* and *BPSK* are presented and it can be seen that *8PSK* experiences higher coincidence with the ideal synchronization, while *BPSK* requires SNR higher than *8[dB]* in order to not increase error probability caused by the synchronization. Fig.11 *b)* shows the changes in the error probability for different number of test subchannels. As can be seen, for higher number of test symbols, the error probability, caused by the synchronization, decreases. In this regard, the results for the system with four test subchannels do not coincidence with the ideal results at any point, which indicates that the time synchronization mechanism cannot be performed for any of the tested SNR ratios. Meanwhile, the system with *50* test subchannels demonstrates results overlapping for SNR higher than *8[dB]*. Additionally, Fig.11 *c)* shows that by adding channel encoding and interleaving will not affect the time synchronization. This is due to the fact that encoding and interleaving are bit level operations and do not affect the signal waveform in any manner.

V. Conclusion and Future Work

The results presented demonstrate the generic OFDM-based system. Functionality of the time synchronization mechanisms was verified by calculating the error probability and comparing the obtained results with the results of the ideal, synchronized system (Fig.11). It was noticed that the synchronization can become dubious if the signal-to-noise ratio and/or the number of the test subcarriers used is to low. Moreover, it was seen from the results presented in Fig.10, changes in system parameters can result in different robustness to AWGN. However, compromise to the robustness, which is a bit throughput that changes as the number of information bits in one OFDM symbol changes, should be considered as an important factor. These changes can be made by applying different modulation schemes or using different number of subchannels per one OFDM symbol. Additionally, performing channel encoding with parameter *k/n* increases robustness, while bit throughput is being decreased by the radio *n/k*.

In addition to the simulation model, made in the LabView application, the connection with a hardware platform, for transmitting and acquiring OFDM symbols to/from real, physical channel, was made. Hardware platform used in this work is *NI PCI 5640R* RIO (*Radio Input-Output*) platform, based on FPGA core with *D/A*, *A/D* converters with multiple channels and other surrounding fixed resources. More detailed description of this platform, as well as in-detail procedure followed in this work for connection of the hardware platform and the LabView application, are presented in [14]. Connecting LabView and hardware platform presents the first stage of developing generic platform for testing OFDM-based systems in real, physical environments. Possibilities for the future work include transferring the physical-layer processing from the LabView application to the FPGA core. Moreover, development of the FPGA logic design can be done in the LabView as it has FPGA Module for transferring graphically design in VHDL (*VH-*

SIC Hardware Description Language) that can be uploaded directly into the FPGA code and define hardware behavior for performing necessary processing [14]. Additionally, specific parts of the system could be replaced with the enhanced ones, such as channel coder and interleaver, or even by adding completely new processing blocks (scrambler or waveform-windowing block).

Higher levels, apart from physical, could be implemented in the FPGA core or LabView application part, for further modem design possibilities. The advantages of using this concept for research and developing various communication systems, are presented in [15]. For instance, one of the today's latest research in the field of mobile communications, developing *5G* standard, includes complete concept development in the LabView environment, and is described in details in [16].

The system in this stage, presented in this work, can serve as a prototype for an enhanced research and development of the platform that could then be used for developing new algorithms or analyzing the influence of the channel to different parameters of the system.

ACKNOWLEDGMENT

D.Hamidović expresses sincere thanks to Medina Džebić for critical reading and reviewing of the paper.

REFERENCES

[1] Lajos Hanzo, Yosef Akhtman, Li Wang and Ming Jiang, *MIMO-OFDM for LTE, WiFi and WiMAX: Coherent versus Non-coherent and Cooperative Turbo Transceivers*, Wiley-IEEE Press, 2010, Chapter 1, ISBN: 978-0-470-68669-0

[2] Fahad Khan, Sobia Baig, Umber Noreen and Adnan Yousaf, *An Overview of OFDM Based Narrowb and Power Line Communication Standards for Smart Grid Applications*, World Applied Sciences Journal 20 (9): 1236-1242, 2012, ISSN: 1818-4952

[3] National instruments Corporation, *OFDM and Multi-Channel Communication Systems*, 2014, opensource material available at: http://www.ni.com/white-paper/3740/en/

[4] Aditya Jain, Bryan Paul, Rice University, Houston, Texas, *Fully Configurable OFDM SDR Transceiver in LabVIEW*, 2012, opensource material available at: cnx.org/content/col11182/1.6/

[5] Fuqin Xiong, *Digital multiplexing techniques*, ARTECH HOUSE, INC., Second Edition, 2006, Chapter 14, pp.750-797, ISBN:1-58053-863-0

[6] Konstantinos Nikitopoulos, Andreas Polydoros, *Post-FFT Fine Frame Synchronization for OFDM systems*, 2010, available at: http://citeseerx.ist.psu.edu/viewdoc/download ?doi=10.1.1.577.250& rep=rep1& type=pdf

[7] Stefan Johansson, Martin Nilsson, Peter Nilsson, *An OFDM Timing Synchronization ASIC*, 2000, conference paper, ISBN:0-7803-6542-9, available at: https://lup.lub.lu.se/search/publication/1034063

[8] Chris Dick, Fred Harris, Michael Rice, *Synchronization in Software Radios-Carrier and Timing Recovery Using FPGAs*, 2000, available at: http://inst.cs.berkeley.edu/∼ n225c/sp08/synch_paper.pdf

[9] Chris Dick, Fred Harris, Michael Rice, *FPGA Implementation of Carrier Synchronization for QAM Receivers, 2004, Journal of VLSI Signal Processing, Kluwer Academic Publishers*

[10] Fachbereich Elektrotechnik und Informationstechnik, opensource material available at: ems.eit.uni-kl.de/fileadmin/ems/user_upload/Appendix_task7_8.pdf

[11] Ahmad Jalali, *Channel coding and interleaving for transmissions on a multicarrier system*, patent: *WO 1998059473 A1*,December 30, 1998

[12] Damir Hamidović, *Implementacija OFDM primopredajnika na FPGA platformi korištenjem LABVIEW okruženja*, master thesis, 2014, Chapter 4 (unpublished material)

[13] Marc Ph. Stoecklin, *Tables of Common Transform Pairs, version v1.5.3*, 2012, available at: www.mechmat.ethz.ch/Lectures/tables.pdf

[14] National Instruments Corporation, *Getting Started with the NI PCI-5640R IF Transceiver and the LabVIEW FPGA Module* 2007, available at: *http://www.ni.com/pdf/manuals/371621b.pdf*

[15] National instruments Corporation, *Advantages of Using LabVIEW in Academic Research*, 2012, available at: *http://www.ni.com/white-paper/8534/en/*

[16] Eric Luther, *5G Massive MIMO Testbed: From Theory to Reality*, 2015, available at: *http://www.ni.com/white-paper/52382/en/*

The Challenge of Cellular Cooperative ITS Services Based on 5G Communications Technology

Zdenko Kljaić, M.Sc.EE / M.Sc.Traff.*, Pero Škorput, Ph.D.Traff. **, Nataša Amin, M.Sc.EE ***
* Ericsson Nikola Tesla d.d. Zagreb, zdenko.kljaic@ericsson.com
** Faculty of Traffic and Transport Sciences, Zagreb, pero.skorput@fpz.hr
*** Ericsson Nikola Tesla d.d. Zagreb, natasa.amin@ericsson.com

Abstract - We live at a time when the automotive industry is going through a technological revolution with the development of vehicles changing into autonomous moving objects having the properties of artificial intelligence. Additionally, cellular communications networks introduce new technologies and concepts: SDN (Software-defined networking), NFV (Network Functions Virtualization). These advanced software-defined communications networks virtualize network functions, allowing a new way of configuration, control and management. Increasing the speed and automation are key requirements to support the most demanding services such as mobile payment of contextual services, as well as the introduction of new "Machine" users. SDN also assists in the implementation of new infrastructure for the dynamic services that are based on the concepts of IoT, Big Data and Everything-as-a-Service. Using NFV enables the Internet of Things services that provide a new way of connecting people, processes, data and devices. It is precisely these requirements that are essential for the introduction of C-ITS systems, i.e., the integration of passengers, drivers, vehicles and transport infrastructure, as well as information, statistics, predictive traffic analytics, all this in real-time. Due to this trend of development of cellular communications networks in the next 5G communications networks, automotive and ITS traffic systems are becoming the most important market for business expansion of telecom operators. Such revolutionary technological changes, together with the integration of various industries entails a series of challenges. The aim of this paper is to define important information, challenges and opportunities for the telecom industry to provide mobility for people and goods.

Keywords:C-ITS, cellular, SDN, NFV, 5G, opportunity, telecom industry.

I. INTRODUCTION

At the beginning of 2016, the Dutch Ministry of Infrastructure and Environment together with HERE company, started a project for solution development and evaluation of a cellular network-based „Cooperative Intelligent Transportation System“ (C-ITS).

The evaluation focuses on a system utilizing standard commercial LTE cellular networks as well as location cloud technology and data analytics from HERE to connect mobile phones and other devices, road infrastructure and traffic management centers. The system would serve as an efficient low latency data exchange mechanism, whereby targeted information - about, for example, an accident or black ice - could be communicated to the right people at the right time and location.

The vast majority of Dutch drivers are using mobile phones and, increasingly, cars and infrastructure are getting connected to the Internet. There is an opportunity to take advantage of this connectivity to improve road safety and traffic flow in the next few years. In particular, in 2016, the goal is to start deploying a road user messaging system that enables information to be gathered, processed and then distributed to the relevant road-users with quality, accuracy and at sub-second delivery rates.

The ministry has been keen to explore the potential of cellular networks in improving mobility in the Netherlands. C-ITS solution can be built on existing Dutch networks, meaning that it would entail large costs, could be quickly rolled out across the whole country and would bring benefits to citizens relatively fast. The Netherlands is regarded as a C-ITS leader in Europe, and we await with great interest the results of the C-ITS project in real urban environment.

II. ITS - INNOVATING FOR THE TRANSPORT OF THE FUTURE

Intelligent transportation systems (ITS) are ICT applications in the areas of transportation and traffic, which manage the collection, processing, presentation and usage of transportation and traffic data, thus providing innovative services related to these fields. ITS intelligence lies in its infrastructure such as sensor networks, communications networks, data processing and adaptive behaviour of systems as such. The aim of ITS management is to ensure that various users have access to better information while emphasizing safety, coordination and smart use of transport networks.

The definition of ITS according to the EU Directive 2010/40/EU (July,2010) is as systems in which information and communication technologies are applied in the field of road transport, including infrastructure, vehicles and users, and in traffic management and mobility management.

In order to transmit ITS knowledge and information to all actors at the local and national country levels and support ITS promotion, the Network of National ITS Associations was formed. The Network, launched in October 2004 in London, currently consists of 27 member organizations. Its Secretariat is at ERTICO – ITS Europe, which is a multi-sector, public/private partnership working on the development and deployment of Intelligent Transport Systems and Services by connecting public authorities, industry players, infrastructure operators, users, national ITS associations and other organizations together to bring "Intelligence into Mobility".

Currently, ERTICO programme focuses on improving transport safety, security and network efficiency while reducing environmental impact. The vision of the future are transport systems working towards zero accidents, zero delays and fully informed people, providing affordable and seamless services, protected environment, while respecting privacy and security.

Several forms of wireless communications technologies have been proposed for intelligent transportation systems, such as radio modem communication, used for short and long range communication within ITS, and short-range communications, accomplished using IEEE 802.11 protocols. Specifically wave or the Dedicated Short Range Communications standard is being promoted by the Intelligent Transportation Society of America. The range of these protocols can be extended using Mobile ad hoc networks or Mesh networking. Longer range communications have been proposed using infrastructure networks such as IEEE 802.16, Global System for Mobile Communications (GSM), or 3G/4G/5G. Long-range communications using these methods are well defined. However, unlike the short-range protocols, these methods require extensive and very expensive infrastructure deployment. There is lack of consensus as to what business model should support this infrastructure. Auto Insurance companies have utilized ad hoc solutions to support eCall and behavioural tracking functionalities in the form of Traffic/Vehicle Telematics ver.2.0. Recent advances in vehicle electronics have led to a move towards fewer, more capable computer processors on a vehicle.

There is a trend towards fewer, more costly microprocessor modules with hardware memory management and Real-Time Operating Systems. The new embedded system platforms enable more advanced software applications to be implemented, including model-based process control, artificial intelligence, and ubiquitous computing. Probably the most important of these for ITS Systems is artificial intelligence.

III. C-ITS Cooperative Intelligent Transport Systems

Cooperative ITS (C-ITS) are Intelligent Transport Systems where ITS stations: vehicles, roadside equipment, traffic control centres and nomadic devices communicate and share information using a standardized communication architecture - the ITS Station Reference Architecture, to provide road users with better road safety, traffic efficiency, comfort, improved mobility and sustainability. It is envisioned that the technology and deployment status of Cooperative Intelligent Transport Systems would use different access technologies: cellular 3G/4G/5G, vehicular WiFi 11p and radio broadcast. Moreover, these technologies will complement each other optimally to provide a set of services, which would evolve over time. Definition of Cooperative ITS (ISO 17465): A co-operative ITS is a subset of the overall ITS that communicates and shares information between ITS Stations to give advice or facilitate actions with the objective of improving safety, sustainability, efficiency and comfort beyond the scope of stand-alone systems. ITS Stations are defined in ISO 21217 / ETSI EN 302 665, as units installed in vehicles, at the road side, in traffic control/management centres, in service centres, or in hand-helds.

The Car 2 Car Communication Consortium and the Amsterdam group ASECAP decided to pursue a gradual deployment approach with an initial deployment of simple use-cases which give user benefits although C-ITS penetration in vehicles is still little and equipped road side ITS-stations in hot spot areas is equally not ample. The level of complexity of use cases will increase as time goes by to include crash avoidance and other “hard safety” applications. Additionally, in the future the number of vehicles with C-ITS equipment and infrastructure coverage will increase as well.

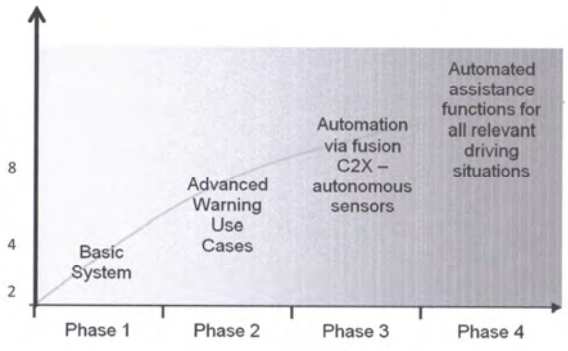


Figure 1: Deployment roadmap by the C2C, ETSI ITS-G5 based C-ITS.

It is important to define the set of applications for initial deployment of cooperative ITS. Since various applications have different operational and functional requirements as well as different latency requirements that would lead to specific communication technologies. Moreover, services related to safety and traffic efficiency as well as use cases based on the exchange of information between vehicles and the roadside infrastructure V2X (V2X - Vehicle-to-Everything, i.e. a collective term for vehicle-to-vehicle (V2V) and vehicle-to-infrastructure

(V2I) technologies), for the time being use the vehicular WiFi technology (ETSI ITS G5 within the 5.9 GHz band). However, in the future a "hybrid" system can be used along with current technologies.

IV. THE DIGITAL TRANSFORMATION OF INDUSTRY AND SOCIETY AND THE NEED FOR RAPID DEVELOPMENT AND IMPLEMENTATION OF 5G MOBILE NETWORKS

In 2008, the ITU-R of the International Telecommun. Union issued requirements for International Mobile Telecommunications – Advanced. This is marketed as 4G mobile phone and Internet access service. The ITU defined 4G data transfer rate: 100 Mbps when user is moving, 1 Gbps when stationary. However, LTE does not yet meet all the required conditions, but confirms the opposite for marketing reasons.

The fifth generation new mobile network is based on a completely new architecture, which has previously been identified by the radio access: analogue, TDMA-Time division multiple access, UMTS-Universal Mobile Telecommunications System and OFDMA-Orthogonal Frequency Division Multiple Access.

The next fundamental step beyond fourth generation OFDMA or LTE networks is referred to as being 5G. 5G requires a new radio access technology, and a new standard to fulfill current subscriber demands that older technologies cannot satisfy. On the other hand, 5G research is driven by current traffic trends and requires a complete network overhaul that cannot be achieved gradually through network evolution. The requirements that cannot be provided by LTE networks include software-driven architectures, fluid extremely dense networks, higher frequency, wider spectrum, devices by the billions and large Gbps capacities.

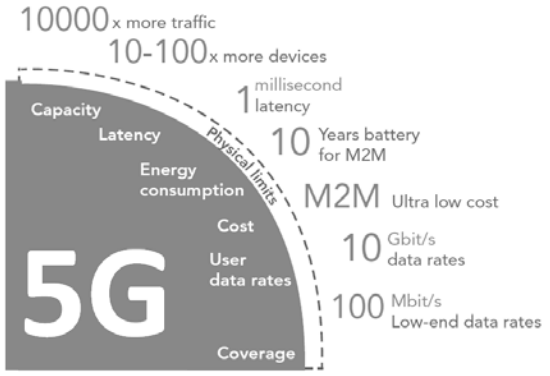


Figure 2: Expected technical capabilities of 5G mobile com. technology.

The evolution of mobile telephony from the analog mobile 1G generation to the last implemented 4G third generation brought about a paradigm shift. The goal of the new mobile technology is not to improve the communication experience but to change the user access into a new global communication reality. This technology shift aims to provide users with a new set of services that support ubiquitous communication (i.e. access that is available every time, everywhere) .

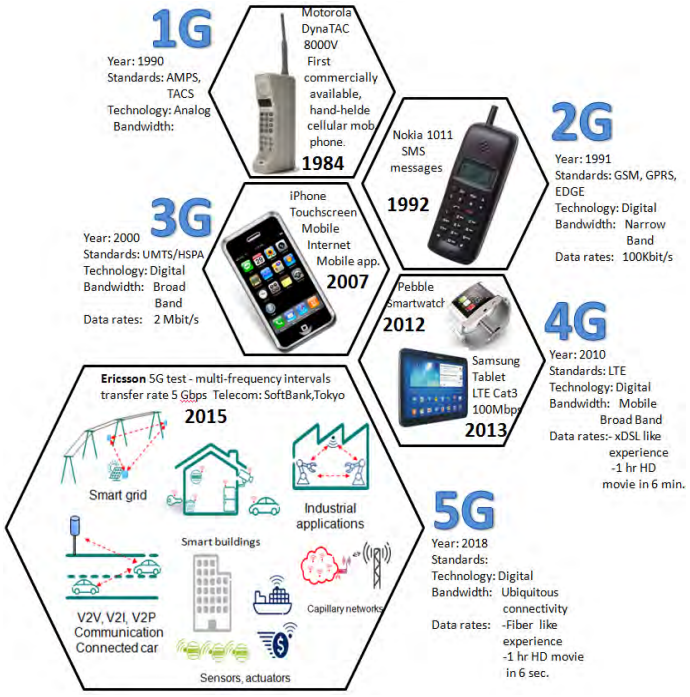


Figure 3: Evolution of Mobile Communications: 1G - 5G Mobile Technologies

The latest step in this evolution, the 5G technology, adds more services, data, use and benefits to the upcoming generation. 5G will be smarter technology with no limits in order to interconnect the whole world seamlessly. The upcoming lifestyle will be different with uninterrupted access of information and interconnection.

V. CAN CONNECTED VEHICLES AND C-ITS DRIVE THE NEED FOR 5G COMMUNICATION TECHNOLOGIES ?

Even though it would seem that 5G communications have been conceived just in order to increase the speed of communication between smart phones, this is not the case. The purpose of 5G is also to provide a wide range of M2M and IoT services and applications. These include support for connected cars and driverless vehicles. The importance of V2V (vehicle to vehicle) and V2I (vehicle to infrastructure) is clearly becoming more evident to the industry, driven especially by the promise of driverless vehicles in which these technologies have a key role. For this, automotive industry such as GM announced a 2017 Cadillac CTS model that will be equipped with V2V technology, combined with the Super Cruise ADAS solution. Additionally, other automotive industries such as Toyota and Audi are also committed to V2V. But while early implementations of V2V-equipped vehicles will start appearing in the next few years, driven by Europe's C2C-CC (car to car communication consortium), wide-scale deployments can be expected around 2020.

Various manufacturers such as Volvo performed LTE trials for testing non-safety critical V2V applications, such as exchanging non-time critical information including hyperlocal weather, road conditions, and traffic data directly between vehicles. However, Dominique Bonte (Managing Director, ABI Research's global automotive) argues that only the arrival of 5G, offering latencies as low as 1ms, will challenge the car industry's dedicated short-range DRSC V2V technologies. According to Alexander Scheidt (Global Automotive Industry Leader at IBM Global Business Services), by 2025, the industry will not only recreate highly personalized and digitized lives inside cars, but will also give consumers a bigger role in defining that experience, whether as a driver or passenger. These transformations in the transport industry can be performed only if 5G communications networks come to life.

It is expected that 5G will see its first deployments in 2020, providing a number of benefits over current 4G networks. This includes downloads speeds of up to 10Gbps, five times lower latencies, and vastly improved network coverage. 5G will also offer more intelligent use of spectrum and greater reliability as service providers deploy software-defined networking and adopt cloud architectures.

With the progress of connectivity solutions in ITS systems and digitalization of traffic and transport, a need arose for change in development and consumption of new services. 5G communications networks are taking over a key role in platform development for the digitalization of different industries. Such platforms opened up a chance for cooperation between telecom operators and service operators, but with it a range of challenges appeared such as interoperability and price pressure. The key to success in the digital market of "traffic and transport" is the ability to adapt, and true business agility requires flexibility in all three dimensions: telecom networks; services and payment; and customers (drivers, travelers, cargo). The main elements of network agility, i.e., the capability of the network to adapt to quick change are: Cloud, SDN and NFV. The network should have the capability to be planned and built efficiently, to adapt to changing requirements, and should provide superior service quality.

VI. THE ROLE OF NFV AND SDN IN 5G

SDN and NFV are vital in advancing the level of network agility in order to explore the opportunities and address the challenges coming from Networked Society and the digital economy.

It is expected from 5G telecom networks to efficiently support a wide range of applications, such as services based on machine-type communications, e.g., intelligent transportation systems and highly-reliable low-latency industrial applications. Providing these new services requires a flexible and programmable infrastructure, which enables agile service composition and scaling in a cost- and resource efficient manner. SDN and NFV are promising technologies with the potential to fulfill the flexibility and programmability requirements of 5G. While SDN separates the network control- from the data plane and introduces abstractions and means for network programmability, NFV separates network functions from hardware appliances enabling the realization of network functions on general-purpose processors (i.e., separating network services from platforms). SDN and NFV are complementary approaches which have the potential to offer together new ways to design, deploy, operate, and manage networks and their services.

Virtualization technologies using software-defined networking and virtualization of network functions will play a significant role in the development of next-generation mobile technology standards expected to rollout under the 5G concept. One reason would be the need for more rapid scalability by using automation designed to allow mobile networks to keep up with growing demand. One of the challenges for the use of virtualization for 5G is moving network elements closer to the base station to allow for lower latency. The need to use virtualization for network resource provisioning is exemplified by a "smart" building that one moment is sending information to the network saying that it is all right, to the next moment saying that a fire broke out and that building is now required to start streaming gigabits of video content to first responders.

Such capabilities have not been entrusted to current network deployments for now due to lack of capabilities. Therefore, the solution is not just to make the pipe bigger, but to offer user experiences and inclusion of vertical sectors that were excluded in 4G up to now.

The industry would reconsider current views on network ownership, since for telecom operators to gain the most efficiency from their resources they may need to give up some of the core operations in order to gain advantages in service offerings.

OSS/BSS natively account for SDN and NFV, since they are designed according to a functional division of network architecture domains. The OSS/BSS and service enablement architecture for SDN/NFV-enabled networks is shown.

The diagram depicts the main functional blocks: OSS/BSS and Service Enablement (SE), network functions (NF), equipment (physical resources), the cloud system infrastructure, and transport. Two types of resources can support an NF: native resources (non-virtualized, physical NF) or by virtual resources (a virtualized application or a virtualized NF). NFs are managed across two planes:
- the netw. function domain manag. plane - shown as NF domain management, for fault management, performance management and specific configuration for NFs
- the supporting resources management plane – shown as vApp management, which handles resources required by a network function throughout its life cycle

The functions in the OSS/BSS and SE plane are:
- experience and assurance – offering service assurance;
- customer and partner interaction – enabling both parties to interact with support systems through multiple communication channels;
- order management;

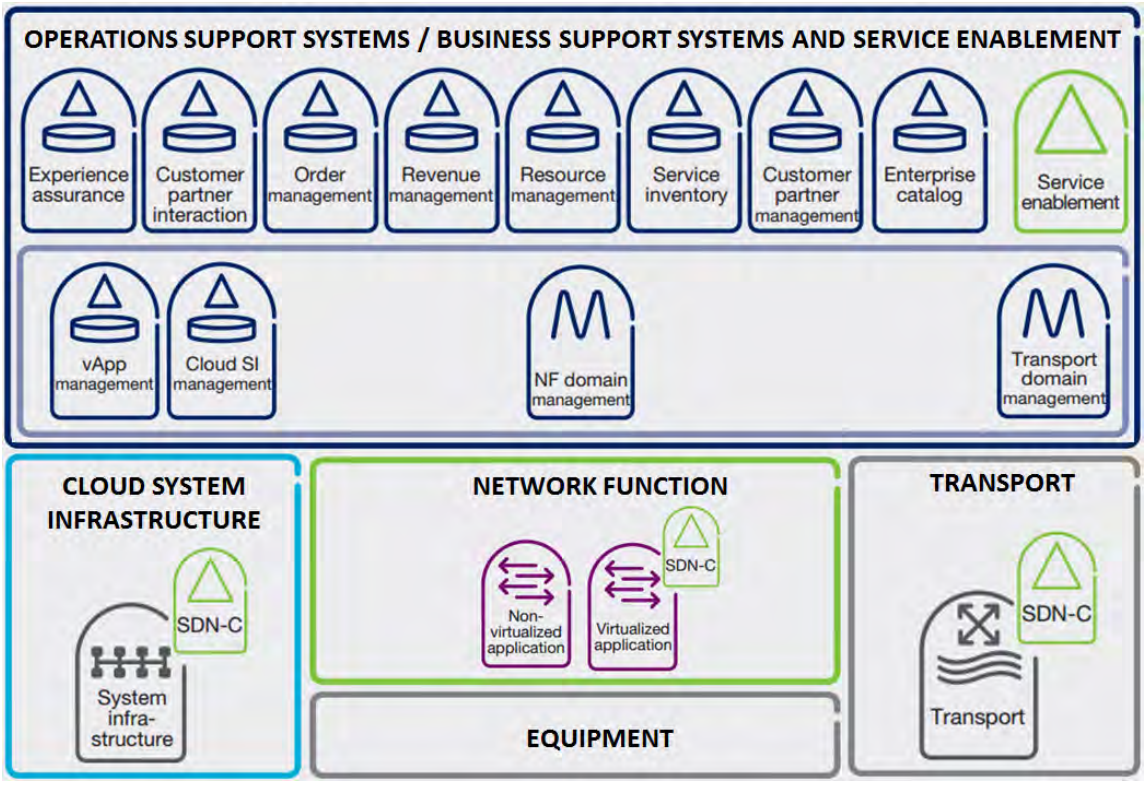


Figure 4: Operations Support Systems/Business Support Systems architecture for Software-defined networking/ Network Functions Virtualization - enabled networks

- revenue managem. - providing the capabilities to charge and invoice for any type of product or service usage;
- resource management – providing a unified resource inventory for both virtual and physical resources;
- service inventory;
- customer/partner management;
- enterprise catalog – consisting of products, services and resources; and
- service enablement – providing service exposure capabilities to partners for service innovation.

Support systems (OSS/BSS – Operations Support Systems/Business Support Systems) provide the capabilities to manage partners and developers, to handle all communication channels, and to organize the administration of products and services. OSS/BSS capabilities can be combined with technologies such as SDN and OpenStack that provide developers with programmable interfaces, so that new services can be deployed in isolated virtual environments. OpenStack and OpenDaylight APIs show network programmability while developers are given access to other services and features like user identification, charging and network policies, and configuration information to program NFs.

Operators and service providers can develop new services for traditional segments by use of NF virtualization. Additionally, they can enter new markets.

One such example is that virtualization enables bundles that include connectivity services to be mixed with value-add services and set up in a one-stop-shop fashion, which can be offered to various industry verticals.

The advantage of NF virtualization can be shown when compared to traditionally offered industry verticals where connectivity services offering provide network connectivity optimized for the specific vertical. In a virtualized environment, optimization is simplified, as NFs can be instantiated for a particular vertical, as shown in the figure.

Also the figure depicts how NFs and support systems interact. NFs provide the connectivity to connect everything in the network together, such as mobile phones and other handheld devices, as well as cars, and health care and transportation equipment. The OSS/BSS support systems manage the NFs translating their capabilities into services that can be offered to industry verticals through operator and service provider capabilities.

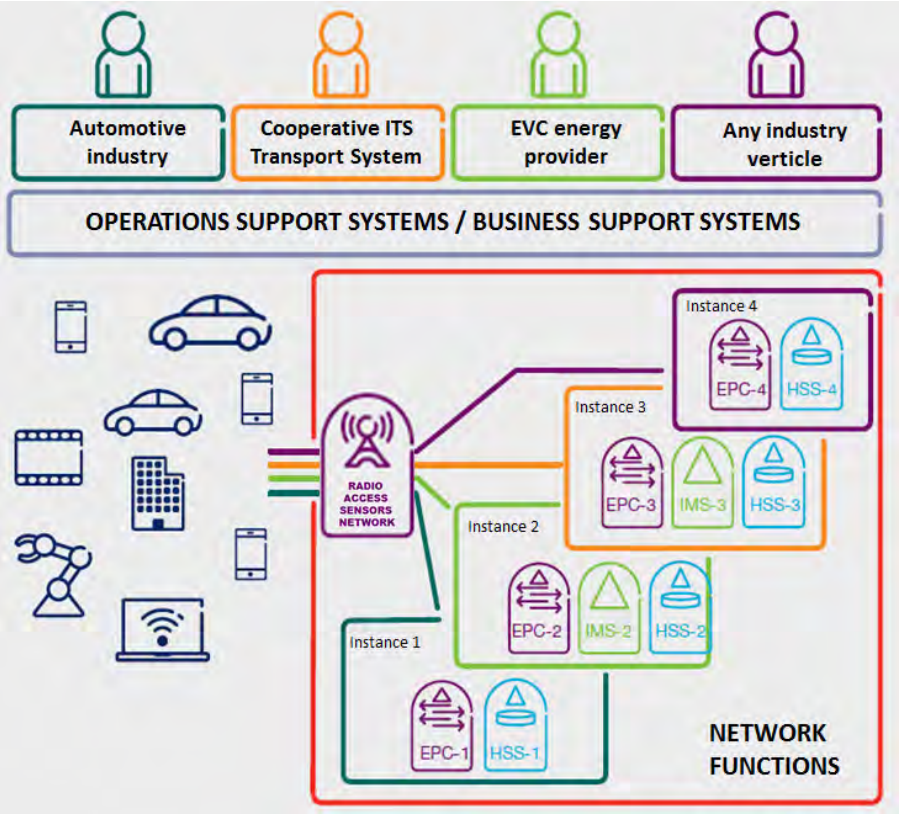


Figure 5: Providing different device connectivity and new services with Network Functions Virtualization

VII. ITS - DIGITALIZATION OF TRAFFIC AND TRANSPORT AS NEW OPPORTUNITY TO TELECOM INDUSTRY

The current trend is that of digitalization of industry and society. Digitalization cuts down the amount of traditional products and services and business models, but offers new opportunities for growth. New industries and products appear on the market while traditional ones disappear. For example, by launching e-watch, Apple has become the world's largest manufacturer of watches in the first year of offering to the market, though until then they did not have any product related to watches in their portfolio.

On average we spend 8.5 hours a week in our cars and by 2018 most new cars sold around the globe will have onboard digital connectivity. Revenue forecasts for connected cars is predicted to reach $172 billion by 2020, so it is little wonder that automotive companies, technology giants like Google and Apple and telecom companies around the world are all looking to carve out a sizeable chunk of this market.

The business trend in EU telecoms markets is in a mindset of accelerating technological change in the adverse economic climate. The competitive race is becoming tougher in the declining classic telecom market, in conditions of income reduction due to stagnating demand and falling prices of electronic communications services.

Telecom operators responded to these adverse market trends by rationalizing costs, and consolidating and focusing operations on specific, more important parts of their core business activities.

Upon digitalization of certain industry branches, there was a danger of repeating the previous "business story" with OTT players, i.e. telecom operators are restored again to being "bitpipes" for new digital services and companies. Telecom operators, however, cannot allow this and are entering new non-core areas such as banking services, sale of electricity, retail and so on. By entering new markets, telecom operators have met several challenges such as lack of competence, lack of knowledge of the market and competition, and regulatory issues. Telecom operators have sought solutions in the fields of Smart City, Electric Energy, environment and so on. But because of these challenges there is a need to find new market niches where telecom operators would have a competitive advantage over global OTT players.

Revolutionary technological changes in the automotive industry, as well as changes in traffic infrastructure, and a large paradigm shift in people's mobility, imposed the market niches of ITS as areas with great potential for business growth for telecom operators.

Traffic and transportation have important local characteristics and thus a specific need for services that

give local telecom operators an advantage over global OTT players.

Smart cars and smart mobility industries provide lucrative new jobs for telecom companies, and these are markets that are undoubtedly on the rise. There are a lot of potential business opportunities for telecommunications companies in the ITS area. This paper will describe only a few that are coming very soon according to market trends.

Automotive manufacturers have less and less profit with their core products, hence they are looking for opportunities for growth in new "after-market" products and services.



Figure 6: Verizon ITS Platform : Smart IoT Vehicle - ITS Platform Makes Every Car Smarter

These "after-market" products and services offer On Board Diagnostics solutions that include in-car wireless connectivity and data transmission over the cars' Controller Area Network. This transmission is achieved via a standard protocol that allows data collection from sensors within the car and their subsequent communication to the telecom network. The collected vehicle parameters could be used for insurance purposes, or for driver or vehicle health monitoring or changing eco or urban characteristics. One such example is to automatically book in a service, another is to monitor the driver's vital signs for signs of fatigue or poor health.

The connected car concept is a market opportunity, but requires collaboration between the automotive companies and the telecom operators offering huge potential for both sides. For connected car analytics automotive companies are required to enable telecom companies to gain access to in-car data. If data transmission is for insurance purposes, or spare part replacement, the system can automatically send information for order and installation of a new part when it detects an impending fault. Another opportunity with huge potential is in e-coaching of drivers. This service analyses and provides customized driver coaching for inexperienced drivers or can be used by emergency services or can improve driving skills for someone that has been caught speeding. The potential for collaboration between equipment manufacturers and automotive companies follows logically since the number of phones outnumbers that of cars. This means that offering such services would greatly amortize the costs if offered by operators rather than by individual car manufacturers.

Another telecom opportunity is the ITS area for "management of the safety of passengers and drivers" that is denoted as 'ITS wearables', which are accessories that can be offered such as safety devices. These devices could include for example an emergency switch for the passenger sitting in the back seat. Operators are well suited to exploit this opportunity since it is not only the service but the value-added component that can be offered with it.

If operators that have been the first to seize the opportunity offered by connected cars are to be mentioned, they would include the following: Verizon that acquired Hughes Telematics in order to offer a range of individual and fleet management services for safety, infotainment and efficiency, and Deutsche Telekom cooperating with Daimler to offer cloud infrastructure, services and connectivity from a single source. At least a dozen of the leading global telecom companies have set up divisions and formed partnerships to participate in the emerging connected car and ITS market.

Since market growth is up-coming, it is expected that competition will be tough and companies such as Google, Apple and Amazon will be formidable telecom opponents. Telecom companies will gain a lot from winning a share of this lucrative new market. Cellular 4G/5G networks have the advantage of being better communications medium than WiFi because they offer high availability and since 24/7 availability is needed, there is an important opportunity for operators to seize. However, operators should avoid the concept of just selling the data and the pipe itself. They should focus on adding real value to this solution. One such way is to offer analytics and intelligence to ensure that the traffic sent through the cloud is handled in an optimized and intelligent manner. For example, instead of streaming all data which would create a large data overload, it is better to identify that a change has taken place and only then transmit this data.

In general, the ITS opportunity for telecom operators is huge. Operators like AT&T expect to earn $5 billion of revenue from the connected car market by 2020. Nowadays only 15 % of new cars are sold with on-board connectivity but this number will rise to 35% by 2017 and by 2018 this will become the norm for all new cars.

VIII. CONCLUSION

In the C-ITS eco-system vehicles occupy the most important place, but have significantly different features and capabilities compared to today's conventional vehicles. Networked smart vehicles equipped with V2V systems have the ability to collect information on the move, store data, perform mathematical operations, communicate with Cloud applications, provide automated data transfer, geo location, and a system for locating surrounding objects. Such large number of activities and information processing sets high requirements on V2V communications with significantly reduced latency with respect to 4G, which leads to a challenging 1-5 ms end-to-end delay. Additionally, the requirements of reliability require from V2V systems access in a variety of cases, including out-of-coverage scenarios, in which one or both devices in communication cannot get network connectivity. C-ITS systems also require integration with various sensor systems or IoT solutions. All these

demands create great challenges for the mobile communications networks, which in addition must achieve profit which in turn means they have to optimizing capacity to the maximum. The current trend of solving these challenges is in the direction of implementation of SDN and NFV technologies, but a complete solution for a large number of mobile smart objects (IoT users) that fulfill the requirements for flexibility, compatibility, security, speed and expandability can be achieved, in the long-term, only by implementing 5G architecture.

LITERATURE

[1] Mandzuka, Sadko; Zura, Marijan; Horvat, Bozica; Bicanic, Davor; Mitsakis, Evangelos, „Directives of the European Union on Intelligent Transport Systems and their Impact on the Republic of Croatia“, Promet - Traffic & Transportation. 25 (2013) , 3; 273-283.

[2] Mandzuka, Sadko; Kljaic, Zdenko; Skorput, Pero. „The Use of Mobile Communication in Traffic Incident Management Process“, Journal of Green Engineering. 1 (2011) , 4; 35-42.

[3] Mandzuka, Sadko; Huzjan, Boris; Kos, Goran. „Cooperative Highway Traffic Safety Management Systems“, Recent Advances in System / Prof. Valeri Mladenov, Technical University of Sofia, Bulgaria (ur.). Zakynthos Island, Greece : Recent Advances in Electrical Engineering Series, 2015. 661-666.

[4] Ericsson, „Architecture Evolution for Automation and Network Programmability“, Ericsson Review, 2014, available at: http://www.ericsson.com/news/141128-er-architecture-evolution_244099435_c.

[5] Ericsson, „Mobility Report“, June 2015, available at: http://www.ericsson.com/mobility-report.

[6] Carlos Bravo, Francesco Caruso, Christian Olrog, Malgorzata Svensson, András Valkó; „Gearing up support systems for software defined and virtualized networks“, Ericsson Review June 2015.

[7] Karthikeyan Natarajan, head of the Engineering and R&D group at Tech Mahindra, „Four ways telcos can profit from the growth of the connected car market“, telecoms.com - news portal covering the latest technological advancements and market trends, focussing on the key business and technology issues facing the industry today, September 2015.

[8] ETSI, „Network Functions Virtualisation (NFV); Architectural Framework“, Group Specification, 2014, available at: http://www.etsi.org/deliver/etsi_gs/NFV/001_099/002/01.02.01_60/gs_NFV002v010201p.pdf.

[9] Ericsson, „Software-defined networking: the service provider perspective“, Ericsson Review, 2014, available at: http://www.ericsson.com/news/130221-software-definednetworking-the-service-providerperspective_244129229_c.

[10] „Here and the Netherlands evaluating 4G/LTE-based C-ITS road messaging system“, Veldhoven, Netherlands; November 2015

[11] ETSI - European Telecommunications Standards Institute, „Cooperative ITS“

[12] European Commission, DG Mobility and Transport, B – 1049 Brussels; „Connected and automated driving (C-ITS)“, January 2016

[13] European Committee for Standardization, „CEN and ETSI deliver first set of standards for Cooperative Intelligent Transport Systems (C-ITS): A key step towards connected cars in Europe“; February 2014.

[14] Ling Sun, Yameng Li, Jian Gao; “Architecture and Application Research of Cooperative Intelligent Transport Systems”, Procedia Engineering 137 (2016) 747 – 753.

[15] D. Perakovic, Z. Kljaic, S. Steficar, „Smart automated platform for telecom operators and enterprise users in industry, transport and energy“, Conference: MIPRO, 2012 Proceedings of the 35th International Convention, January 2012.

[16] Ertico-ITS Europe, European Commission,5GPPP; “5G Automotive Vision”, October 2015.

[17] Ericsson White paper, “5G systems - Enabling Industry And Society Transformation”, Uen 284 23-3244 | January 2015

[18] Vensi Inc., „The Mobile App That Allows You to Monitor Your Car“, February 2016.

[19] Erik G. Ström, Petar Popovski, Joachim Sachs; „5G Ultra-Reliable Vehicular Communication“, Cornell University Library, October 2015. arXiv:1510.01288

[20] Global mobile Suppliers Association; “The Road to 5G: Drivers, Applications, Requirements and Technical Development”, November 2015.

[21] Andreas Festag, “Cooperative Intelligent Transport Systems Standards in Europe”, IEEE Communications Magazine, Volume:52 , Issue: 12, December 2014.

[22] David Soldani, Antonio Manzalini, „Horizon 2020 and Beyond: On the 5G Operating System for a True Digital Society“, IEEE Vehicular Technology Mag., Volume:10 Issue:1, March 2015.

[23] Shahid Mumtaz, Kazi Mohammed Saidul Huq, Muhammad Ikram Ashraf, Jonathan Rodriguez, Valdemar Monteiro, and Christos Politis; „Cognitive Vehicular Communication for 5G“, IEEE Communications Magazine, Volume:53 , Issue: 7. July 2015.

NETWORKS PERFORMANCES

Performance Evaluation of Different Scheduling Algorithms in LTE Systems

A. Marinčić , D. Šimunić
University of Zagreb, Faculty of Electrical Engineering and Computing,
Department of Wireless Communications, Green Engineering Laboratory
Unska 3, 10000 Zagreb, Croatia

Abstract – Nowadays, LTE is widely used mobile communication technology. The development of mobile communications has never been faster. The main goal of mobile communications is to increase the data throughput and to ensure the best possible quality of communication between users, regardless of environment in which they are situated and type of data they share. In this paper, performance evaluation of different scheduling algorithms used in LTE systems is given. The following scheduling algorithms are considered: Round Robin, Proportional Fair, Best CQI, Resource Fair and MaxMin. Fractional Frequency Reuse (FFR) method, which represents a new way of resource allocation in LTE systems, is also evaluated in this paper. System performance was observed according to the following parameters: average throughput of user equipment, peak throughput of user equipment, edge throughput of user equipment and fairness in radio resources allocation. Vienna LTE System Level Simulator was used for conducting evaluation of LTE system.

I. INTRODUCTION

Development of mobile communications requires ever higher transmission rates and emerging various multimedia services. The number of applications and services, which require a great number of transmitted data (e.g. searching Internet or using social networks), has been rising. As a result, mobile communication networks of new generations have been developed. New technologies and systems should provide transmission rates and network capacity that is necessary for these kind of multimedia services. In this paper, performance evaluation of Long Term Evolution (LTE) system based on different scheduling algorithms is presented. Due to the need of efficient usage of scarce resources, scheduling algorithms become more important area of interest in mobile communication. It is important to use scarce resources effectively in order to increase network capacity and preserve needed quality of data transmission. In this paper following scheduling algorithms were evaluated: Round Robin, Proportional Fair, Best CQI, Resource Fair and MaxMin. Also, a Fractional Frequency Reuse (FFR) method, which represents a new way of resource allocation in LTE systems is evaluated in this paper. Fairness and throughput are used as parameters for evaluation of scheduling algorithms.

II. VIENNA LTE SYSTEM LEVEL SIMULATOR

In this section Vienna LTE System Level Simulator [1], which is developed at Vienna University of Technology, is described. This simulator was used for evaluation of scheduling algorithms in this paper. The goal of Vienna LTE System Level Simulator is to enable analysis of network performances. In LTE technology this network consists of big number of base stations (eNodeB), which are covering specified area. That area is called region of interest (ROI). In Figure 1 eNodeB and users layout is depicted.

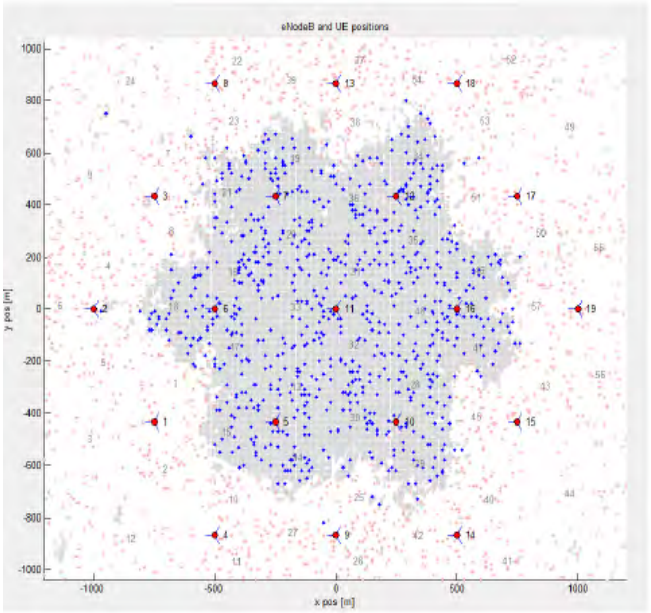


Figure 1. eNodeB and users layout

Vienna LTE System Level Simulator consists of two parts [1]:

- Link measurement model
- Link performance model

Link measurement model is in charge of measuring link parameters. This model demonstrates link quality based on measurements conducted using user's equipment, which are later sent to base stations in form of measurement report. This part of simulator is in charge of link adaptation and resource allocation. Based on the link measurement model, the link performance model predicts the block error rate (BLER) of the link, based on the receiver signal to interference ratio (SINR) and the transmission parameters (e.g., modulation and coding).

In Figure 2 schematic block diagram of the LTE system level simulator is given.

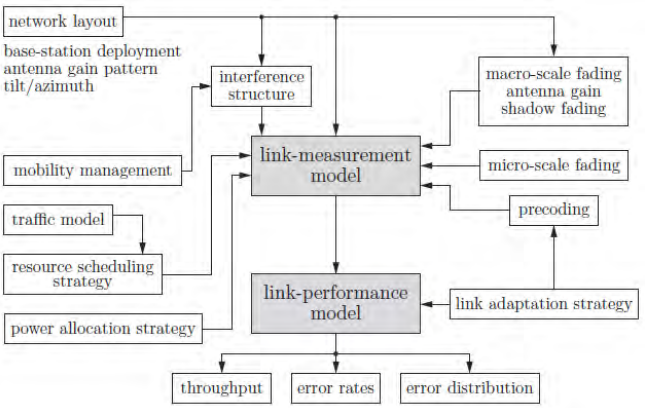


Figure 2. Schematic block diagram of LTE system level simulator [1]

III. SCHEDULING ALGORITHMS IN LTE NETWORKS

Scheduler is a part of Medium Access Control (MAC)[2] layer and it allocates resources to users. A group of resource blocks and associated transmission format are allocated to the user equipment. Transmission format represents transmission block size, used modulation scheme and coding type. Resource block is a unit of transmission resource. It consists of 12 subcarriers in the frequency domain and 1 time slot (0.5ms) in the time domain. The bandwidth of resource block is 180 kHz [2]. Scheduler can use following algorithms for allocating resources.

A. Round Robin Algorithm

Round Robin [3] algorithm represents the simplest way resource allocation. This algorithm allocates resources cyclically. Scheduler generates a list of active users, who are randomly sorted. After list of user's equipment is generated, resource blocks are allocated to users in order in which they are sorted in list. After all users have been assigned, the resource allocation starts from the beginning of the list. While allocating resources Round Robin algorithm doesn't consider channel conditions. It can be thought of as one of the fairest algorithms.

B. Best CQI Algorithm

In contrast to Round Robin algorithm, while allocating resources Best CQI algorithm takes into account current state of channel. Current channel state is expressed by channel quality indicator (CQI) parameter. Better channel state is described with higher values of CQI. Best CQI algorithm assigns resources to users with highest CQI value, which means that user is located in part of channel with best conditions. Due to that users can achieve high throughput. When resources are allocated using Best CQI algorithm high capacity of the whole system can be accomplished. Best CQI algorithm can be expressed as [3]:

$$k = \arg\max_i R_i \tag{2.1}$$

where R_i is current transmission rate for user i.

C. Proportional Fair Algorithm

Before mentioned algorithms provide high throughput with low fairness (Best CQI) or low throughput with high fairness (Round Robin). Due to that, need of new algorithm which will be a combination of these two algorithms emerged. A Proportional Fair algorithm was developed and today it is one of the mostly used algorithms. The goal of this algorithm is to enable adequate throughput and maintain fairness. Proportional Fair algorithm assigns resources to users, who have relatively the best channel conditions. The resources are assigned to a user with the relative best channel condition. Resources are assigned according to following expression [3]:

$$k = \arg\max_i \frac{R_i}{\overline{R_i}} \tag{2.2}$$

where R_i is current transmission rate for user i and R_i is average data transmission rate for user i.

D. Resource Fair Algorithm

Similar to Proportional Fair Algorithm, another algorithm called Resource Fair algorithm puts emphasis on fairness in resource allocation. This algorithm is also based on sum rate maximizing approach. In order to fulfil the goal of Resource Fair algorithm, which is to enable sum throughput maximization with satisfactory fairness, following condition needs to be accomplished [4]:

$$\|b_k\|_1 = \frac{N}{K} \forall k \tag{2.3}$$

where $||b_k||_1$ is number of resource blocks allocated to user k. This condition is used in order to ensure every user to get the equal number of resource blocks. From this condition it can be seen that the number of allocated resource block to user k is equal to ratio of total number of resource blocks N and total number of users K.

E. MaxMin Algorithm

Last of the considered algorithms, also based on the sum throughput maximizing, is MaxMin algorithm. The goal of this algorithm is to maximize minimum user's throughput. By using this algorithm, when throughput of one user is increased, throughput of other user with lower throughput of first user lowers.

F. Scheduling algorithms comparison results

In this section comparison of different scheduling algorithms is given in order to evaluate LTE system. Depending on scheduling algorithms and other parameters, the results of LTE system performances are given. Average, peak, and edge throughput results of user equipment are given. Also, average cell throughput and fairness comparison results are given.

Table 1 presents simulation parameters of LTE system evaluation when using different schedulers.

Table 1. Simulation parameters of LTE system based on used scheduler

System frequency	2 GHz
Bandwidth	20 MHz (100 resource blocks)
Transmission mode	SISO
Simulation time	10 TTI
Distance between base stations	500 m
Number of base stations	57
Number of users for every base station	20
Transmitter power	49 dBm
Network geometry	Hexagonal geometry
User equipment speed	5 km/h
User equipment distribution	Constant distribution
Scheduler	Round Robin, Best CQI, Proportional Fair, Resource Fair, MaxMin

In the first case, average, peak and edge user equipment throughputs were considered. Figure 3 gives average throughput of user equipment depending on different scheduling algorithms.

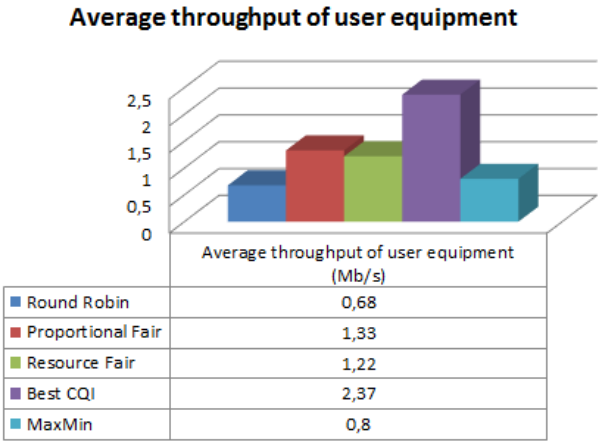


Figure 3. Comparison of average throughput of user equipment based on used scheduler

Average user throughput is defined as the amount of data sent successfully for user over a period of time (i.e., one TTI). Results on Figure 3 show that the highest average throughput of user equipment is achieved with the Best CQI algorithm. This result is expected, because with usage of this algorithm resources are allocated to user with the best channel state. Those users located in part of channel with the best conditions can achieve highest throughput. Resource Fair and Proportional Fair algorithms have somewhat worst performances than Best CQI algorithm. This is due to their requirement that high fairness needs to be achieved, and because of that lower throughput is achieved. The worst performances are achieved with MaxMin and Round Robin algorithm. It is not possible to achieve high performances with Round Robin algorithm, because with this algorithm channel state is not taken into consideration. MaxMin algorithm can't achieve high throughput because the main goal of this algorithm is to maximize minimum throughput of user equipment, which means that it maximizes edge throughput. In Figure 4 peak throughput of user equipment is given based on used scheduling algorithms. Peak user throughput is defined as maximum achievable throughput per user.

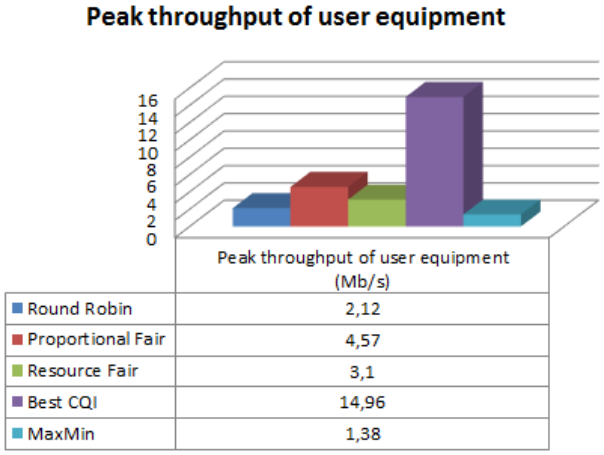


Figure 4. Comparison of peak throughput of user equipment based on used scheduler

Figure 4 shows that the peak throughput of user's equipment complies with results of average throughput of user equipment. Best CQI algorithm has the highest peak throughput. Figure 5 shows edge throughput of user equipment based on used scheduler.

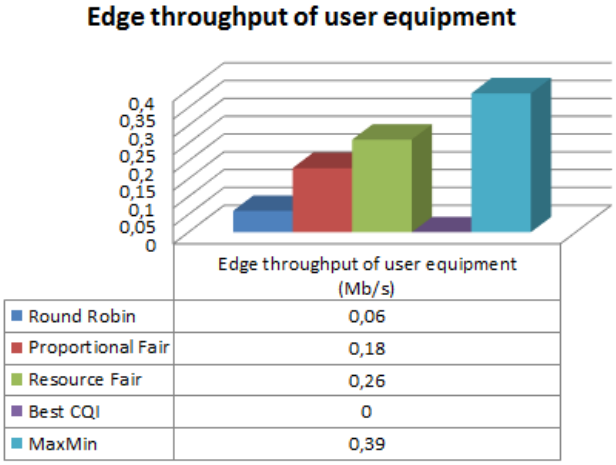


Figure 5. Comparison of edge throughput of user equipment based on used scheduler

Edge throughput is defined as the 5% point of cumulative distribution function (CDF) of the user throughput normalized with the overall cell bandwidth. Figure 5 shows that the highest edge throughput is achieved with MaxMin algorithm. This is due to algorithm design, which enables schedulers to allocate resources to users with

poor channel quality. The worst performances are achieved with Round Robin and Best CQI algorithm due to design of these algorithms. Figure 6 gives comparison of fairness results for different scheduling algorithms.

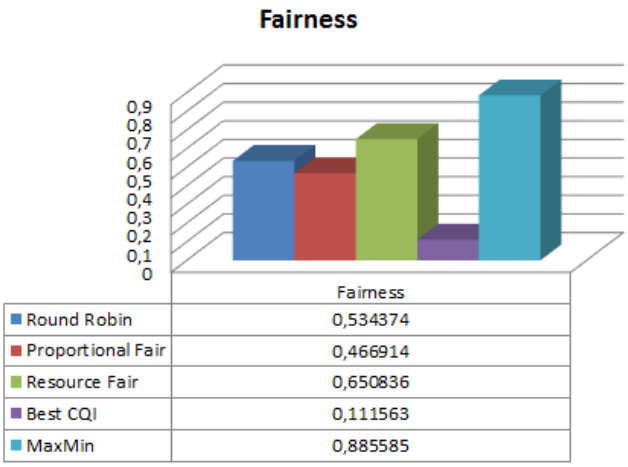


Figure 6. Comparison of fairness based on used scheduler

Fairness is defined as a metrics which determines if resources are shared uniformly between users. Fairness is calculated according to the following expression [4]:

$$J(\mathbf{x}) = \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N\sum_{i=1}^{N} x_i^2} \qquad (2.4)$$

where $\mathbf{x}$ is a vector of length N (number of users), which contains fairness value for every user (x_i). From Figure 6 it can be seen that the best performances are achieved by MaxMin algorithm. This algorithm is the fairest algorithm because its goal is to maximize edge throughput of different user equipment. From (2.4) expression follows that the fairness is higher when the throughput of different schedulers is similar. The worst fairness is achieved by Best CQI algorithm due to the fact that this algorithm allocates resources to users with the best channel state. Because of that, there is a big difference between throughputs of different user equipment.

IV. FRACTIONAL FREQUENCY REUSE METHOD

A. *Fractional Frequency Reuse scheduler*

Fractional Frequency Reuse scheduler is implemented in Vienna LTE System Level Simulator as a new way of resource allocation.

This scheduler divides cell on two parts [5]:

- Full Reuse (FR), which is a central part of cell.
- Partial Reuse (PR), which is edge part of cell.

In central part of cell, interference is not expressed and it doesn't have a great impact on users. In this part of cell only one frequency is used (reuse-1) [5]. In edge part of the cell higher factor of frequency reuse is used. In our case, edge part of the cell uses three frequencies (reuse-3) [5]. With this kind of frequency reuse, the goal is to lower interference and increase cell capacity. That can be obtained by using one frequency in central part of cell and a bigger number of frequencies in edge part of the cell. This results in maintaining of high spectral efficiency of central part of cell and increasing spectral efficiency of cell edge. Cell division is realised by allocating β_{FR} number of slots to FR cell part, and 1- β_{FR} slots to PR cell part [5].

B. *FFR method results*

This section gives analysis of performance of LTE system based on usage of FFR scheduler. The goal is to show the improvements that can be achieved using FFR scheduler. Two parameters that have a great impact on results which are obtained with FFR scheduler are bandwidth ratio β_{FR} and SINR value Γ_{thr}. SINR value defines FR and PR boarder. In this work, FFR scheduler performance were observed with β_{FR} =0.01 and Γ_{thr} = 15dB. Since 20 MHz bandwidth was used (100 resource blocks), it can be concluded that one resource block is allocated to FR part and 33 resource blocks are allocated to every PR part. Used simulation parameters are the same as in Table 1. In the first case, comparison of Round Robin algorithm with and without FFR is given. Results for that comparison are shown on Figure 7 to Figure 10. In the second case, comparison of Proportional Fair algorithm with FFR and without FFR is given. Results are shown on Figure 11 to Figure 14. Also, a comparison of Round Robin with FFR and Proportional Fair with FFR is described.

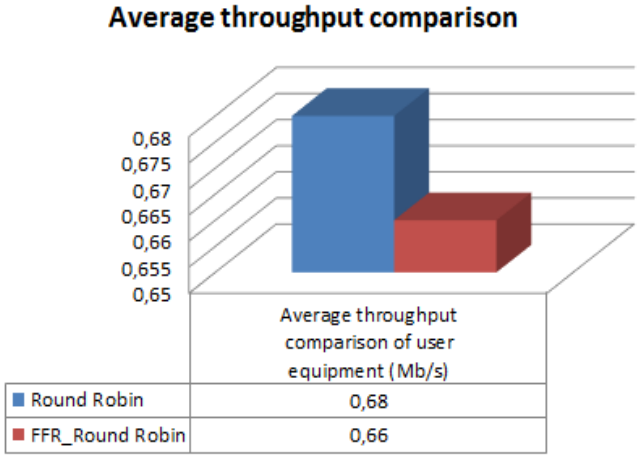


Figure 7. Average throughput comparison

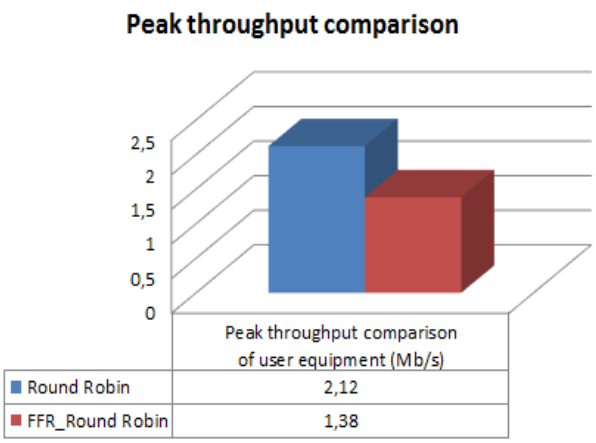


Figure 8.Peak throughput comparison

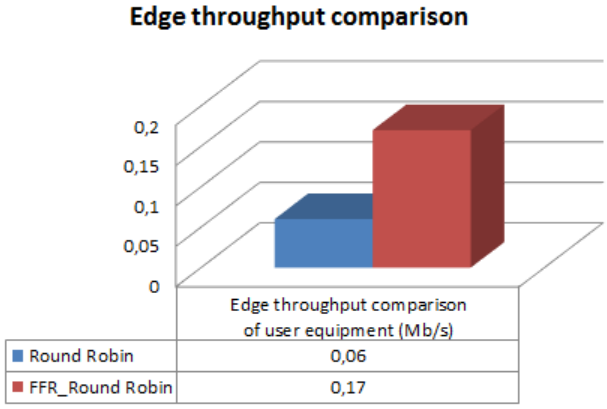


Figure 9. Edge throughput comparison

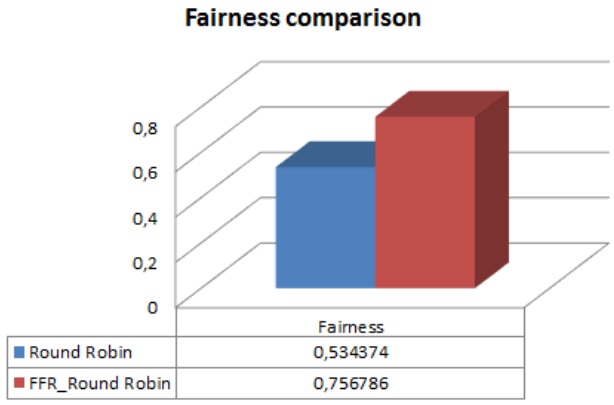


Figure 10. Fairness comparison

Based on obtained results, it can be seen that by using FFR method of resource allocation fairness is improved with minimal throughput decreasing. From Figure 10 it can be seen that by using FFR fairness is improved by 42%. While fairness is improved, peak throughput is lowered by 35%, average throughput is lowered by 3% and edge throughput is higher for 183%.

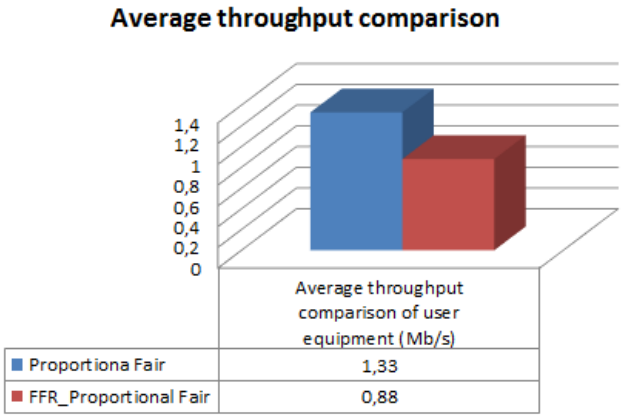


Figure 11. Average throughput comparison

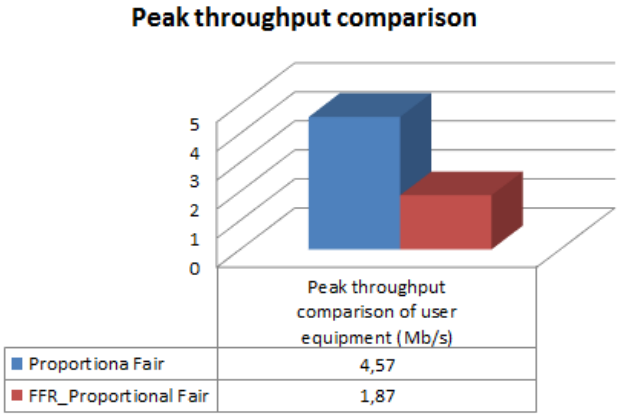


Figure 12. Peak throughput comparison

In second case Proportional Fair scheduler with use of FFR method was studied. From Figure 14 it can be seen that by using FFR method fairness is improved for 50%. In this case average and peak throughput are also lowered. From Figure 13 it can be seen that average throughput is lowered for 34% and peak throughput is lowered for 59%. On the other side, edge throughput is increased for 28%.

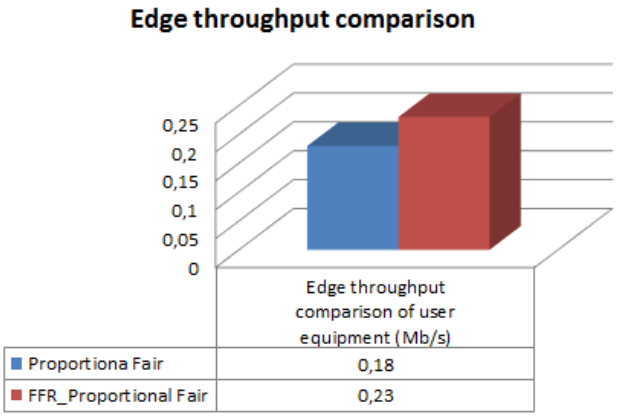


Figure 13. Edge throughput comparison

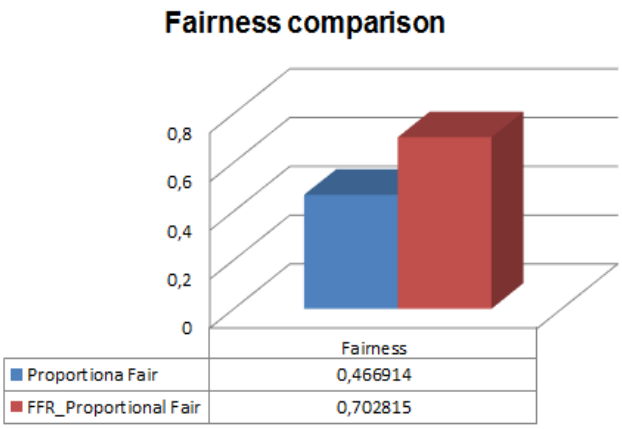


Figure 14. Fairness comparison

Simulation results show that better improvements are gained by using Round Robin algorithm and FFR than by using Proportional Fair algorithm with FFR, although fairness was improved for a larger percentage in the second case. Also, higher losses are achieved in the second case for average and peak throughput. On the other side, if we compare performances of Round Robin with FFR and Proportional Fair with FFR, better LTE system performances are achieved using Proportional Fair algorithm with FFR.

V. CONCLUSION

In this paper we evaluated LTE systems regarding the used scheduling algorithms. Vienna LTE-A System Level Simulator was used for evaluation. The following algorithms were considered: Round Robin, Best CQI, Proportional Fair, Resource Fair and MaxMin. Also, FFR method, as a new method in resource allocation was examined. The following parameters were considered: average throughput of user equipment, peak throughput of user equipment, edge throughput of user equipment, spectrum efficiency of user equipment, average cell throughput and fairness in radio resources allocation.

Results show that Best CQI algorithm has the best performance in a case of average and peak throughput of user's equipment and about average throughput of cell and average spectral efficiency. On the other side, Best CQI algorithm has the worst results considering fairness and edge throughput of user's equipment. The highest fairness and edge user's equipment throughput is gained by MaxMin algorithm.

Fractional Frequency Method was also examined in this paper. This is a new method in the resource allocation. Results show that improvements of throughput and fairness can be gained by using this method with Round Robin algorithm. However, if the results are compared with results of throughput and fairness for Proportional Fair algorithm, then the last scheduling algorithm provides better results.

REFERENCES

[1] Mehlführeh Ch., Ikuno J.C., Šimko M., Schwarz S., Wrulich M., Rupp M. „The Vienna LTE Simulators - Enabling reproducibility in wireless communications research.“ EURASIP Journal on Advances in Signal Processing, 2011., 2011:29, str. 1-14.

[2] Geßner, Ch. „Long Term Evolution: A concise introduction to LTE and its measurement requirements. Second Edition.“ Mühldorfstraße 15, 817671 Munich, Germany : Rohde & Schwarz GmbH&Co, KG, 2011.

[3] Dahlman E., Parkvall S., Sköld J. „4G: LTE/LTE-Advanced for Mobile Broadband. Second Edition.“ The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom: Elsevier, 2014.

[4] Schwarz S., Mehlfuhrer Ch., Rupp M. „Low Complexity Approximate Maximum Throughput Scheduling for LTE.“ Conference Record of the Forty-fourth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, Nov 2010., str. 1563-1569.

[5] Ikuno J.C., Taranetz M., Rupp M. „A Fairness-based Performance Evaluation of Fractional Frequency reuse in LTE.“ 17th International ITG Workshop on Smart Antennas (WSA)., Stuttgart, Germany, March 2013., str.1-6.

[6] Sun Z., Yin C., Yue G. „Reduced- Complexity Proportional Fair Scheduling for OFDMA System.“ IEEE International Conference on Communications, Circuits and Systems. vol 2., Guilin, China,June 2006, str. 1221-1225.

[7] Caban S., Mehlführer Ch., Rupp M., Wrulich M. „Evaluation of HSDPA and LTE: From Testbed Measurements to System Level Performance.“ The Atrium, Southern Gate, Chichester,West Sussex, PO19 8SQ, United Kingdom: John Wiley & Sons Ltd, 2012.

[8] Ikuno J.C., Wrulich M., Rupp M. „System Level Simulation of LTE Networks.“ IEEE 71st Vehicular Technology Conference (VTC 2010-Spring),Taipei, Taiwan, May 2010,str 1-5.

Performance Analysis of LTE Networks with Random Linear Network Coding

Tewelde Degefa Assefa, Katina Kralevska and Yuming Jiang
Department of Telematics
Norwegian University of Science and Technology, Trondheim, Norway,
Email: tewetgrt@gmail.com, katinak@item.ntnu.no, jiang@item.ntnu.no

***Abstract*—Random Linear Network coding (RLNC) has emerged as a promising solution for reliable multimedia delivery over mobile cellular networks. In this paper, we deploy Application Layer-RLNC (AL-RLNC) on the top of the existing Hybrid Automatic Repeat Request (HARQ) in 4G Long Term Evolution (LTE) networks. A simple implementation scenario composed of a user equipment, an eNB and a remote host is considered. Our results show that AL-RLNC improves the throughput and the coverage at a cost of a higher packet delay. In addition, we compare the performance when AL-RLNC is used to the performance with the advanced LTE system antenna technique Multiple-Input Multiple-Output (MIMO).**

Keywords: RLNC, AL-RLNC, LTE, MIMO, HARQ

I. Introduction

The number of mobile users and advanced multimedia delivery services offered over mobile cellular networks have increased rapidly. It is expected that the total worldwide mobile traffic will reach more than 127 Exabytes in 2020, representing 33 times increase compared to 2010 [3]. Thus, there is an inevitable need for high capacity networks.

Long Term Evolution (LTE) is a standard that provides high-speed data for mobile phones and data terminals. The key features of LTE that enable to meet the strict Quality of Service (QoS) demands are: scalable carrier bandwidths (from 1.4 MHz to 20 MHz), support both Frequency Division Duplexing (FDD) and Time Division Duplexing (TDD), Orthogonal Frequency Division Multiplexing (OFDM) in combination with higher order modulation techniques and Adaptive Modulation and Coding (AMC).

Still novel solutions for reliable and fast delivery of mutimedia are needed. Random Linear Network Coding (RLNC) has become a promising approach for improvement of the network throughput, efficiency and scalability [4], [7]. In RLNC [6], a coded packet y_j is a linear combination from input packets x_i, multiplied by coefficients $c_{j,i}$ from a Galois Field of size q, $GF(q)$, where $i = 1, \ldots, k$ and $j = 1, \ldots, n$. As the number of x_i packets or the size q of the Galois Field increases, the probability that each newly generated packet y_j is linearly independent is higher. Given the unique flexibility of RLNC to efficiently bridge the upper layer content packetization and the lower layer packet transmission, RLNC is considered as a powerful cross layer solution for a reliable mutimedia delivery over the LTE/LTE Advanced networks. Recently, there have been theoretical and practical evaluations on RLNC integration in LTE networks such as the work presented in [5], [9], [12], [14], [15]. The authors in [14], [15] proposed a modified LTE protocol stack by using RLNC on the Media Access Control (MAC) layer as a replacement for Hybrid Automatic Retransmission reQuest (HARQ) in LTE/LTE-A networks. The performance of RLNC by implementing an inter-flow and intra-flow MAC-RLNC for a single-user and multi-users scenarios is evaluated in [5]. Performance analysis and energy efficiency for integration of RLNC in the MAC layer is presented in [8]. Practical implementations of RLNC at the application layer for multimedia streaming on Apple iPhone platform have been presented in [11], [13].

A. Our Contribution

This paper analyzes the performance of LTE networks with intra-flow RLNC over unicast flows. Only packets within one flow are coded when intra-flow RLNC is used [2]. We implement RLNC at the application layer as a complement of the HARQ mechanism. To the best of our knowledge, this is the first work that uses the Kodo network coding library [10] to examine the performance of LTE networks with AL-RLNC. At the time of conducting this work, Kodo library functionalities are implemented in the transport layer. In comparison to MAC-RLNC, AL-RLNC provides a simple integration of RLNC into LTE networks without any modification to the LTE protocol stack. In addition, it turns out that deployment of network coding is less complex with AL-RLNC in terms of compatibility with current devices and applications. Moreover, MAC-RLNC requires packets to be coded at the MAC layer which Kodo library does not support.

We analyze the throughput and delay performance which are the key design parameters of any system subject to strict QoS requirements. The results show that RLNC improves the throughput at the cost of a higher packet delay. In addition, we compare the performance when Single-Input Single-Output (SISO) and Multiple-Input Multiple-Output (MIMO) antenna techniques are used.

The paper is organized as follows. In Section II, we present the background of AL-RLNC. Section III describes the simulation setup and the system flow. In Section IV we discuss the results obtained from the simulations. Conclusions and future work are given in Section V.

II. Background

When LTE networks support AL-RLNC, the standard LTE protocol layer stack remains unchanged as shown in figure 1. AL-RLNC is deployed on top of the existing MAC layer HARQ based packet transmission process. In this solution, RLNC encoded IP packets enters the eNB Packet Data Conversion Protocol (PDCP) layer. PDCP layer performs header compression and ciphering then the PDCP encapsulated IP packets are delivered to the RLC layer. The RLC layer performs segmentation/ concatenation of IP packets into RLC packets to fit the MAC frame size requirements. Each MAC frame is allocated a single physical layer transport block for transmission over the eNB/UE interface. The physical layer carries all the information from the MAC transport channels over the air interface. In addition, it has link adaptation functionality that provides a matching of the modulation and the coding techniques to the radio link interface condition [8]. In this work packets are coded at the application layer by using the RLNC funcionalites from the Kodo network coding library.

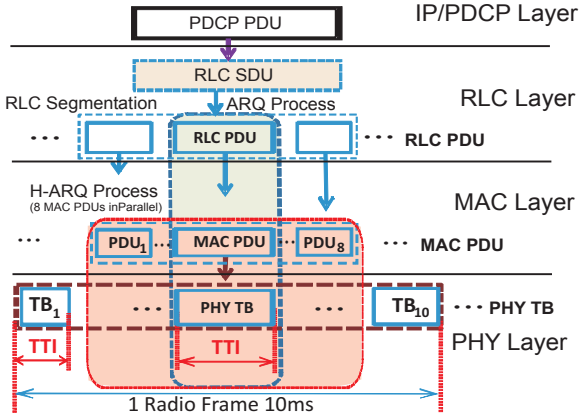


Fig. 1. eNB E-UTRAN protocols: MAC-HARQ solution.
Fig. 1. eNB RAN protocol, MAC-HARQ solution, AL-RLNC is deployed on top of the standard MAC-HARQ solution [8]

We have decided to implement AL-RLNC solution due to two reasons. First, the Kodo library is applicable only to the upper layers and it is not designed to use the encoding and decoding functionalities of the library in the lower layers. Second, from an implementation point of view, deploying the RLNC in the application layer allows to simply integrate the Kodo library coding scheme on top of the current stack and without affecting the functionality of the LTE protocol stack and/or compatibility of UEs and eNBs.

III. AL-RLNC

In order to deploy RLNC in LTE networks, we used ns-3[1] and Kodo[2] library.

The appropriate size RLC PDUs are forwarded to the MAC layer which transmits groups of 8 MAC PDUs using 8 parallel HARQ processes within blocks of 8 consecutive TTIs (Fig. 1). If any of the MAC PDUs is not received correctly, its new incremental redundancy (IR-HARQ) based version is transmitted within the same slot of the following TTI octet. The TTI octet period (8 ms) between retransmissions is sufficient for ACK/NACK feedback reception and generation of a new IR-HARQ-based MAC PDU description, if needed. The maximum number of retransmissions is three.

Finally, each MAC PDU is allocated a single PHY TB trans-

[1] A C++ based open source simulation library for networking research. We used ns-3 version 3.22

[2] A C++ library for implementing random linear network coding. We used Kodo version 19.0.0



Fig. 2. LTE-EPC simulation topology. The remote host sends coded packets to UE through the SGW/PDN-GW

A. Topology and Architecture

In ns-3, only FDD for LTE is supported and all the obtained results in this work are based on FDD operation mode.

The QoS aware scheduler is a key component in LTE for the achievement of a fast adjusted and efficiently utilized radio resource. In this experiment we have used a QoS aware scheduler called Priority Set Scheduler (PSS) from the ns-3 module. During the TTI duration, i.e. 1 ms, the UEs report their perceived radio quality as an input to the eNB scheduler to decide which AMC should be used. The scheduler prioritizes the QoS requirements amongst the UEs. Then it informs the UEs of the allocated radio resources both in downlink and uplink direction.

Figure 2 shows the topology of the LTE network used for the performance analysis. The topology consists of a remote host located in an external network, SGW/PDN-GW as one entity, eNB and a UE. Two applications are implemented at the remote host and the UE for encoding and decoding packets, respectively. On each of these applications a UDP socket is created to transmit and receive the encoded IP packets over the LTE network. The coded packets are generated from a large file with a specified generation size (number of source packets), packet size and coding over a specific Galois field using the RLNC functionality in the Kodo library. We use this simple topology because the LTE-EPC simulation module in ns-3 supports only a point-to-point connection between remote host and UEs/eNBs located in different networks, i.e., it does not support broadcasting.

The architecture in Figure 3 shows how the encoded packets are transmitted from the remote host and decoded at the receiver, UE. As shown in Figure 3, a packet to the SGW/PDN-GW node which is connected to the Internet by the SGi interface is forwarded by Internet routing. SGW/PDN-GW determines the eNB to which the UE is connected to by looking at the UE IP destination address. It classifies the packet by using traffic flow templates to identify to which Evolved packet system (EPS) bearer it belongs. Each EPS bearer has a one-to-one mapping to the S1-U interface. Following the mapping it sends out the coded packet to the intended eNB via the S1-U interface. Upon a reception of the coded packet, the eNB forwards it to the UE over the LTE-Uu interface based on the bearer ID. Finally the UE receives the encoded packet. The application installed on the UE checks whether the received packet is linearly independent. It drops the linearly dependent packets and stores the linearly independent packets in a matrix form for decoding. The remote host (encoder)

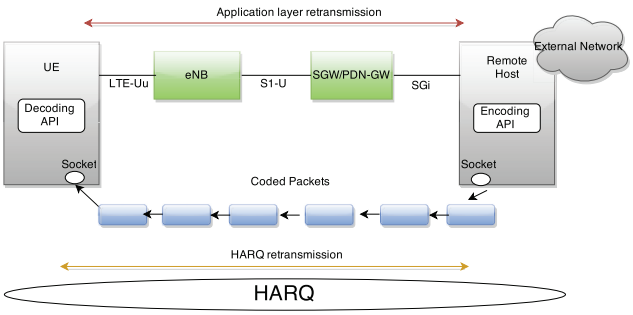


Fig. 3. Architecture showing how coded packets are sent down to the UE using RLNC

is transmitting encoded packets until it gets a notification from the UE (decoder) that signals the reception of sufficient linearly independent packets to recover the whole file. In this process the HARQ performs its functions independently from the application layer, i.e., it retransmites a coded packet after unsuccessful reception.

Figure 4 gives an overview of the steps for creating the simulation architecture.

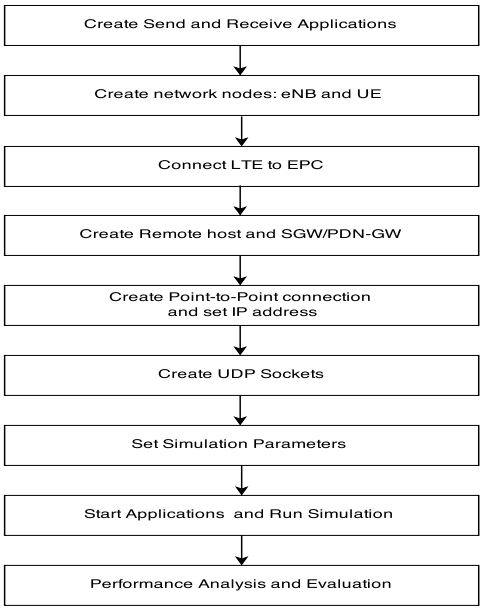


Fig. 4. Overall system flow for setting up the simulation

B. Simulation Setup and System Flow

We use the simulation parameters shown in Table I in order to simulate and collect data for the performance analysis and evaluation of the defined network topology. As shown in the table an EPS bearer is established to classify the packets according to TFTs and the defined QoS class, as they cross the EPC core towards the UE in the LTE access network. The simulation time is 20 seconds. The run-times

TABLE I
LTE-EPC PERFORMANCE SIMULATION SETUP PARAMETERS

Parameter	Value
Simulation time (s)	20
Application run-time (s)	20
Downlink Bandwidth (MHz)	5
Resource Block	25
Distance dependent pathloss model	COST231 propagation loss model
Antenna type	2x2 MIMO, SISO
eNB Tx power (dBm)	30, 46 and 60
Fading model	Fast fading
Shadowing deviation	7
UE speed of interest (Kmph)	0
Operating frequency band	2GHz
Downlink EARFCN	100
Traffic model	video
Video packet generation interval (ms)	100
EPS bearer type	NGBR
MAC Scheduler Type	PSS
Encoding scheme	Systematic RLNC
Galois field	Binary8
Generation Size	100
Packet Size (byte)	1024

of the applications on the remote host and the UE are the same as the simulation duration so that enough packets are generated and encoded within the given duration at a packet generation interval of 100ms. Furthermore, the used distance dependent propagation model is COST231 propagation loss model[3]. In addition, a fast fading model generated with a Rayleigh channel is included in order to consider the impact of the fading on the radio signals. Some of the values for the parameters are taken from [1] so that we simulate a topology that is as close as possible to a real urban area LTE network topology. The table also shows the coding parameters. A systematic coding is performed over the Galois Field GF(256). The generation size is 100 packets where each packet has size of 1024 bytes.

IV. SIMULATION RESULTS AND ANALYSIS

We evaluate the performance when encoded and non-encoded packets are sent. Accordingly a comparison between the throughput and the delay performance for the both schemes is made. The throughput is defined as the rate of successful data delivery, while the delay is defined as the end-to-end delay for all the packets sent from the remote host to the UE are successfully decoded.

A. Throughput performance analysis

To analyse the throughput performance with RLNC, 100 packets each of size 1024 bytes, are encoded and sent to the UE from the remote host. The simulation is performed for three different eNB transmission powers, UE-eNB distance in the range between 0.3km and 4km and bandwidth of 5MHz. We run the simulation with these parameters so that we can

[3]COST231 propagation loss model is applicable to urban areas to evaluate path loss of radio signals in frequency range 1500 MHz to 2000 MHz and link distance of up to 20 km.

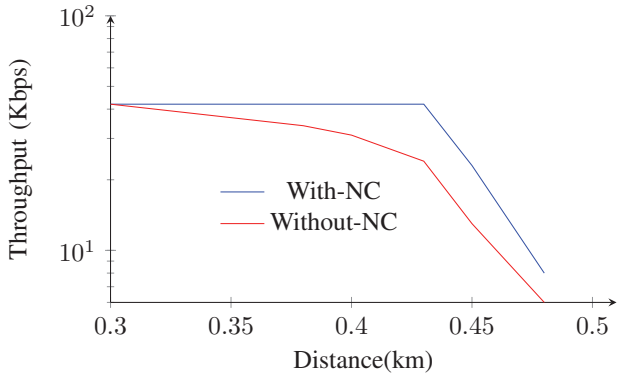


Fig. 5. Throughput performance for 30dBm eNB transmission power: SISO, 5MHz bandwidth

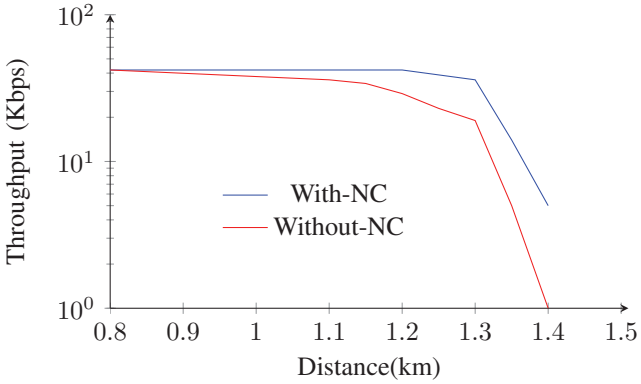


Fig. 6. Throughput performance for 46dBm eNB transmission power: SISO, 5MHz bandwidth

observe the effect of the distance on the number of successfully received packets. If the UE is close enough to the eNB there will be no packet loss since there is only one UE for the given system bandwidth.

Figures 5, 6 and 7 show the difference in the throughput performance obtained with and without RLNC, for 30, 46 and 60dBm eNB transmission power, respectively. We can conclude from these figures the throughput is improved with RLNC. When RLNC is used, the only requirement is enough linearly independant packets to be received at the UE, i.e., an exact packet should not be received in a specific order. Note that RLNC is performed over HARQ which by itself is an efficient way of recovering lost packets with retransmission requests. The lowest value of the distance on the x-axis in Figures 5, 6 and 7 indicates that with the given power level all the sent packets are fully recovered up until that distance. However, as we increase the distance, the number of lost packets is increasing. For instance, for 30dBm power the throughput is the same with and without RLNC up until 0.3km. When the distance is more than 0.3km, the throughput performance is better with RLNC.

The reason why we depict the throughput performance when different eNB transmission power levels are used is to check the effect of network coding at different distances and to show

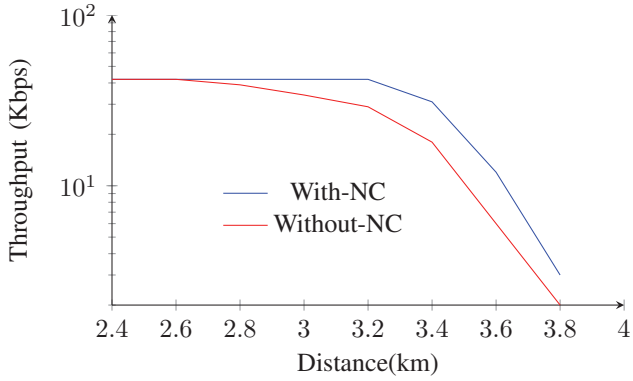


Fig. 7. Throughput performance for 60dBm eNB transmission power : SISO, 5MHz bandwidth

how the performance can be improved not only with network coding, but also by increasing the transmission power. As we can see better throughput performance in LTE networks is achieved with increasing the transmission power. For example the throughput and the coverage that can be obtained with 30dBm and RLNC can easily be achieved by increasing the transmission power to 46dB and not using RLNC. However this is only for the downlink communication where eNB power transmission is not a big concern, but for UEs increasing the transmission power could have its own effect on the battery life of the UE. Thus, we can say that without having to deal with the complexity of deploying new RLNC scheme on the existing HARQ scheme the required performance can be achieved with increasing the transmission power.

Note that the throughput is in Kbps, because only 100 packets each of size 1024 bytes are sent during 20s simulation time.

B. Delay performance analysis

Figures 8, 9 and 10 show the delay performance at 30, 46 and 60 dBm eNB transmission power levels. As shown the delay is higher with RLNC. This is because in RLNC in addition to the transmission delay, the delay includes the time to encode a packet, receive enough linearly independent packets and decode them. We also see how the delay reduces for higher distance between the UE and the eNB. This is in line with the decreasing throughput which indicates the total number of received packets are less so the delay as well.

The way how we calculate the delay is imposed by ns3. The flow monitor tracks the received packets and calculates the delay for these packets. When the distance is big, the receiver does not receive packets during the simulation time. That is the reason why the delay goes towards zero. In real-world scenarios, the delay actually goes towards infinity.

C. AL-RLNC vs. MIMO

In this part, it is shown that MIMO could be an alternative to deploying RLNC in LTE networks. MIMO techiques are one of the major enablers for LTE. They allows higher data rate transmission through the use of multiple antennas at the

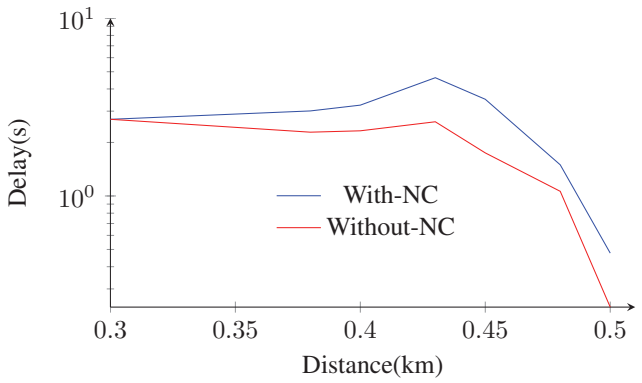


Fig. 8. Delay performance for 30dBm eNB transmission power: SISO, 5MHz bandwidth

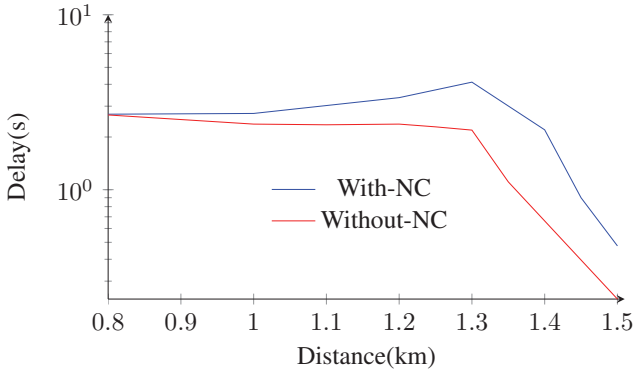


Fig. 9. Delay performance for 46dBm eNB transmission power: SISO, 5MHz bandwidth

transmitter and the receiver. Figure 11 shows the throughput obtained for 60dBm eNB transmission power for MIMO and Single-input Single-output (SISO) systems with and without network coding. As shown in the figure the throughput obtained with MIMO system without using RLNC is higher at any given distance as compared to SISO with network coding. We can also see that the MIMO system increases the coverage of the network up until 5km. Thus, we can expect a better

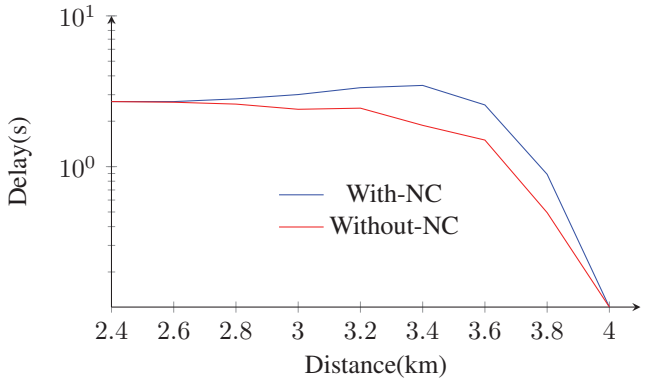


Fig. 10. Delay performance for 60dBm eNB transmission power: SISO, 5MHz bandwidth

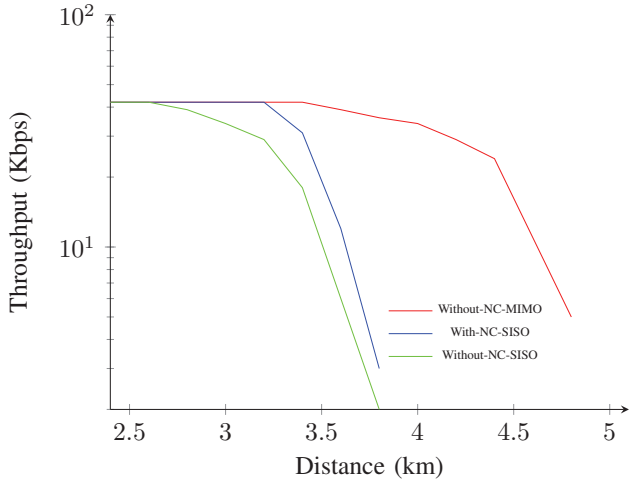


Fig. 11. MIMO, SISO and RLNC performance comparison for 60dBm eNB transmission power

throughput performance than the graph shows if we combine MIMO with network coding.

V. Conclusions

This paper presented an AL-RLNC integration in LTE networks deployed on top of HARQ with a goal of achieving an efficient and flexible multimedia delivery mechanism. The presented results indicate that the throughput performance of LTE network is improved by using RLNC but at the cost of a higher packet delay. We have also shown that the performance improvement with RLNC can also be obtained either with increasing the eNB transmission power or by using MIMO techique. The results presented in this paper are obtained for a simple topology and the performance can further be analysed and evaluated with multiple UEs and remote hosts.

References

[1] Vodafone Alcatel-Lucent, picoChip Designs. Simulation assumptions and parameters for fdd henb rf requirements. *3GPP TSG RAN WG4 (Radio) Meeting*, May 2009.

[2] Szymon Chachulski, Michael Jennings, Sachin Katti, and Dina Katabi. Trading structure for randomness in wireless opportunistic routing. In Jun Murai and Kenjiro Cho, editors, *SIGCOMM*, pages 169–180. ACM, 2007.

[3] UMTS Forum. Mobile traffic forecasts 2010-2020 report. *UMTS Forum Report 44*, January 2011.

[4] Christina Fragouli, Jean-Yves Le Boudec, and Jörg Widmer. Network coding: an instant primer. *Computer Communication Review*, 36(1):63–68, 2006.

[5] Pavel Loskot Hassan Hamdoun. Implementing network coding in LTE and LTE-A. *International Workshop on Smart Wireless Communications-SWICOM2012*, 2012.

[6] Tracey Ho, M. Medard, R. Koetter, D.R. Karger, M. Effros, Jun Shi, and B. Leong. A random linear network coding approach to multicast. *Information Theory, IEEE Transactions on*, 52(10):4413–4430, Oct 2006.

[7] Sachin Katti, Hariharan Rahul, Wenjun Hu, Dina Katabi, Muriel Médard, and Jon Crowcroft. XORs in the air: practical wireless network coding. *IEEE/ACM Trans. Netw*, 16(3):497–510, 2008.

[8] Chadi Khirallah, Dejan Vukobratovic, and John S. Thompson. Performance analysis and energy efficiency of random network coding in LTE-advanced. *IEEE Transactions on Wireless Communications*, 11(12):4275–4285, 2012.

[9] Ashutosh Kulkarni, Michael Heindlmaier, Danail Traskov, Marie-José Montpetit, and Muriel Médard. An implementation of network coding with association policies in heterogeneous networks. In Vicente Casares Giner, Pietro Manzoni, and Ana Pont, editors, *Networking Workshops*, volume 6827 of *Lecture Notes in Computer Science*, pages 110–118. Springer, 2011.

[10] Morten Videbæk Pedersen, Janus Heide, and Frank H. P. Fitzek. Kodo: An open and research oriented network coding library. In Vicente Casares Giner, Pietro Manzoni, and Ana Pont, editors, *Networking Workshops*, volume 6827 of *Lecture Notes in Computer Science*, pages 145–152. Springer, 2011.

[11] Hassan Shojania and Baochun Li. Random network coding on the iphone: fact or fiction? In Wei Tsang Ooi and Dongyan Xu, editors, *NOSSDAV*, pages 37–42. ACM, 2009.

[12] Surat Teerapittayanon, Kerim Fouli, Muriel Médard, Marie-José Montpetit, Xiaomeng Shi, Ivan Seskar, and Abhimanyu Gosain. Network coding as a wiMAX link reliability mechanism. *CoRR*, abs/1208.4766, 2012.

[13] P. Vingelmann, F.H.P. Fitzek, M.V. Pedersen, J. Heide, and H. Charaf. Synchronized multimedia streaming on the iphone platform with network coding. *Communications Magazine, IEEE*, 49(6):126–132, June 2011.

[14] Dejan Vukobratovic, Chadi Khirallah, Vladimir Stankovic, and John S. Thompson. Random network coding for multimedia delivery over LTE-advanced. In *ICME*, pages 200–205. IEEE Computer Society, 2012.

[15] Dejan Vukobratovic, Chadi Khirallah, Vladimir Stankovic, and John S. Thompson. Random network coding for multimedia delivery services in LTE/LTE-advanced. *IEEE Transactions on Multimedia*, 16(1):277–282, 2014.

VoLTE E2E Performance Management

D.Klobučarević, Ž. Klobučarević and D. Belošić
Hrvatski Telekom/T5, Zagreb, Croatia
Hrvatski Telekom/T4, Zagreb, Croatia
Nokia-HR, Zagreb, Croatia
danijela.klobucarevic@t.ht.hr
zarko.klobucarevic@t.ht.hr
darko.belosic@nokia.com

Abstract – By introducing new Services on IMS (IP Multimedia Subsystem) platform, there is also a need for Quality of Service (QoS) monitoring. There is an overall opinion that QoS is main and key differentiator on telecommunication market. HT has a quite big experience in providing telecommunication services and along with this comes quality monitoring. These days focus is ensuring QoS monitoring in new technology environment. Working in multivendor and multinational environment, which is very demanding and challenging becomes important to have strong support from international recommendations dealing with Performance Management of new Services. Focus of this paper is to describe methodology of introducing necessary Key Performance Indicators (KPIs), supporting and reporting systems for monitoring and collecting and benchmarking of KPIs. In addition, some results from previous international projects which were the base of future VoLTE E2E Performance Management are given.

I. INTRODUCTION

As VoLTE has become very present service in many worldwide Telecom operators, introducing Performance Management System (PMS) for purpose of surveillance and reporting of KPIs becomes necessary.

By introducing VoLTE service in Croatian Telecom Network, there are in parallel activities going on considering PMS.

Previous experience from international IMS Projects has given us knowledge what will be the main challenges for VoLTE QoS Project.

Working in multivendor environment is the next big challenge: it is necessary to have strong backup in ITU and ETSI recommendations in order to have same or almost same definitions for the KPIs. Another strong requirement is comparison of KPIs between different technologies and services (2G, 3G and VoLTE). This is sometimes not easy to achieve irrespective of same or near the same definition. In addition to service level KPIs, there are other KPIs needed for monitoring health of the Network Elements (NE), such as Capacity, CPU, Overload KPIs, etc.

II. VOLTE SERVICE IN IMS ENVIRONMENT

VoLTE is based on technology specification that defines the standards and procedures for delivering multimedia communication over LTE access networks. For VoLTE service realization there are several subsystems involved:

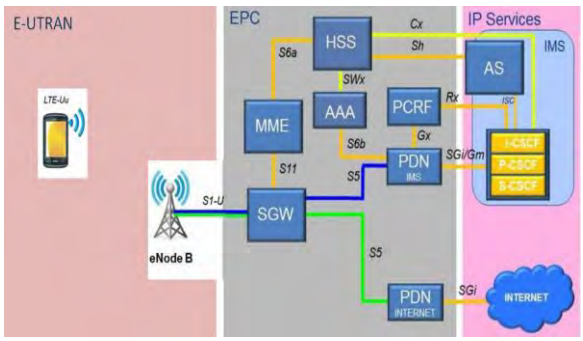


Figure 1. VoLTE architecture

IMS core - LTE access network is native IP technology and perfectly fits to IP Multimedia Subsystem (IMS) framework (Figure 1). Here VoLTE service is considered as full network concept which covers creating, provisioning and managing voice, video and messaging services in a 4G access network. In short, IMS core network is taking care about authentication between user and service, registration of user, routing of sessions and breakout of calls and transcoding if necessary.

On top of the HT IMS Core Network there will be VoLTE Application Server (AS) which is responsible for originating and terminating service execution.

Evolved Packet Core (EPC) - together with E-UTRAN (Evolved Terrestrial Radio Access Network) is responsible for authentication between user and network, collecting, transporting, aggregating and policing user IP traffic. Depending of the service chosen by customer IP, user traffic is to be delivered to different domains. On Figure 1. delivering of user traffic (media and signaling) to IMS network for further processing is shown.

IP transport network – responsible for connecting above-mentioned subsystems and transporting IP packets in accordance with predefined IP Quality of Service per service. IP transport quality of service KPIs which have to be obeyed per service are: packet loss, delay, jitter and availability of IP transport network.

III. KEY PERFORMANCE INDICATORS FOR VOLTE

According to ETSI and 3GPP recommendations, several basic KPIs should be monitored and reported for Performance Management of VoLTE service. Monitoring of these KPIs will be done on AS, IMS Core, EPC and E-UTRAN. Some KPIs can be derived directly from one subsystem, but there are also cases when KPIs has to be

obtained and calculated from several different subsystems. To correlate those KPIs it is important to fully understand the KPI triggering points from each network element involved.

A. Top Key Performance Indicators

KPIs can be divided in several main groups (Top KPIs): Availability, Accessibility, Retainability, Utilization, Mobility and Integrity / Quality.

Availability is total availability of the network elements including EPC, IMS Core and VoLTE AS considering overall network redundancy architecture.

Accessibility consists of several KPIs considering possibility and fastness of subscribers to access the network; Attach Success Ratio, Initial Registration Success Rate, Session Setup Time (Mean), Session Establishment Success Rate, Re-registration Success Rate, Session Setup Time Originated and Session Setup Time Terminated .

Retainability KPIs are those one which show the possibility of keeping service going on; Call drop ratio and Mean Holding Time of the Sessions.

Utilization KPIs show usage of service; Number of Users Registrated and Number of Successful Sessions.

Mobility KPIs describes Quality of Sessions while subscriber is moving through network and changing service areas or doing handover to legacy networks; EPC Handover Success Ratio, eSRVCC Success Rate to legacy (2G&3G) Circuited Switched (CS) network.

Integrity / Quality KPIs are Voice Quality parameters; calculated Mean Opinion Score (MOS) is based on QoS Parameters relevant for IP transport and EPC networks (Latency, Jitter, Packet Loss).

B. End to End Quality Parameters

One of the main drivers for introducing VoLTE service is to improve quality perceived by customer: fastest call setup, evolved HD voice (eHD voice) and increased battery life for user equipment compared to OTT VoIP.

Different KPIs can be measured on different Network Elements and some of them can be measured on more than one Network Element. This is very important in order to have possibility for future benchmarking of parameters and complete Performance Management introduction for VoLTE service.

Availability KPI is obtained from Fault Management Systems and represents Total Time of Failure over total observed time.

Accessibility KPIs are mostly measured on Call Session Control Function (CSCF), but some of them like Attach Success Ratio could be measured on E-UTRAN or on EPC. Session Setup Time is obtained from VoLTE AS or from R-UTRAN.

Retainability parameters are obtained from CSCF and E-UTRAN.

Utilization KPIs are obtained from capacity measurement on VoLTE AS and CSCF.

Mobility KPIs are regularly measured on E-UTRAN and EPC and additionally eSRVCC KPIs can be measured on legacy CS network element (MSC Server) and on VoLTE AS.

Integrity/Quality KPIs are measured on eNodeB, MGW and BGW.

C. Results from previous international projects

Participating in some international projects and defining necessary KPIs according to VoLTE PM Policy, we have managed to introduce common reporting tool for future benchmarking between different vendors and National Companies (NatCos) in DT group. This approach makes availability of reports much easier for all participants in process. According to previous defined KPIs on different NE, it would be much easier to introduce Performance Management for new service.

In addition, there are some results existing for different KPIs according to definition of Accessibility, Retainability, Performance and Mobility TOP KPIs.

In Figure 2. the parameter Accessibility is presented. It contains average value of IMS Registration Success, IMS Re-registration Success and percentage of IMS Session Set-up Time which are less than 3s. It can be seen that Accessibility is increased as Service is improved.

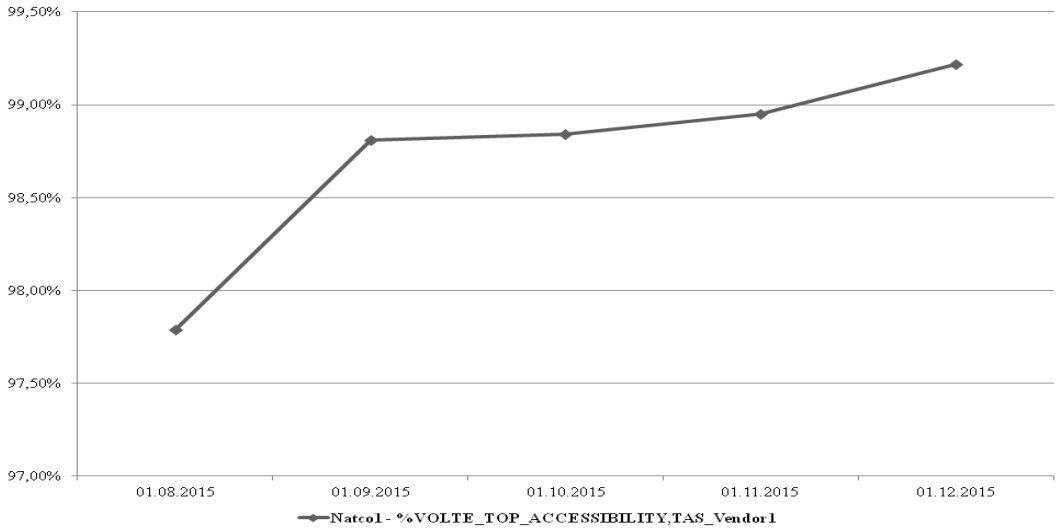


Figure 2. Top Accessibility

In Figure 3. the parameter Retainability can be seen. It consists of two parameters; Call Drop Ratio Originating and Terminating Call Drop Ratio is rather high but it is normal situation because every time new service is introduced it takes time for solving unexpected errors that cannot be anticipated in the test phase.

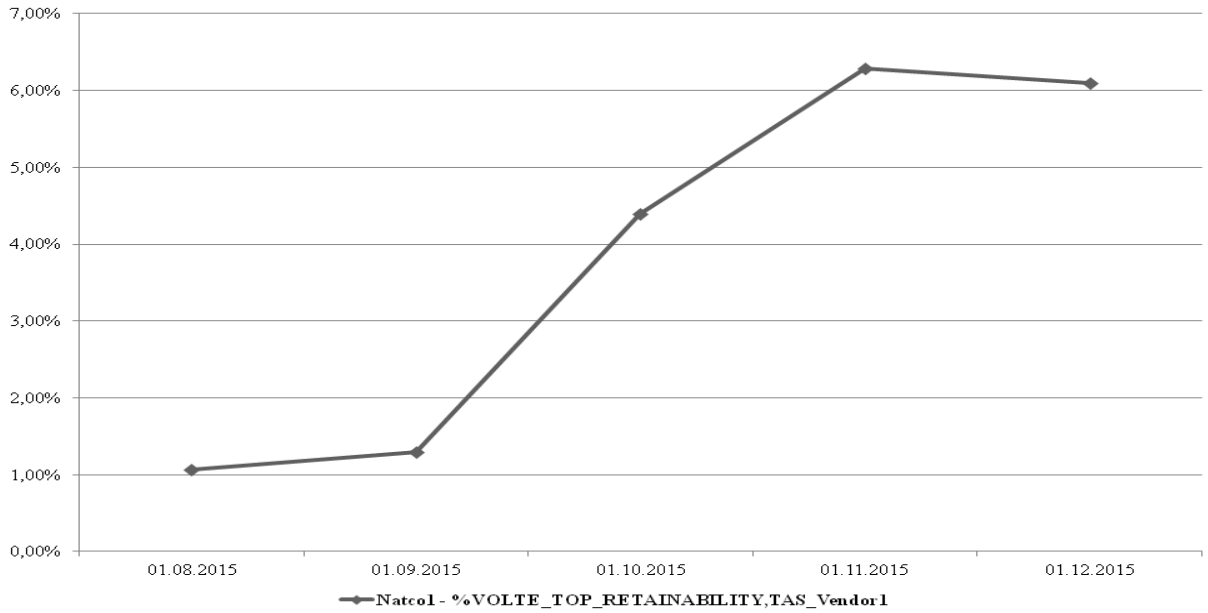


Figure 3. Top Retainability

In Figure 4. the parameter Performance is presented. It consists of parameter Ineffective Session Attempts Ratio, and it is defined according to RFC 6076 as inversely proportional value of this parameter. SIP Cause Codes which are calculated in this parameter are: 408 Request Timeout, 500 Server Internal Error, 503 Service Unavailable and 504 Server Time-out.

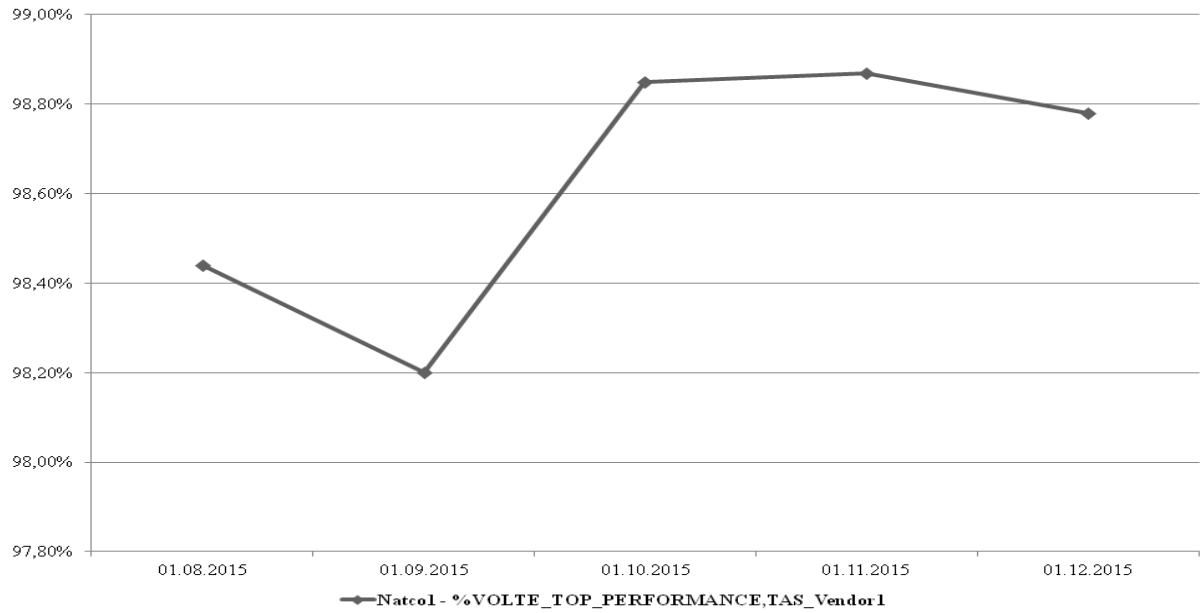


Figure 4. Top Performance

In Figures 5. and 6. the parameter Mobility is presented; eSRVCC Success Rate to 2G and eSRVCC Success Rate to 3G. It can be seen that parameters are slightly worse for 2G Service, as well as for Vendor 3. Those parameters are very good example of comparison between different Vendors, and future negotiation for Service Level Agreements.

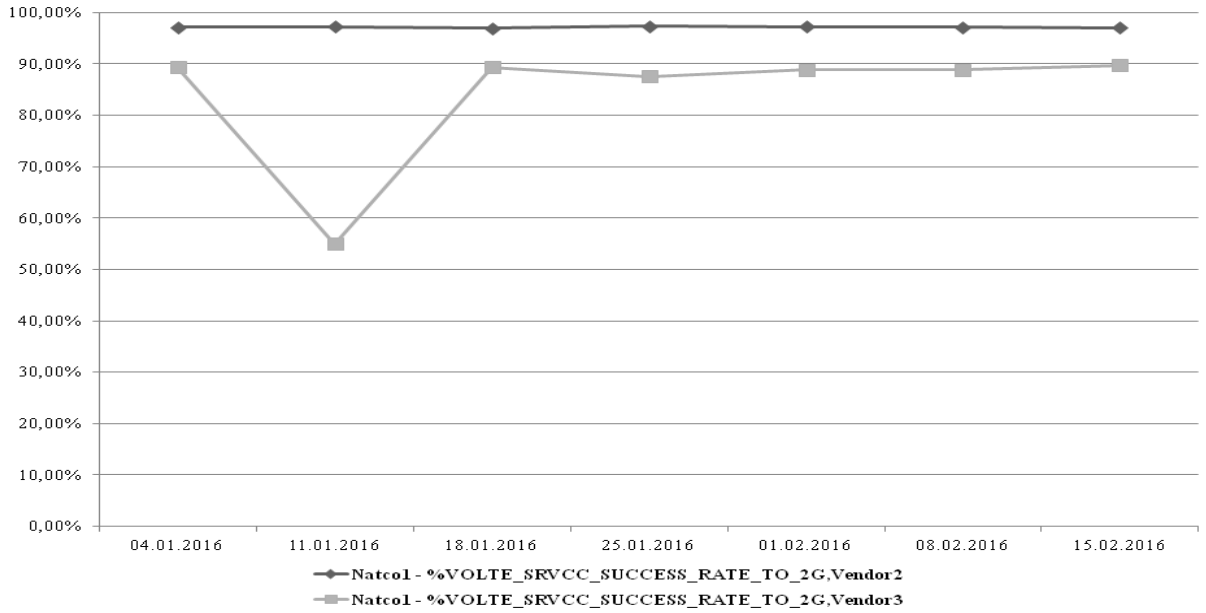


Figure 5. eSRVCC Success Rate to 2G

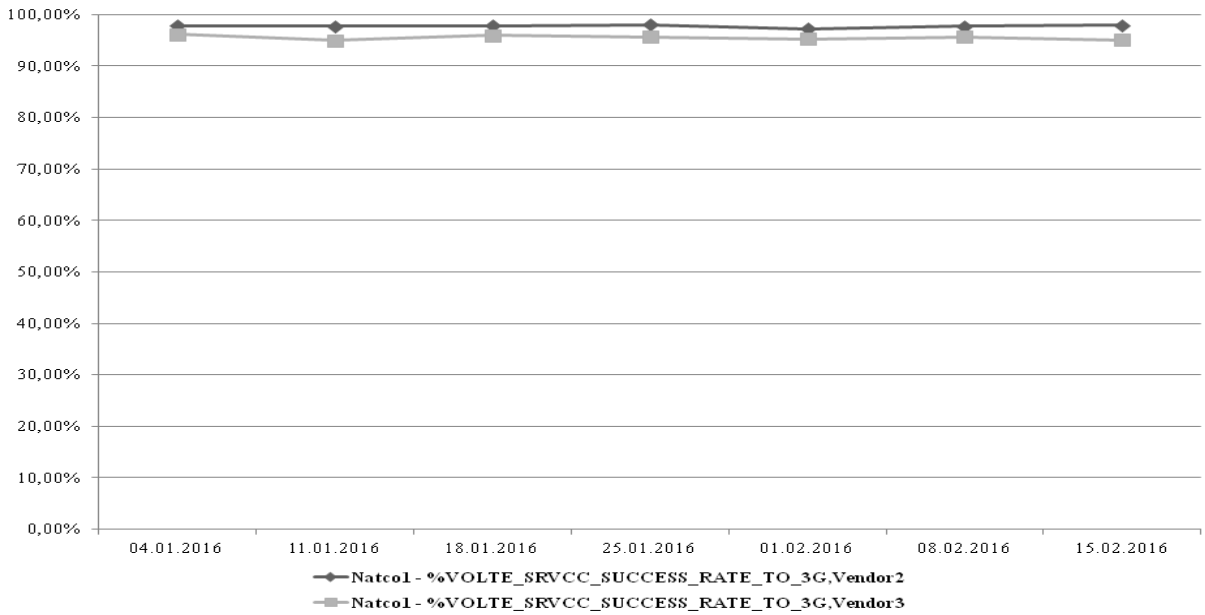


Figure 6. eSRVCC Success Rate to 3G

D. Additional Performance Management Systems

One more possibility to have overall Network view and complete Performance Management Policy is to have additional system for monitoring such is signaling monitoring system. Introducing signaling probes on different Network Elements, it is possible to obtain detailed information of each session, as well as overall statistics necessary for Performance Management. According to obtained values and differences from the expected values, Alarm Policy could be defined in purpose of proactive Network Monitoring.

IV. CONCLUSION

Working in multivendor environment is very challenging and demanding, but by using strictly regulated definitions of KPIs following international recommendations, it is possible to obtain high quality Performance Management System in spite of different vendors for different Network Elements (NE).

Obtaining results from different NE and different Vendors, it is shown that E2E Voice Quality for VoLTE is possible. The next valuable step is benchmarking of KPIs between NatCos and Vendors for purpose of improving Service E2E Quality.

In spite of very complex and demanding architecture and functionality, there is a need for obtaining quality Performance Management System for VoLTE service for proactive Network and Service Management. Methodologies described in this paper are proposal of how it could be done in the future.

LITERATURE

[1] ETSI TR 103 219 V1.1.1 (2015-04), "Speech and multimedia Transmission Quality (STQ); Quality of Service aspects of voice communication in an LTE environment"

[2] 3GPP TS 32.454 V10.0.0 (2011-03), "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Telecommunication management; Key Performance Indicators (KPI) for the IP Multimedia Subsystem (IMS); Definitions (Release 10) "

[3] ETSI TS 132 409 V7.3.0 (2009-01), " ETSI TS 132 409 V7.3.0(2009-01) Technical Specification Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Telecommunication management; Performance Management (PM); Performance measurements IP Multimedia Subsystem (IMS) (3GPP TS 32.409 version 7.3.0 Release 7) "

[4] K. Silvennoinen, H. Holma and J. Reunanen, Nokia, "VoLTE Performance", unpublished

Balancing Security and Blocking Performance with Reconfiguration of the Elastic Optical Spectrum

Sandeep Kumar Singh, Wolfgang Bziuk, and Admela Jukan
Technische Universität Carolo-Wilhelmina zu Braunschweig, Germany
{sandeep.singh, w.bziuk, a.jukan}@tu-bs.de

***Abstract*—Reconfiguration of connections along frequency axis, – similar to frequency hopping in wireless systems, has been shown to improving security in elastic optical networks (EONs). However, the effect of spectrum reallocation on the connections' blocking and the associated connections' security is still under-explored. In this paper, we examine the effect of reconfiguration of connections on its security and blocking performance of an elastic optical link (EOL). We propose a Reconfiguration-as-a-Service (RaaS) method, which randomly reconfigures connections to improve its security, and for that scenario we model the connection patterns of an elastic optical link. We apply a multi-class continuous-time Markov chain under the first-fit and random-fit spectrum allocation (SA) method. Analytical results show that the random reconfiguration scheme can improve security at the cost of only a slight increase in blocking, depending on the rate at which it is performed, the link load and the capacity of an EOL. We hence show that it is indeed possible to balance the security and blocking performance with RaaS.**

I. Introduction

The emergence of elastic optical network (EON) has opened the door for flexible spectrum resource allocation to meet the increasingly diverse and bandwidth-intensive traffic demands [1]. At the same time, however, EON inherited some of the security challenges from its predecessors, wavelength division multiplexing networks, and that is the optical layer security [2]. In general, the light can be coupled from the optical fibers using tapping devices by an eavesdropper. Another approach of eavesdropping is to listen to the residual crosstalk from an adjacent channel while spoofing one of the users in EONs. In EONs, different orthogonal frequency-division multiplexing (OFDM) subcarriers are allocated to different users of variable demands. In a practical OFDM system, where perfect channel isolation (or orthogonality) is absent, an eavesdropper can listen to interchannel-crosstalk by dropping a desired channel belonging to a particular user.

In EONs, the security challenge can be mitigated by using the dimensional transformation of encrypted data along the time and frequency domains, which makes it difficult to decode data. In the past studies, the optical layer security has been widely studied [2], where data encryption is augmented with optical coding using optical code division multiple access. Recently, authors in [3] permuted the encrypted data in frequency and time domains to generate a large key space to resist brute-force attacks. However, reconfiguration of connections along frequency is challenging to model and evaluate analytically as one needs to take into account of spectrum continuity and contiguity constraint. Additionally, the traffic of live connections are disrupted due to the retuning of transceivers and reconfiguration of optical switches. Hence, a few studies showed that disruption can be minimized using "make-before-break" [4] or hitless defragmentation method [5], [6]. To the best of our knowledge, there have not been much effort in this field to analyze how a reconfiguration process impacts the connections' security and blocking.

The analytical blocking performance evaluation of EONs has only started recently [7]–[11]. In [7], authors used the Erlang-based loss model to calculate the blocking probability for different spectrum allocation policies under time-varying traffic conditions. In [8], exact blocking probability of an elastic optical link (EOL) was evaluated for the first time by modeling the bandwidth occupancy as continuous-time Markov chain under the first-fit (FF) and random-fit (RF) spectrum allocation methods, which is later extended by [9] to suit for a general EON. Since the number of states grows exponentially, the paper presented two approximation methods for the large-scale networks. Recently, in [10] the same model was used to present resource and fragmentation blocking separately. In [11], we proposed a defragmentation-as-a-service (DaaS) method to analyze the effect of defragmentation on the overall blocking performance, and showed that the overall blocking can be reduced if DaaS rate is higher than the connections' service rates.

In this paper, we propose a reconfiguration-as-a-service (RaaS) method, and analytically examine its effect on the connections' security and blocking. When an attacker eavesdrops a range of spectrum, say f_0 to f_n (assuming that eavesdropper has access to a transponder with fixed capacity $f_n - f_0$), it can access the data of the connections allocated over these frequencies. Therefore, RaaS can be triggered regularly to enhance the connections' security by randomly reallocating the spectrum to the existing connections. Depending on the number of reconfigurations during the lifetime (service) of a connection, an eavesdropper can access a part of the total transmitted data volume of those connections. To see the effect of RaaS on security, we calculate the fraction of observed data in a spectrum window (f_0 to f_n) that an eavesdropper can access during the average service time of a connection. Note that if RaaS is not performed within the lifetime of a connection that can be listened by an eavesdropper, then the all data volume of the connection is observed. Therefore, in this case, security of those data volumes depends if the eavesdropper can decode the data or not. However, in our

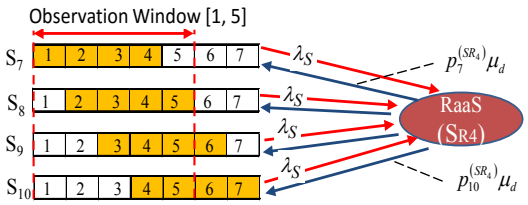


Fig. 1. Possible state transition from and into a RaaS state.

model, only a part of the connection data will be accessed by the attacker if RaaS is performed frequently, which enhances the data security on top of the encryption.

For the analytical modeling of RaaS, we model an EOL using a multi-class continuous-time Markov chain (CTMC), and include RaaS states in addition to the regular (data) service states. To give some insight into the cause of connection blocking, we differentiate between resource blocking and blocking due to fragmentation, which are caused by the unavailability of free spectrum slots, and the unavailability of free consecutive slots to satisfy the demand, respectively. Additionally, in our model, all requests arriving during the RaaS processing times are blocked, which are termed reconfiguration blocking. The results show that although reconfiguration blocking is added to the resource and fragmentation blocking, the overall blocking can be close to a system without reconfiguration, if the reconfiguration service rate is higher than the normal service rates of the connections. We hence show that it is indeed possible to balance the security and blocking performance with RaaS. Furthermore, we show that the fraction of observed data decreases with the increase in the number of reconfiguration performed in an average lifetime of a connection.

The rest of this paper is organized as follows. Section II presents an example to understand the behavior of our model. We analyze the overall blocking in Section III, and present the security analysis in Section IV. In Section V, we evaluate the performance, and conclude the paper in Section VI.

II. System Description

A reconfiguration of the spectrum resources of existing connections can result in two types of SA patterns: fragmented and defragmented patterns. This is illustrated in Fig. 1, where the spectrum assigned to a connection with demand $d = 4$ slot is randomly reconfigured by entering the RaaS process (state SR_4), and as a result, one of the states $\{S_7, S_8, S_9, S_{10}\}$ is chosen with equal probability. Here, $S_8 and S_9 (S_7 and S_{10})$ are fragmented (not fragmented) states, which will block (accept) an incoming request with demand $d = 3$ slots. In our model, RaaS process is triggered to randomize the spectrum assigned to the connections. This makes it more difficult to rapidly detect the spectrum assignment pattern, and to successfully demodulate the spectrum, thus increases the system security.

A. Underlining Assumptions

We have the following assumptions for our model. i) The RaaS process is triggered at random points in time, and the reconfiguration results into a spectrum pattern which is chosen randomly out of all possible rearrangement of connections. ii) The services of all existing requests are interrupted during the reconfiguration time of the RaaS process. iii) Further incoming requests are blocked during the RaaS period. Assumption (ii) is disadvantageous, because traffic interruption of some or all existing connections does happen during reconfiguration in the real system. But there are a lot of techniques to reduce the effective disruption time. Due to assumption (iii) we would experience an increase in overall call blocking probability.

B. System Model

We adopt a multi-class CTMC to model the occupancy pattern of an EOL [8]–[10]. Here, we assume that both inter-arrival times T_{E_k} and call holding times T_{H_k} of class-k requests are independent and exponentially distributed with average rates λ_k and μ_k, respectively. To model the reconfiguration of the occupied slots in RaaS states, we introduce reconfiguration times T_{RT}, which are exponentially distributed with rate μ_d. The RaaS process is randomly triggered at security interarrival times T_S, which are exponentially distributed with average rate λ_S. A specific RaaS state SR_ν is triggered by states belonging to the set $\Gamma(SR_\nu)$. In Fig. 1, all states out of set $\Gamma(SR_4) = \{S_7, S_8, S_9, S_{10}\}$ transit to RaaS state SR_4 at rate λ_S. If the system is in a RaaS state SR_ν, after reconfiguration it transits back to a randomly chosen system state $S_i \in \Gamma(SR_\nu)$ with probability $p_i^{SR_\nu}$, e.g. in Fig.1 from state SR_4 to state S_{10} with rate $\mu_d p_{10}^{SR_4} = \mu_d/4$. Due to the randomization strategy we assume equal transition probabilities $p_i^{SR_\nu} = 1/m(\Gamma(SR_\nu)$, where $m(\Gamma(SR_\nu)$ is the number of elements in $\Gamma(SR_\nu)$. Although all states $S_j \in \Gamma(SR_\nu)$ have different spectrum pattern, they are characterized by the same connection pattern $\mathbf{n}(S_j) = (n_1(S_j), ..., n_K(S_j))$, where a pattern is defined by the number of connections per class. Based on a chosen state S_j let us to define the set of states with same connection pattern as $\Gamma(S_j) = \{S_i | \mathbf{n}(S_i) = \mathbf{n}(S_j), i = 1, ..., N_{SA}\}$. Furthermore, connection pattern associated to a RaaS state is defined by the states of their target transitions. Thus $\Gamma(SR_\nu)$ defines all S_i associated for reconfiguration to RaaS state SR_ν.

The function $fr(\cdot)$ is used to identify a fragmented state, which can not allocate a block $FS_l(S_i)$ of $l \geq d_k$ consecutive free slots to a class-k request of demand d_k, even though it contains equal or more than d_k free slots, given as

$$fr(S_i, k) = \begin{cases} 1 & \text{if } d_k > max_l FS_l(S_i) \text{ and} \\ & d_k \leq C - \sum_j d_j n_j(S_i) \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

This allows us to define the set of fragmentation states for class-k requests as follows.

$$\mathbb{FB}(k) = \{S_i | fr(S_i, k) = 1, i = 1, \cdots, N_{SA}\}, \forall k \quad (2)$$

Resource blocking states, on the other hand, do not have sufficient free slots to satisfy an incoming class-k request, i.e.,

$$\mathbb{RB}(k) = \{S_i | d_k > C - \sum_j d_j n_j(S_i), i = 1, ..., N_{SA}\}. \quad (3)$$

Now, we consider an example of an EOL with capacity $C = 7$ slots under RF policy in Fig. 2. There are two

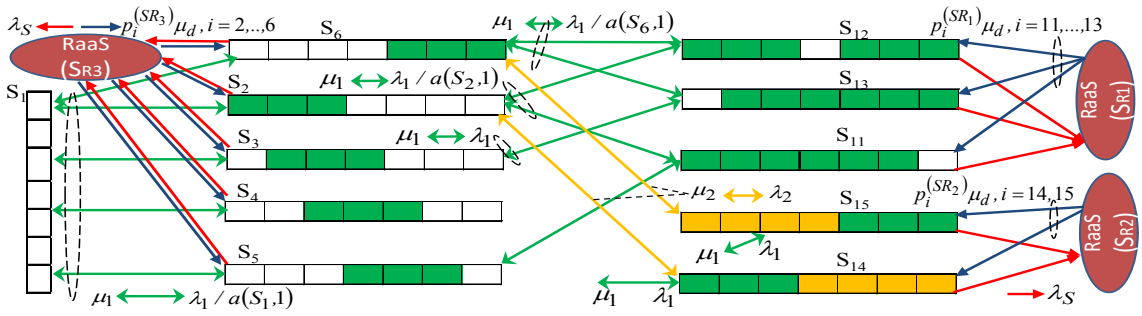


Fig. 2. Partial state diagram with transitions leading to spectrum reconfiguration for RF spectrum allocation method.

different classes of requests with demands 3 and 4 consecutive slots. Here, we only show a few states and transitions under RF policy. When a call of class $k = 1$ arrives in state S_1 with demand $d_1 = 3$, it can be allocated in one of the possible assignments shown by the transitions from S_1 to states $\{S_2, S_3, S_4, S_5, S_6\}$. In the target states, we have $n_1(S_i) = 1, i = 2, ..., 6$ class-1 calls, which will leave $F_1 = 4$ free slots, thus there are $a(S_1, k = 1) = \frac{(n_1+F_1)!}{n_1!F_1!} = 5$ different ways to allocate the 7 free slots, and transitions occurs with rates $\lambda_1/5$. Similarly, a class $k = 2$ arrival in state S_6 with demand $d_2 = 4$ will cause $S_6 \rightarrow S_{15}$ transition. On the other hand, transitions out of states $S_i, i = 2, ..., 15$, will occur due to the departure of a class-1 or class-2 requests with rate $\mu_k, k = 1, 2$. As we can see, when a class-2 request (4 slots) arrives, blocking of the request due to fragmentation will occur in the set of states $\mathbb{FB}(2) = \{S_3, S_4, S_5\}$. The set of states $\mathbb{RB}(1) \cap \mathbb{RB}(2) = \{S_{11}, \cdots, S_{15}\}$, on the other hand, are resource blocking states for both classes of requests. To improve the connections' security, we introduce RaaS states in our model which randomly reconfigures the existing connections. Here, for example, the arrivals of security indicator calls with rate λ_S in the set of states $\Gamma(SR_2) = \{S_{14}, S_{15}\}$ will transit to a RaaS state SR_2, and because $m(\Gamma(SR_2)) = 2$ the process transit back to a randomly chosen state $S_i \in \Gamma(SR_2)$ with rate $\mu_d/2$. Due to simplify our model, we do not differentiate the assignment of connections per class to spectrum slots, e.g. in state S_{12} we can simple change the assignment of connections to slots without changing the state.

Note that during the RaaS processing time, existing connections effected by spectrum reconfiguration are interrupted for a time duration equal to the exponentially distributed reconfiguration time T_{RT} with mean $1/\mu_d$. Due to the memoryless property of the exponential distribution, the remaining call holding times of interrupted calls are again exponentially distributed with mean $1/\mu_k$. Thus, all transitions due to call terminations after interruption are also modeled with rates μ_k.

III. System State Probability Analysis

In this section, we show how to derive the global balance equations (GBEs) of the Markov chain. The notation and parameters are given in Table I. Furthermore, we define the following terms to calculate the transition rates of the system.

$$A(S_i, k) = \begin{cases} 1 & \text{if a class-k request is accepted in } S_i \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

TABLE I
Notations and the parameters used in the model

Notation	Description
C	Total number of spectrum slots (or capacity units)
λ_k (μ_k)	Arrival (service) rate of class k calls, where $k = 1, 2, \cdots, K$
λ_S	Security indicator arrival rate
$1/\mu_d$	Mean reconfiguration time
d_k	Number of slots requested by class-k calls
$\mathbf{n}$	$\equiv (n_1, n_2, \ldots, n_K)$, where n_i is the number of class i calls
S_i	Occupancy state for normal operation $i = 1, ..., N_{SA}$
SR_ν	Reconfiguration state $\nu = 1, ..., N_R$
$p_i^{SR_\nu}$	transition probability from state SR_ν to target state S_i
$n_k(S_i)$	Number of class-k calls in state S_i
$\mathbf{n}(S_i)$	$\equiv (n_1(S_i), \ldots, n_K(S_i))$, realization of $\mathbf{n}$ in state S_i
$a(S_i, k)$	Number of different ways class-k call can be allocated in state S_i
$FS_l(S_i)$	Size of lth fragment of free slots in state S_i
$\mathbb{FB}(k)$	set of fragmentation states for class-k requests
$\mathbb{RB}(k)$	set of resource blocking states for class-k requests
$\Gamma(S_i)$	set of states S_j with same connection pattern $\mathbf{n}(S_i) = \mathbf{n}(S_j)$
$m(\Gamma(SR_\nu))$	number of elements of the set $\Gamma(SR_\nu)$
N_R (N_{SA})	Number of RaaS states (regular states without RaaS service)

$A(S_i, k)$ determines if a transition from state S_i is possible due to the arrival of a class-k call. Additionally, $T^{\pm}(S_j, S_i, k)$ determines the cause of transition if it is due to allocation (+) or deallocation (-), given as

$$T^{\pm}(S_j, S_i, k) = \begin{cases} 1 & \text{if } S_j \rightarrow S_i \text{ is due to class-k} \\ & \text{arrival (departure) in } S_j(S_i) \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

For simplicity, we have introduced a different state probability notation for RaaS states $P(SR_\nu), \nu = 1, ..., N_R$ in addition to the regular states $P(S_i), i = 1, ..., N_{SA}$. To model the different many-to-one and one-to-many mapping between Raas and regular states S_i, we introduce an indicator function, which compares the connection pattern of states given by

$$\delta(\mathbf{n}(S_i), \mathbf{n}(SR_\nu)) = \begin{cases} 1 & \text{if } \mathbf{n}(SR_\nu) = \mathbf{n}(S_i) \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

To illustrate the mapping between regular states and RaaS states, consider states S_{11}, S_{12} and S_{13} in Fig.2, which have equal number of class-1 and 2 calls, i.e. $\mathbf{n}$ = (2, 0). Therefore, the many-to-one mapping to RaaS state SR_1 can be defined

by $\delta(\mathbf{n}(S_i), \mathbf{n}(SR_1)) = 1$ for $i = 11, 12, 13$. Similarly, $\delta(\mathbf{n}(SR_\nu), \mathbf{n}(S_i))$ establishes the one-to-many mapping from RaaS to regular states. In Fig.2, all possible transitions from the RaaS state SR_1 to all states $S_i \in \Gamma(SR_1)$ are defined by $\delta(\mathbf{n}(SR_1), \mathbf{n}(S_i)) = 1$ for $i = 11, 12, 13$.

Finally, the global balance equations (GBEs) for normal occupancy states S_i under the RF policy are given as follows.

$$\left(\sum_{k=1}^{K} A(S_i,k)\lambda_k + n_k(S_i)\mu_k + \lambda_S\right)\pi(S_i) = \sum_{j=1, j\neq i}^{N_{SA}} \sum_{k=1}^{K} \left(\frac{\lambda_k T^+(S_j, S_i, k)}{a(S_j,k)} + T^-(S_j, S_i, k)\mu_k\right)\pi(S_j) + \sum_{\nu=1}^{N_R} \delta\left(\mathbf{n}(SR_\nu), \mathbf{n}(S_i)\right) \frac{\mu_d \pi(SR_\nu)}{m(\Gamma(SR_\nu))}, i = 1, \cdots, N_{SA} \quad (7)$$

In Eq. (7), left hand side represents the output flow rate from the state S_i including transitions to RaaS states (rate λ_S), while the right hand side represents input flow rate into the sate S_i. More precisely, the second line of Eq. (7) represents the input flows from other regular states S_j, while the last line of Eq.(7) defines the input flows from one RaaS state SR_ν which is mapped to S_i. The indicator $\delta\left(\mathbf{n}(SR_\nu), \mathbf{n}(S_i)\right)$ selects the correct RaaS state SR_ν with connection pattern equivalent to the state S_i, where the factor $m(\Gamma(SR_\nu))$ takes into account that the RaaS state only randomly selects a target state S_i with the probability $1/m(\Gamma(SR_\nu))$. Similarly, GBE of a RaaS state SR_ν, $\nu = 1, 2, \cdots, N_D$ can be given as follows.

$$\mu_d \pi(SR_\nu) = \lambda_S \sum_{S_j \in \Gamma(SR_\nu)} \pi(S_j), \nu = 1, 2, \cdots, N_R \quad (8)$$

As an example, the GBE for state S_6 in Fig.2 is given by $(\lambda_1 + \lambda_2 + \mu_1 + \lambda_S)\pi(S_6) = \lambda_1/5\pi(S_1) + \mu_2\pi(S_{15}) + \mu_d/5\pi(SR_3)$. The left (right) side characterizes the probability flow out of (into) S_6. The input rate $\lambda_1/a(S_1, 1) = \lambda_1/5$ takes into account that 7 free slots of S_1 results into 5 different spectrum pattern for class-1 (3 slots), one of which is S_6. Similarly, after reconfiguration in the RaaS state SR_3 with mean time $1/\mu_d$, there are $m(\Gamma(SR_3)) = 5$ possible transitions, one of which is $SR_3 \rightarrow S_6$ occurs with rate $\mu_d/5$. For the output flow, a class-1 (class-2) arrival with rate λ_1 (λ_2) can be accepted into S_6, and an existing connection can also depart with rate μ_1. More importantly, a security indicator call with rate λ_S can also trigger $S_6 \rightarrow SR_3$ transition. Hence, the output flow rate from state S_6 is $\lambda_1 + \lambda_2 + \mu_1 + \lambda_S$.

Similarly, we can write GBE for a RaaS state SR_2. The RaaS-state SR_2 is reached from a set of state $\Gamma(SR_2) = \{S_{14}, S_{15}\}$ when a security indicator call arrives with rate λ_S in one of these states. After reconfiguration in SR_2, system will return back to one of the possible states in $\Gamma(SR_2)$. However, the output flow rate from SR_2 is still μ_d. Hence, the GBE of state SR_2 is given by $\mu_d\pi(SR_2) = \lambda_S\left(\pi(S_{14} + \pi(S_{15})\right)$.

Under the stationary condition, the state probabilities $\pi = [\pi(S_1), \pi(S_2), \cdots, \pi(S_{N_{SA}}), \pi(SR_1), \cdots, \pi(SR_{N_R})]$ can be calculated by solving $\pi\ Q = 0$ subject to $\sum_i \pi(S_i) + \sum_\nu \pi(SR_\nu) = 1$, where $N = N_{SA} + N_R$ is the total number of states, and Q is the transition rate matrix.

The FF policy is also modeled by Eq. (7) - (8), taking into account, that some transitions are not allowed due to the assignment strategy. Here, a transition which leads to the first available consecutive slots allocation are only allowed, so we have $a(S_i, k) = 1$ for FF method.

The resource blocking of class-k is given as

$$P_B^R(k) = \sum_{S_i \in \mathbb{RB}(k)} \pi(S_i) \quad (9)$$

In additional, blocking also occurs due to the fragmentation, which is given for class-k by summing over the states which are fragmentation states for the class-k requests.

$$P_B^F(k) = \sum_{S_i \in \mathbb{FB}(k)} \pi(S_i) \quad (10)$$

Since RaaS states are blocking states for all classes of request, the blocking is given as follows.

$$P_B^D = \sum_{\nu=1}^{N_R} \pi(SR_\nu) \quad (11)$$

The total blocking in the RaaS model can be calculated as,

$$P_B = P_B^D + \frac{\sum_k \lambda_k (P_B^R(k) + P_B^F(k))}{\sum_k \lambda_k} \quad (12)$$

In the regular system i.e., without RaaS states, the GBE equations can be simplified. In Eq. (7), all transitions to and from states SR_ν are not present, i.e. setting $\lambda_S = 0$ and $\pi(SR_\nu) = 0$. Furthermore, Eq. (8) has to be omitted. The blocking for regular system can be calculated as follows.

$$P_B^{regular} = \frac{\sum_k \lambda_k (P_B^R(k) + P_B^F(k))}{\sum_k \lambda_k} \quad (13)$$

IV. Security Analysis

We included RaaS states to perform random reconfiguration of existing connections, which is triggered upon arrival of a security indicator request, and the security interarrival time is exponentially distributed with rate λ_S. And, after reconfiguration, it returns to one of all possible states out of set $\Gamma(SR_\nu)$ having same number of connections (but different SA patterns) as shown in Fig. 1 and Fig. 2. Let us calculate the total number of possible rearrangements of connections in a state S_i. If the state S_i contains $n_k(S_i)$ connections of $class - k$ with demand d_k, and there are $E(S_i) = C - \sum_{k=1}^{K} n_k(S_i) d_k$ empty slots. Then, for the RF policy the number of all possible rearrangements of connections of S_i with spectrum contiguity constraint can be given as follows.

$$R_{\mathbf{n}}(S_i) = \frac{(E(S_i) + \sum_{k=1}^{K} n_k(S_i))!}{E(S_i)!} \quad (14)$$

In contrast to RF, the possible rearrangements are much lesser in FF policy, for details see [9]. Now, let us assume that an eavesdropper can access a range of spectrum $[1, W]$, where $W \leq C$. An attack (eavesdrop) is defined as successful if the observed connections within the spectrum window ($\mathbf{n_{in}}$) before and after a reconfiguration remains same, and the

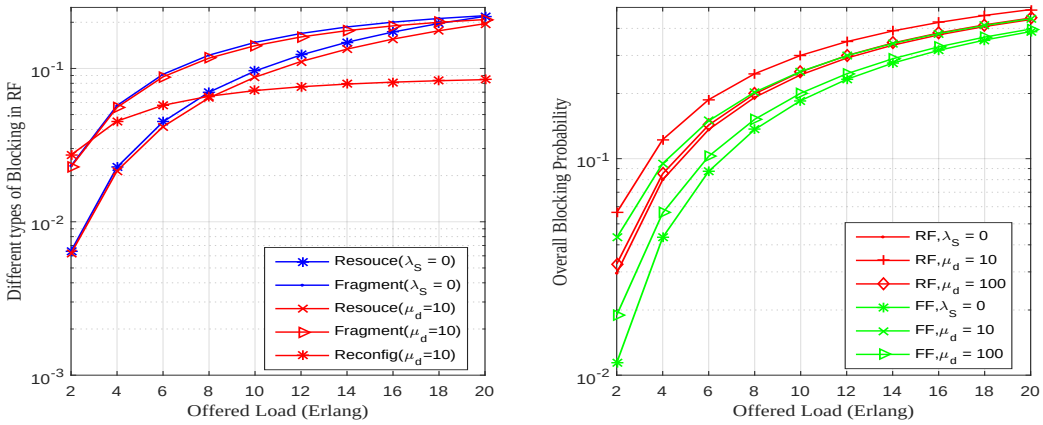


Fig. 3. (left) Different types of blocking in RF policy; (right) overall blocking in FF and RF policies. Security arrival rate $\lambda_S = 1$.

probability of a successful attack within a spectrum window $[1, W]$ (i.e., $P_{SA}^{[1,W]}$) is given as follows.

$$P_{SA}^{[1,W]} = \sum_{i=2}^{N_{SA}} \left[\frac{R_{\mathbf{nw}}(S_i)}{R_{\mathbf{n}}(S_i)} \times \frac{\pi(S_i)}{\sum_{j=2}^{N_{SA}} \pi(S_j)} \right] \quad (15)$$

In Eq. (15), $R_{\mathbf{nw}}(S_i)$ is the all possible rearrangements of connections of $\mathbf{n}(S_i)$ which results in same $\mathbf{n_{in}}$ before and after the reconfiguration. As we can see that Eq. (15) depends mainly on the window size W, and the connections within it. This can be explained with an example in Fig. 1, where an attacker can access data in the observation window [1, 5]. Here, there are four possible states $S_7 - S_{10}$ which can be reached if a reconfiguration is performed due the arrival of a security indicator request in any one of these states. Now, let us assume that an attacker listens to a connection (with demand 4-slots) within its window when the system is in state S_7, and the reconfiguration is triggered (i.e., $S_7 \rightarrow SR_4$). If the system returns to state S_7 or S_8 after the reconfiguration, the attacker can still access data. Therefore, the successful attack probability given that system is in state S_7 is 2/4. On the other hand, if there is no connection within $[1, W]$, then for a given state (e.g., S_9), the successful attack probability is zero.

The average number of reconfigurations during a lifetime of a connection (N_r) can be calculated as the ratio of the average holding time of a connection and the average reconfiguration interarrival time, i.e. $N_r = \frac{1/\mu_k}{1/\lambda_S} = \frac{\lambda_S}{\mu_k}$, where we restrict only to integer $N_r = 1, 2, \cdots$. Until reconfiguration is triggered for the first time, an attacker observes data successfully with probability 1. Therefore, the amount of data successfully observable before reconfiguration is $b \times \frac{1}{\lambda_S} \times 1$, where b is the average data rate observed in the window $[1, W]$. Similarly, until 2nd reconfigurations, successfully observable data is $b \times \frac{1}{\lambda_S} \times [1 + P_{SA}^{[1,W]}]$. In general, the amount of successfully observable data until N_r^{th} reconfigurations is given as follows.

$$D_{SA}^{[1,W]} = b \times \frac{1}{\lambda_S} \left[1 + P_{SA}^{[1,W]} + \cdots + [P_{SA}^{[1,W]}]^{N_r - 1} \right]$$
$$= b \times \frac{1}{\lambda_S} \times \frac{1 - [P_{SA}^{[1,W]}]^{N_r}}{1 - P_{SA}^{[1,W]}} \quad (16)$$

Now, the total amount of data transmitted in during the mean holding time $1/\mu_k$ is $\frac{b}{\mu_k}$. Therefore, the fraction of successfully observed data is given as the ratio of successful observed data and total transmitted data, and given as follows.

$$\Lambda^{[1,W]} = \frac{\mu_k}{\lambda_S} \times \frac{1 - [P_{SA}^{[1,W]}]^{N_r}}{1 - P_{SA}^{[1,W]}} \quad (17)$$

V. Numerical Results

In this section, we present the analytical results for an EOL having 20 spectrum slots. We consider three different classes ($K = 3$) of requests with demands $d_1 = 4, d_2 = 6$ and $d_3 = 8$ slots [10]. Arrival rates are uniformly distributed i.e., $\lambda_k = \frac{\lambda}{K}$, where λ is the total arrival rate. We assume mean holding time as one unit for all classes of requests [8]–[10]. Load on the link is calculated as $\sum_{k=1}^{K} \frac{\lambda_k d_k}{\mu_k}$. We consider two different RaaS rates of $\mu_d = 10$ and $\mu_d = 100$ to compare the RaaS model with the regular system without RaaS, which is shown by $\lambda_S = 0$. The security arrival rate is varied as an integer multiple of mean service rate, i.e. $\lambda_S = n\mu_k, n \in \mathbb{N} = \{1, 2, \cdots\}$. Therefore, the mean number of reconfigurations ($N_r = \frac{\lambda_S}{\mu_k}$) is $n, n \in \mathbb{N}$.

In Fig. 3(left), we compare the blocking parts under RF policy due to the effects of lack of resources (Eq. 9), fragmentation (Eq. 10) and reconfiguration (Eq. 11). As expected, for traffic load range shown here, the blocking is dominated by fragmentation and not by the resource unavailability. Note that in the RaaS model, in addition to fragmentation and resource blocking, we have blocking due to reconfiguration (Eq. 11). Due to this additional reconfiguration blocking, the overall blocking in both SA policies is higher than the regular system, as shown in Fig. 3(right). However, blocking in RaaS significantly reduces, and approaches the blocking of regular system when RaaS rate is increased. The reason is that reconfiguration blocking reduces significantly by increasing RaaS rates from μ_d =10 to 100 (i.e., reducing reconfiguration time $t_{RT} = 1/\mu_d$ from 0.1 to 0.01). We see the similar trend for the FF allocation method, which exhibits much lower blocking than RF method in all scenario.

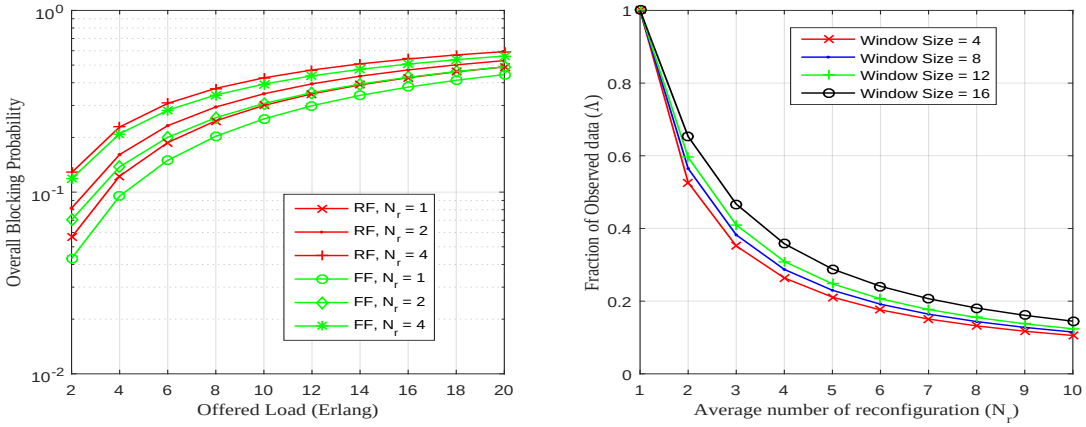


Fig. 4. (left) Overall Blocking in FF and RF policies for various N_r; (right) the fraction of observed data for different observation windows.

In Fig. 4(left), we show the effect of mean number of reconfigurations during the lifetime of a connection ($N_r = 1, 2$, and 4) on the overall blocking for a fixed $\mu_d = 10$ in the RaaS system. As we can see, under both SA policies the overall blocking increases with the increase in N_r. However, again, FF outperforms RF. In Fig. 4(right), we plot the fraction of observed data for RF policy (Eq. (17)) with respect to mean number of reconfigurations (N_r) for different observation windows ($W = \{4, 8, 12, 16\}$) at fixed link load of 18 Erlangs. As expected, with the increase in the mean number of reconfigurations, the fraction of observed data decreases exponentially, which means security also increases exponentially. Although observation window is increased to 80% of the capacity, the security is doubled when only two reconfigurations are performed during the connections' lifetime. When the size of the observation window increases, the fraction of observed data also increases due to the increase in the probability of success (Eq. (15)). Note that when the observation window size is equal to the capacity of the EOL, full spectrum can be observed with probability 1 (Eq. (15)), and the attacker will observe 100% of data.

The most important findings of this paper are: i) irrespective of the RaaS mean reconfiguration time ($t_{RT} = 1/\mu_d$), overall blocking is always higher for both spectrum allocation methods, as compared to the system without reconfiguration; ii) however, if the RaaS mean reconfiguration time is much lower (e.g. $\mu_d = 100$) than the mean holding time t_{H_k}, the overall blocking in RaaS is comparable to the overall blocking of the system without reconfiguration; and iii) the security can be improved by many folds with only a smaller number of reconfigurations. Based on our findings, network operator can tune the RaaS rate and mean number of reconfigurations for a particular load value of the network in order to experience the higher gain from the RaaS.

VI. Conclusion

In this paper, we proposed reconfiguration-as-a-service (RaaS) for an elastic optical networks, where link security can be improved using random reconfiguration scheme. Based on our findings, network operator can tune the RaaS interarrival time ($\sim$ 1/security rate) and the the reconfiguration time for a particular load value of the network to experience the higher gain from the RaaS model. In our model, we added few RaaS states to reconfigure the existing connections, and showed its effect on the overall blocking probability of the system for first-fit and random-fit spectrum allocation methods. More importantly, we showed the correlation among the loads, RaaS rates, the number of reconfigurations, and the fraction of observed data. For future work, we plan to extend our model to a multi-hop EONs. This is no easy task, however, since, again, both criteria of spectrum continuity as well as contiguity must be considered in the analysis which can make it rather complex.

References

[1] M. Jinno *et al.*, "Spectrum-efficient and scalable elastic optical path network: architecture, benefits, and enabling technologies," *Communications Magazine, IEEE*, vol. 47, no. 11, pp. 66–73, 2009.

[2] M. P. Fok, Z. Wang, Y. Deng, and P. R. Prucnal, "Optical layer security in fiber-optic networks," *Information Forensics and Security, IEEE Transactions on*, vol. 6, no. 3, pp. 725–736, 2011.

[3] B. Liu, L. Zhang, X. Xin, and Y. Wang, "Physical layer security in OFDM-PON based on dimension-transformed chaotic permutation," *Photonics Technology Letters, IEEE*, vol. 26, no. 2, pp. 127–130, 2014.

[4] T. Takagi *et al.*, "Disruption minimized spectrum defragmentation in elastic optical path networks that adopt distance adaptive modulation," in *ECOC*. OSA, 2011.

[5] F. Cugini *et al.*, "Push-pull defragmentation without traffic disruption in flexible grid optical networks," *Lightwave Technology, Journal of*, vol. 31, no. 1, pp. 125–133, 2013.

[6] R. Proietti *et al.*, "Rapid and complete hitless defragmentation method using a coherent rx lo with fast wavelength tracking in elastic optical networks," *Optics express*, vol. 20, no. 24, pp. 26 958–26 968, 2012.

[7] K. Christodoulopoulos *et al.*, "Time-varying spectrum allocation policies and blocking analysis in flexible optical networks," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 1, pp. 13–25, 2013.

[8] Y. Yu *et al.*, "The first single-link exact model for performance analysis of flexible grid wdm networks," in *proc. OFC*. OSA, 2013.

[9] H. Beyranvand, M. Maier, and J. Salehi, "An analytical framework for the performance evaluation of node-and network-wise operation scenarios in elastic optical networks," *Communications, IEEE Transactions on*, vol. 62, no. 5, pp. 1621–1633, 2014.

[10] A. Rosa *et al.*, "Statistical analysis of blocking probability and fragmentation based on markov modeling of elastic spectrum allocation on fiber link," *Optics Communications*, vol. 354, pp. 362–373, 2015.

[11] S. K. Singh, W. Bziuk, and A. Jukan, "Defragmentation-as-a-Service (DaaS): How Beneficial is it?" in *proc. OFC*. OSA, 2016.

Ensuring continuous operation of critical process of remote control system at the level of network connectivity

I. Fosić*, D. Budiša **
* HEP Telekomunikacije d.o.o. PS Osijek, Osijek, Croatia, igor.fosic@hep.hr
** HEP ODS d.o.o. Elektroslavonija Osijek, Osijek, Croatia, domagoj.budisa@hep.hr

Abstract - Continuity of critical process is extremely important in the business of each company and depends on many different parameters, such as uninterruptible power supply, application stability, continuous interconnection of devices that are in the system and so on. If a failure occurs in one of the system parameters, there is a downtime of critical process that will result in losses of various types. Continuity of communication can be accomplished in several ways, using redundant cards, devices, communication paths, but they are all based on an uninterrupted power supply. This paper will describe how to improve the communication of network devices that are a part of non-critical systems and have an impact on the operation of critical system. For that purpose, the paper will describe the use of HSRP (Hot Standby Router Protocol) routing protocol feature employed to ensure smooth traffic routing of the critical system to an IP network in case of the main network device failure.

I. INTRODUCTION

Continuity of critical process is extremely important in the business of each company and depends on many different parameters, such as uninterruptible power supply, application stability, continuous interconnection of devices that are in the system and so on. If a failure occurs in one of the system parameters, there is a downtime of critical process that will result in losses of various types. To make the system continuous and uninterrupted, all system parameters must be protected from unwanted downtime. In the remote control system, the most relevant parameters of the system are power supply and mutual communication. Uninterrupted power supply devices usually use a redundant power supply module and uninterruptible power supply system with the same autonomy as UPS (Uninterruptible Power Supply) devices with batteries to provide power supply during the malfunction. Continuity of communication can be accomplished in several ways, using redundant cards, devices, communication paths, but they all are based on an uninterrupted power supply. Therefore, during the system design, more attention is paid to the power supply rather than communication. In complex systems, where critical processes are dependent on non-critical processes and components, the improvement of uncritical parts of the system leads to the desired level of continuity to target critical systems. This paper will describe how to improve communication of network devices that are a part of non-critical systems and have an impact on the operation of critical system. For that purpose, the paper will describe use of HSRP (Hot Standby Router Protocol) routing protocol feature employed to ensure smooth traffic routing of the critical system to an IP network due to the main network device failure.

II. REMOTE CONTROL SYSTEM AS THE CRITICAL SYSTEM OF THE POWER SYSTEM

Global trends, as well as trends in the electrical power industry, represent the transferring of communication onto the IP technology. Accordingly, the process systems are beginning to use devices that are connected to the IP network. IEC 60870-5-104 protocol is set to connect the SCADA (Supervisory Control And Data Acquisition) systems over the IP protocol [1].

Functions of the system for the power system control in the majority of modern dispatching centers can be divided into three groups:

1. SCADA system for the monitoring of the power system with various displays and alarms;

2. EMS (Energy Management System) which entails various calculations and analyses necessary for the management of power system (current state estimation, power flow, N-1 security analysis, optimal power flow, short circuit, etc.);

3. Automatic control systems.

Their concept is generally based on the hierarchy principle. The measurements and signals from transformation substations are transmitted by Remote Terminal Units (RTU) via regional network centers (RC) to the National Dispatch Centre (NDC), which carries out the functions of the power control system. The transmission speed of this data depends on its size as well as the quality of the process and communication equipment [2].

III. LINKING FACILITIES FOR THE COMMUNICATION WITHIN THE REMOTE CONTROL SYSTEM

The basic technology that allows for logical grouping of network users connected to the L2 (layer 2 - Data Link

Layer of the Open Systems Interconnection model) switches is the VLAN (virtual local area network) technology. Each VLAN is a separate logical network that is partitioned and isolated in a computer network at the data link layer, and they can be spread out over a large number of L2 switches, thus enabling communication within the same VLAN between the users on different L2 switches. If several VLANs are defined on the switch, the traffic between different VLANs remains separated, and the users of a particular VLAN can only see the traffic within their own VLAN. This makes the VLAN technology suitable for the separation of the traffic pertaining to a groups of users whose traffic should stay out of other user's sight. A firewall presents a solution for connecting these separate users in different VLANs, through which only a certain number of users can access certain services. When wanting to establish communication between different VLANs, the L3 (layer 3 - Network Layer of the Open Systems Interconnection model) device is necessary, e.g. the L3 switch or router. This technology is reliable and widespread, and applies to Ethernet technology only.

Since the remote control system network is composed of many VLANs, they are integrated into the VPN system via VRF-lite (Virtual Routing and Forwarding) technology. This technology allows for the existence of multiple instances on a single L3 routing device. Routing tables of individual instances are separated and, except for special cases, it is not possible to route traffic from one VRF to another. Thus, a private and independent routing traffic of different VPNs is enabled. The VRF-lite technology enables VPN traffic routing regardless of the address scheme used, making it possible to use the same address scheme within two or more different VPNs. VPN traffic varies based on incoming/outgoing interface/sub-interface. Meanwhile, an interface/sub-interface can only pertain to a single VRF instance, while multiple instances of VRF may be defined in a single L3 device [3].

Since non-critical business systems and the remote control system which is a critical system use the same devices in one segment of network, their functionality affects both systems. Figure 1. shows a simplified situation of the use of network devices in the same environment that is used by non-critical and critical systems. Normally, the primary router is the so-called gateway for hosts and devices to all other locations in the corporate LAN/WAN (local area network/wide area network) network. The backup router used to divert the traffic only if the communication of the primary router with the rest of the LAN/WAN network is interrupted. All non-critical equipment is supplied via central uninterruptible power supply, while all critical system devices are additionally secured with a local uninterruptible power supply. Maintenance and functionality of critical and non-critical systems are divided in different working units; therefore availability of all systems in case of a power failure was not taken

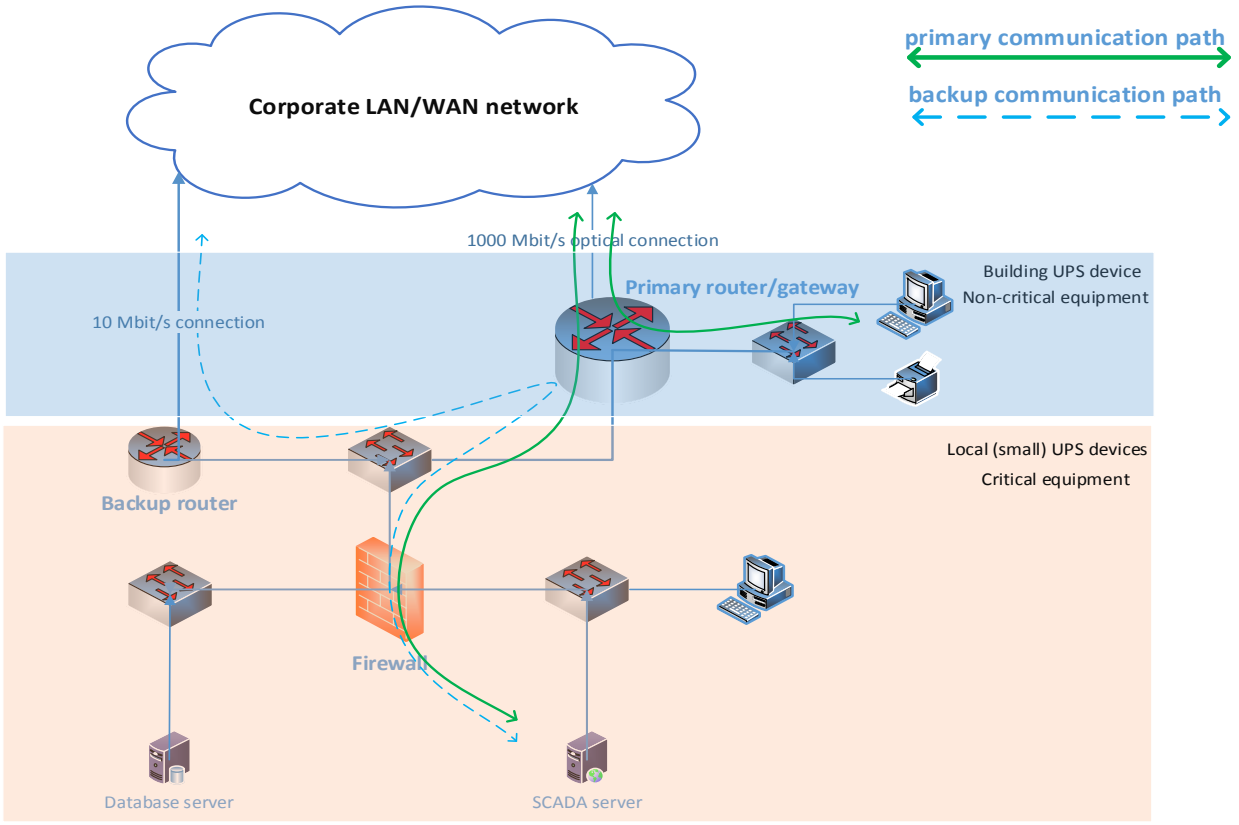


Figure 1. Simplified network equipment scheme of critical and non-critical system without HSRP functionality

into account. The primary router is a device with redundant power supply modules, redundant IP interface cards and redundant supervisor modules. Backup communication path is active only if communication of primary router/gateway lose communication with rest of LAN/WAN network. Then primary router forwards IP packets to the backup router and critical system have connectivity through backup router to the corporate LAN/WAN network. Non-critical system have only primary communication path without backup communication path to ensure enough bandwidth for critical system. When a failure occurred in the central (building) uninterruptible power supply shown in figure 2. marked with red color, the primary router ceased to work and threatened the critical system connectivity, and although all the devices of the critical system continued to work, the access to corporate LAN/WAN network was disabled. The figure 2. shows the status and the inability of the critical system to communicate with other locations in the corporate LAN/ WAN network due to failure of work of the primary router.

Since the backup router was functional, but could not divert the traffic, the solution was to allow a gateway functionality of the backup router for critical system if the issue with the power supply to the primary router reappears. There are several redundant routing protocols which can solve this problem like HSRP (Hot Standby Router Protocol), GLBP (Gateway Load Balancing Protocol) or VRRP (Virtual Redundant Routing Protocol). GLBP protocol is an enhancement over HSRP, in that it offers load sharing by default. VRRP is typically used with non-Cisco devices and it is similar to HSRP [4]. HSRP protocol is described with more detail in chapter IV. of this paper. Reasons for HSRP usage is that no need for load balancing and all network devices are Cisco products. The backup router was enabled with the HSRP functionality and paired with the primary router which is also get HSRP functionality.

IV. BACKUP ROUTER HSRP FUNCTIONALITY CONFIGURATION

In computer networking, the Hot Standby Routing Protocol (HSRP) is a Cisco proprietary redundancy protocol for establishing a fault-tolerant default gateway, and has been described in detail in RFC 2281. [5] The Hot Standby Router Protocol, HSRP, provides a mechanism which is designed to support non-disruptive failover of IP traffic in certain circumstances. In particular, the protocol protects against the failure of the first hop router when the source host cannot learn the IP address of the first hop router dynamically. [6]

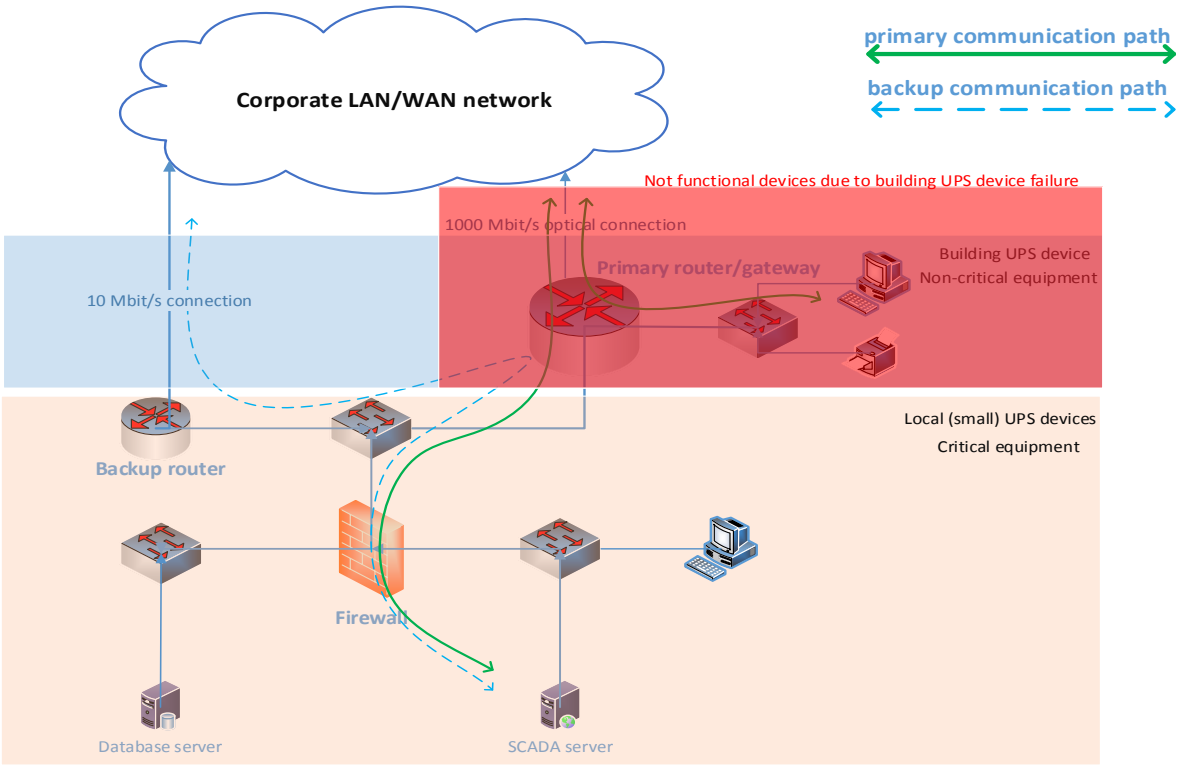


Figure 2. Simplified network equipment scheme of critical and non-critical system without HSRP functionality in the event of the main router failure

TABLE I. HSRP INTERFACES CONFIGURATION

Primary router	Backup router
interface Vlan45 *description SDV VK DUC* *ip vrf forwarding SDV_VK* *ip address 10.224.3.2* *255.255.255.0* *no ip redirects* ***standby 45 ip 10.224.3.1*** ***standby 45 priority 105*** ***standby 45 preempt***	interface FastEthernet0.45 description SDV_VK LAN encapsulation dot1Q 45 ip vrf forwarding SDV_VK ip address 10.224.3.3 255.255.255.0 ***standby 45 ip 10.224.3.1*** ***standby 45 preempt***

By sharing an IP address and a MAC (media access control) layer 2 address, two or more routers can act as a single "virtual" router. The members of the virtual router group continually exchange status messages. This way, one router can assume the routing responsibility of another, should it go out of commission for either planned or unplanned reasons. Hosts continue to forward IP packets to a consistent IP and MAC address, and the changeover of devices doing the routing is transparent. [7]

Here are several parameters for the HSRP functionality configuration of both routers:

- Standby group number
- Virtual MAC address
- Priority
- Authentication Data
- Hellotime
- Holdtime

During the HSRP functionality configuration the basic parameters were used, as shown in Table I, and were applied to the interface used by critical system, while other parameters were set to default values like Hellotime and Holdtime. The values of these parameters can be seen in Table II.

TABLE II. HSRP INTERFACE STATUS

Primary router
Vlan45 - Group 45 **Local state is Active,** priority 105, may preempt Hellotime 3 sec, holdtime 10 sec Next hello sent in 2.250 Virtual IP address is 10.224.3.1 configured **Active router is local** Standby router is 10.224.3.3 expires in 8.564 Virtual mac address is 0000.0c07.ac2d 1 state changes, last state change 20w3d IP redundancy name is "hsrp-Vl45-45" (default)
Backup router
FastEthernet0.45 - Group 45 **State is Standby** 7 state changes, last state change 20w3d Virtual IP address is 10.224.3.1 Active virtual MAC address is 0000.0c07.ac2d Local virtual MAC address is 0000.0c07.ac2d (v1 default) Hello time 3 sec, hold time 10 sec Next hello sent in 2.544 secs Preemption enabled Active router is 10.224.3.2, priority 105 (expires in 9.632 sec) **Standby router is local** Priority 100 (default 100) IP redundancy name is "hsrp-Fa0.45-45" (default)

The primary and backup router configurations differ from one another because the devices feature are different classes and opportunities. The primary router is the L3 switch from Cisco 4500 series with a routing functionality, while the backup router belongs to Cisco 1700 series. Different series of the devices and different IOSs (Internetwork Operating System) allow for version 1 of HSRP, which is not specified in the configuration since it represents the default version of HSRP. The usage of the version 2 of HSRP would further improve the performance, since according to [8], HSRP version 2 features are as follows:

- advertises and learns millisecond timer values. This change ensures stability of the HSRP groups in all cases.
- expands the group number range from 0 to 4095.
- provides improved management and troubleshooting. With HSRP version 1, you cannot use HSRP active hello messages to identify which physical device sent the message because the source MAC address is the HSRP virtual MAC address.
- packet format includes a 6-byte identifier field that is used to uniquely identify the sender of the message. Typically, this field is populated with the interface MAC address.
- uses the new IP multicast address 224.0.0.102 to send hello packets instead of the multicast address of 224.0.0.2, used by HSRP version 1. This new multicast address allows CGMP leave processing to be enabled at the same time as HSRP.

The configuration shows that the so-called common IP address is 10.224.3.1, representing the gateway IP address for the devices in that particular segment of the network and critical system. Additional parameters for the functioning of the HSRP are the standby group number which is assigned a value of 45, and the priority which has a value of 105 at the primary router, while with the backup router the parameter is omitted, bearing the value of 100 which is the default value unless indicated otherwise. The preempt option allows for a device with a higher priority parameter value to be chosen as the primary router, i.e. that is the common IP address 10.224.3.1 is given to the primary router. If there is a change in the status of the primary router, the HSRP will ensure that the backup router is given the 10.224.3.1 gateway address.

During the testing of HSRP configuration on the network and critical system for errors, results were obtained which enable the critical system function to be unhindered and without interruption, i.e. with a reasonable interruption duration, if a problem occurs with the primary router. The Figure 3. shows the target environment that satisfies the continuity of the operation of critical systems featuring the occurrence of issues with the primary router. The Table III. shows the time when an interruption occurred, as well as the time of the communication re-establishment, which are obviously within the set time limits. By simulating a

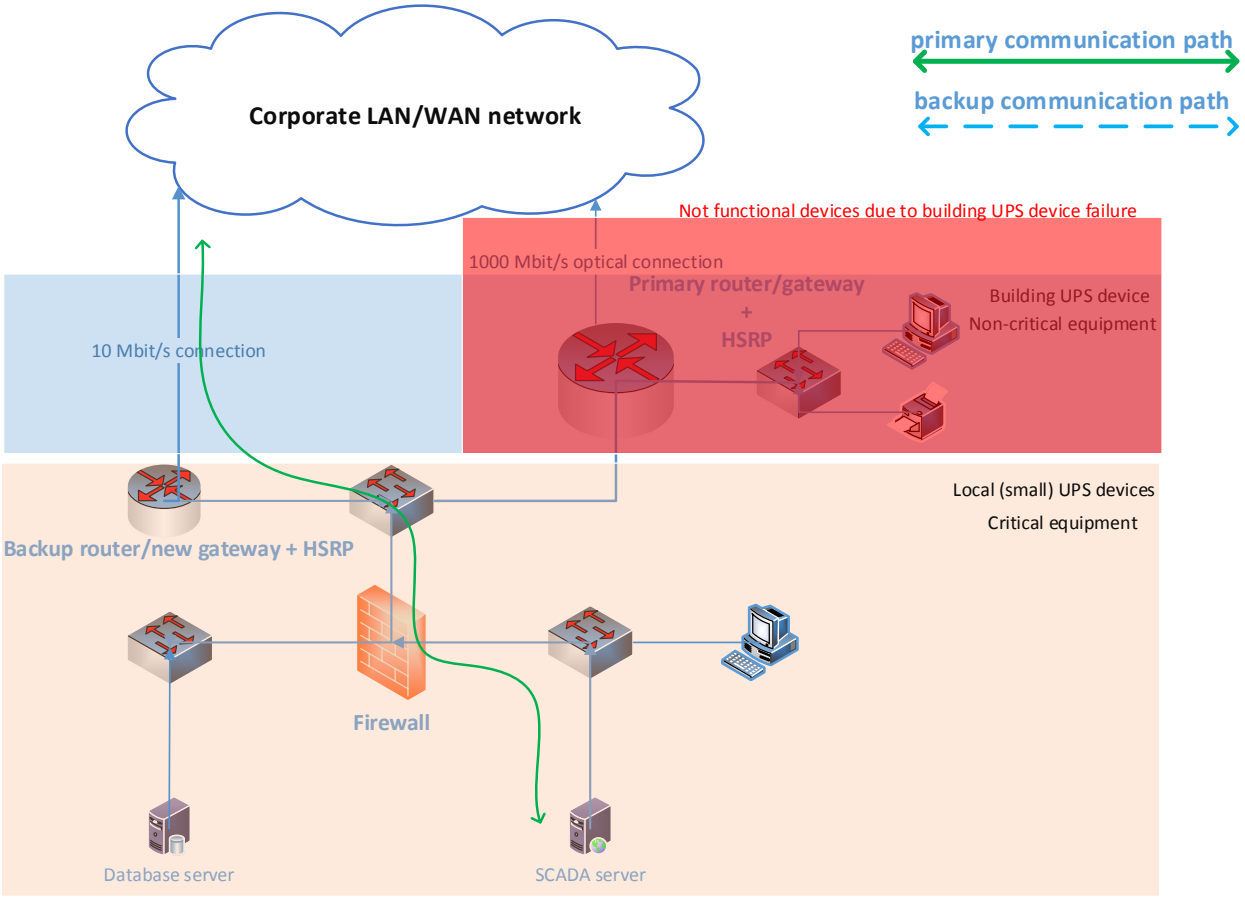


Figure 3. Simplified network equipment scheme of critical and non-critical systems with HSRP functionality of main and backup router

failure of the primary router in a controlled environment, the following results were obtained:

The duration of the routing protocol transfer on the standby HSRP router is negligible as can be read from the backup router log.

Feb 4 08:32:32.458: HSRP: Fa0.45 Grp 45 Standby: c/Active timer expired (10.224.3.2)

Feb 4 08:32:32.458: HSRP: Fa0.45 Grp 45 Active router is local, was 10.224.3.2

Feb 4 08:32:32.458: HSRP: Fa0.45 Grp 45 Standby router is unknown, was local

Feb 4 08:32:32.458: HSRP: Fa0.45 Grp 45 Standby -> Active

Feb 4 08:32:32.458: HSRP: Fa0.45 Grp 45 Redundancy "hsrp-Fa0.45-45" state Standby -> Active

Feb 4 08:32:32.626: HSRP: Use HSRP virtual address 10.224.3.1 as ICMP src

As for the users, interruptions have been noted referring to devices in remote locations. The transition state has lasted for a very short time and within acceptable limits. The Figure 3. shows the targeted state and HSRP functionality of network equipment in case of primary router issues. All communication is transferred onto the backup router and very soon everything was functional in regard to the users. From a total of 37 devices in remote locations

TABLE III. INTERRUPTION DURATION OF REMOTE LOCATIONS

STATION	OBJECT	SYSTIME	TEXT
NONE	RTUZUPANJA1	4.2.2016 9:31	Not Ready
NONE	RTUFARMASLATINE	4.2.2016 9:31	Not Ready
NONE	RTUTOVARNIK	4.2.2016 9:31	Not Ready
NONE	RTUVUKOVAR3	4.2.2016 9:31	Not Ready
NONE	RTUZUPANJA1	4.2.2016 9:31	Ready
NONE	RTUFARMASLATINE	4.2.2016 9:31	Ready
NONE	RTUVUKOVAR3	4.2.2016 9:31	Ready
NONE	RTUTOVARNIK	4.2.2016 9:31	Ready
NONE	RTUNIJEMCI	4.2.2016 9:32	Not Ready
NONE	RTUNIJEMCI	4.2.2016 9:32	Ready
NONE	RTUDELETOVCI	4.2.2016 9:32	Not Ready
NONE	RTUBOROVONASELJE	4.2.2016 9:32	Not Ready
NONE	RTUNIJEMCINOVA	4.2.2016 9:32	Not Ready
NONE	RTUZUPANJA2	4.2.2016 9:32	Not Ready
NONE	RTUZUPANJA2	4.2.2016 9:32	Ready
NONE	RTUNIJEMCINOVA	4.2.2016 9:32	Ready
NONE	RTUBOROVONASELJE	4.2.2016 9:32	Ready
NONE	RTUDELETOVCI	4.2.2016 9:32	Ready

representing a part of the critical system, 10 of them (27%) suffered an interruption in duration of less than 1 minute.

Although the HSRP process lasted briefly, the convergence in the OSPF (Open Short Path First) routing protocol and communication of the application in the central location with the devices at remote locations resulted in the detection of a failure in communication. Although interruptions were recorded, they were all within acceptable limits.

V. CONCLUSION

Very often during design and sizing equipment of critical systems do not take into account all possible issues. More attention is focused on the well-known problems, such as power redundancy and redundant components. Incoherence and unequal levels of availability of common components in different systems can reflect on the availability of critical systems. With no requirements for traffic load balancing and with fact that all network devices are Cisco products the HSRP router functionality proved to be a solution which meet set requirements. There were no need for new equipment, no extra cost only better and new configuration of existing devices. The room for improvement of this solution is presented in the fact that the HSRP's upgrade to version 2 which would further improve the system performance. Consequently, an upgrade of the software version of the primary router should also take place. The solution provided has fully meet the requirements of continuous availability of critical systems in regard to users' needs.

REFERENCES

[1] B. Turk, R. Bobanović, D. Janković, Povezivanje EE objekata multiservisnim IP mrežnim sustavima, Mipro 2012.

[2] I. Pavić, Napredne funkcije sustava za vođenje EES-a, Mipro 2011.

[3] B.Turk, Ž.Kovač, Povezivanje SDV sustava putem višeuslužne širokopojasne HEP-ove IP mreže, Cigre 2008.

[4] http://www.infocellar.com/networks/Routers/HSRP-GLBP-VRRP.htm

[5] https://en.wikipedia.org/wiki/Hot_Standby_Router_Protocol

[6] https://tools.ietf.org/html/rfc2281

[7] http://www.cisco.com/c/en/us/support/docs/ip/hot-standby-router-protocol-hsrp/9234-hsrpguidetoc.html

[8] http://www.cisco.com/c/en/us/td/docs/ios-xml/ios/ipapp_fhrp/configuration/xe-3s/fhp-xe-3s-book/fhp-hsrp-v2.html

IoT PLATFORM AND APPLICATIONS

Requirements and Challenges in Wireless Network's Performance Evaluation in Ambient Assisted Living Environments

Ana Koren, Dina Šimunić
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000
Zagreb, Croatia
E-mail(s): {ana.koren, dina.simunic} @fer.hr

***Abstract*— Ambient Assisted Living environments rely heavily on use of technology for providing its users with assistance or supervision in their Activities of Daily Living (ADLs). Thus, it is crucial to ensure that wireless sensor networks (WSN) have high reliability and availability in such systems. Performance of networks is measured by the packet delivery ratio, average end-to-end delay and power consumption in order to ensure QoS, especially for time sensitive traffic. During the past thirty years of research, a colossal number of MAC protocols have been developed, differing by the type of media they use as well as the performance requirements for which they have been optimized. Design of MAC layer has a major role in energy efficiency of the wireless sensor network as energy efficient protocols improve the lifetime of the network. This paper explores different protocol implementations for WSN with the goal of achieving best possible QoS in context of Ambient Assisted Living environments.**

Keywords*— *Ambient Assisted Living; eWALL; Wireless Sensor Networks; Person-Centered Approach; MAC layer; Energy Efficiency

I. INTRODUCTION

The sensor node spends most of its time in sleep mode in which the power consumption is of a much lower level then in the active state. Hence, applying an optimized sleep and wake-up mechanism and appropriate power control scheme would greatly impact the total power consumption of the node. Furthermore, the number of transmitted data packets can be minimized to avoid redundant and converged data; one possible solution being fusing different data with high correlation into a single value. Packet drops, losses and interference of nodes obstructs the communications and causes retransmission of packets so it is necessary to ensure these issues don't drain the power of nodes by avoiding them as much as possible. Design of MAC layer has a major role in energy efficiency of the wireless sensor network as energy efficient protocols improve the lifetime of the network. For example, using Time Division Multiple Access (TDMA) and polling channel access mechanisms on MAC layer, deterministic packet loss and packed delay can be attained. On the other hand, Carrier-Sense Multiple Access with Collision Avoidance (CSMA/CA) can be used at MAC layer for dynamic allocation of transmission channel to the end-nodes. eWALL (project no. 610658: "eWALL for Active Long Living") is a framework consisting of the sensing environment which collectively realizes the paradigm of Ambient Intelligence, providing long-term care for the elderly and assistance to their caregivers.

II. RELATED RESEARCH

Medium Access Control (MAC) protocols coordinate nodes' access to a shared communication medium. During the past thirty years of research, a colossal number of MAC protocols have been developed, differing by the type of media they use as well as the performance requirements for which they have been optimized. MAC layer design plays a pivotal role for optimizing energy efficiency in wireless sensor networks, as the MAC protocol controls directly the radio of the nodes. For MAC protocol design, several attributes need to be considered; energy efficiency, effective collision avoidance, scalability, and adaptability to changes of network topology (i.e. density and the number of nodes) being the most important. Other attributes include channel utilization, latency, throughput, and fairness. However, different attributes require trade-off or compromise; for instance, in contention-based protocols like CSMD, adaptive sleep cycles can improve energy efficiency at the cost of increased packet drop and latency. MAC protocols for wireless sensor networks can generally be divided into those based on on-demand wake-up (contention-based) and those based on scheduled wake-up (scheduling-based). Each can have different strengths and weaknesses, depending on the network topology or the applications. In contention-based protocols, the node that wants to send data occupies the channel through the competition method, without pre-coordination with other nodes. These protocols also require a form of signaling the nodes to wake up, usually through a separate radio (which implies higher costs), while in the case of a scheduled wake up, the nodes are assigned time slots in which they need to transmission the data. MAC protocols based on scheduled wake-ups can be synchronous and

asynchronous. While asynchronous protocols are easier to implement than synchronous ones (as they don't require periodic synchronization), they are not as efficient. The most commonly used technique for synchronous scheduled wake-up protocols is TDMA. Since sensor nodes consume more power when in active state (sensing and transmitting) then when in sleep mode, significant amount of energy can be saved by carefully optimizing the sleep cycles of the nodes, leveraging between lower energy consumption and degraded delay performance (which occurs when nodes are put to sleep). Some of the solutions for the MAC layer have been presented in [1], [2], [3], [4], [5], and [6]. Scalable and efficient Medium Access Control (MAC) protocol is crucial for maximizing the performance of wireless sensor networks [2]. Furthermore, Low-Energy Adaptive Clustering Hierarchy (LEACH) [7] is a self-organizing, adaptive, clustering-based protocol that uses randomized rotation of cluster-heads to evenly distribute the energy load among the sensor nodes in the network [6].

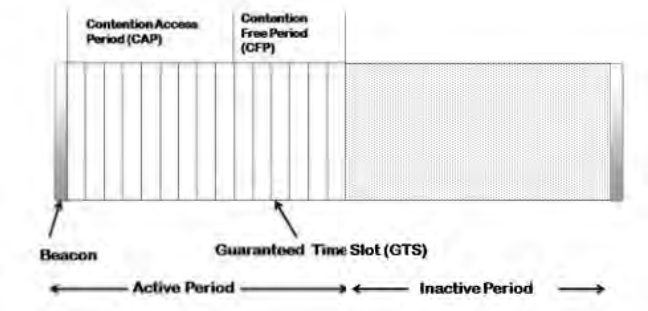


Figure 1. IEEE 802.15.4 Superframe structure [8]

Lastly, the MAC protocol should provide collision free communication between the sensor nodes in such a way which increases the life time of the network, reduces energy consumption and minimizes delay [8]. In a home sensing environment, some of the information collected is time-sensitive. Hence, there's the need for Guaranteed Time Slots (GTS). To transmit through a GTS, sensor node needs to wait for one beacon interval [9]. This is shown in Figure 1. A thorough overview and performance analysis of a IEEE 802.15.4 for medium access protocol non beacon-enabled mode which uses unslotted carrier-sense multiple access with collision avoidance (CSMA/CA) is given in [10]. Similarly, a mathematical model for the beacon-enabled mode of the IEEE 802.15.4 medium-access control (MAC) protocol is provided in [11].

III. Simulation

Ambient Assisted Living environments collect various types of data, many of which are time-sensitive and, as such, require a Guaranteed Time Slots (GTS) for transmission. Thus, beacon-enabled mode has been used in the following model on which the simulations have been executed. The most significant metrics of a WSN performance are packet delivery ratio, power consumption, and average end-to-end delay [12]. Wireless sensor network (shown in Figure 2) consists of one hundred equally spaced static nodes with a coordinator node in the center.

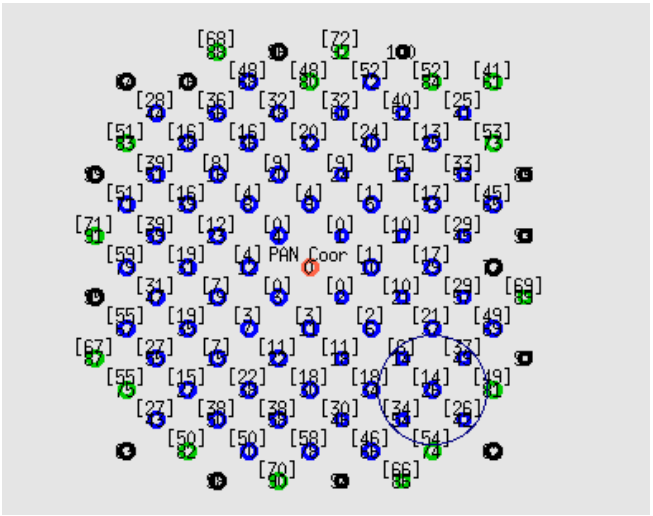


Figure 2. Topology of wireless sensor network used in simulation consisting of one hundred nodes (numbered) and a coordinator

The full characteristics of ZigBee nodes used are given in the table below (Table 1).

Table 1. Characteristics of ZigBee sensor nodes used in simulations

Channel type	Channel/WirelessChannel
Radio-propagation model	Propagation/TwoRayGround
Network interface type	Phy/WirelessPhy/802_15_4
MAC type	Mac/802_15_4
Interface queue type	Queue/DropTail/PriQueue
Antenna model	Antenna/OmniAntenna
Max packet in IFQ (Interface priority queue)	50
Routing protocol	AODV (Ad hoc On-Demand Distance Vector Routing)

Simulation was executed using ZigBee routing (ZBR) and peer-to-peer, beacon enabled topology and lasted one hundred seconds. Results in Figure 3 show the ratio of sent, received and dropped packets. Number of packets sent is 33734, received 31032, and dropped 2702, making the packet drop rate 8.7% which is within the acceptable threshold of 15% – everything above that rate noticeably deteriorates the

availability of the network as the MAC layer retransmissions cannot compensate fast enough [13].

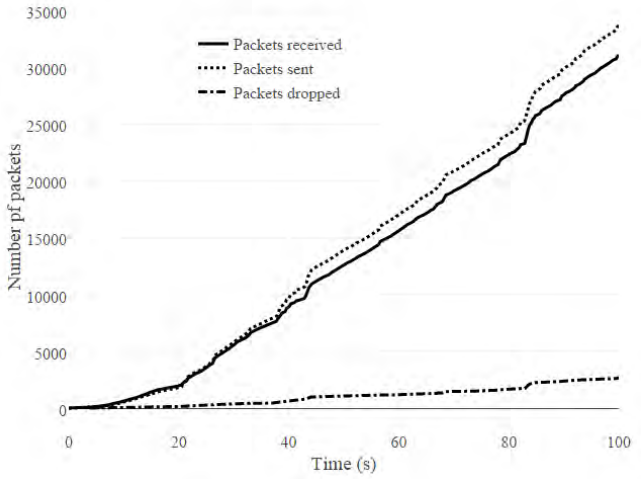


Figure 3. Packets sent, received and dropped

Breakdown of the cause of the packets dropped is given in the Table 2 below.

Table 2. Cause of packet drops

Number of packets dropped	Percentage of packets dropped	Cause of packet drop
18	0.7%	ARP (dropped by ARP)
22	0.8%	BSY (drop due to medium being busy)
72	2.7%	CBK (drop RTR MAC callback)
23	0.9%	DUP (drop MAC duplicate)
65	2.4%	IFQ (drop due to no buffer space in queue)
2478	91.7%	LQI (Link Quality Indication packet drop)
14	0.5%	NRTE (dropped due to no route being available)
10	0.4%	TTL (dropped because time to live has reached zero)

The vast majority of packets (91.7%) have been dropped due to the link quality at MAC layer (LQI). This happens when the link quality of the packets arriving at the MAC layer is not within acceptable threshold [13] which results in packets getting dropped. LQI in ZigBee networks is a computed value, based on the received signal strength as well as the number of errors received and it indicates the strength of the communications link. Furthermore, the distribution of LQI-caused packet drops is even throughout the simulation, as shown in Figure 4.

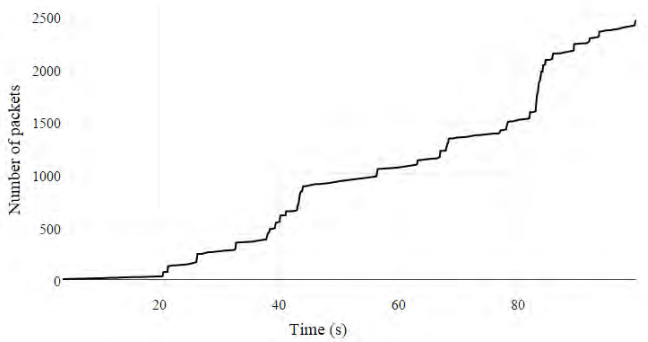


Figure 4. LQI-caused packet drops

Thus, routing protocols based on link quality are needed. Using Link Quality Estimation (LQE) algorithms to estimate the link quality provides the necessary information to evaluate and choose the best path available. However, many of the existing routing methods are strongly influenced by estimation errors and, as such, are non-applicable in large-scale networks [14] and should be further explored in the future since it's essential for an energy-efficient WSN. In many wireless network scenarios rapid variation in the channel hinders the solution of an efficient mechanism for estimating the link quality at the time of transmission [15]. However, link quality tracking is crucial as it enables the routing layer to select links that would maximize the packet delivery rate and minimize traffic congestion [16] and energy consumption. This proves to be a challenge, especially in Ambient Assisted Living environments which include wearable sensors, due to the networks being dynamic, heterogeneous, and consisting of interconnected miniaturized nodes with sensor or actuator capabilities in, on, or around a human body [17]. Mobility is a key factor as the human body may be experiencing posture changes and is moving around in the environment [17]. With the goal of maximizing the overall network's performance for Wireless Body Area Networks (WBANs), in paper [18] a link-quality-aware resource allocation scheme is proposed, while in [19] new protocol which would exploit channel diversity is proposed. Ultimately, the solution for delivery of packets in wireless sensor network in an AAL environments should be reliable and scalable.

IV. CONCLUSION

The most challenging task in designing a MAC protocol for Ambient Assisted Living (AAL) environments is satisfying different needs of different sensor nodes generating varying types of traffic. This often means trade-offs between energy-efficiency and QoS are necessary. This paper gives comparison and overview of QoS and performance of wireless sensor networks and evaluates them in the context of an Ambient Assisted Living environments. Results of the simulation show that major part of the packets dropped in such networks are dropped due to the link quality at MAC layer (91.7% in the simulated scenario). This means the link quality of the packets arriving at the MAC layer wasn't within the acceptable threshold. In spite of all the solutions presented so far, this continues to be a challenge especially for mobile wireless body area networks (WBANs) as well as large-scale WSNs, both of which are applicable for AAL environments. The priority should be to develop a protocol for WSN which would enable creating a comprehensive system that would offer easily accessible and affordable continuous services to its users. The solution should be reliable, scalable and take into account different types of coexisting sensors in AAL and their mobility, which will be further explored in future work.

Acknowledgement - We thank our colleagues from the eWALL project who provided insight and expertise that greatly assisted the work presented in the paper.

REFERENCES

[1] Yanping Yu, „A Cross-layer Approach for Energy-efficient MAC Layer in Wireless Sensor Networks“, 2008 4th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '08), October 2008, pp. 1-3.

[2] Razaque A., Elleithy K., "Scalable and Energy-Efficient Medium Access Control Protocol for Wireless Sensor Networks", Systems, Applications and Technology Conference (LISAT), 2015 IEEE Long Island, May 2015, pp. 1-4.

[3] Yang G., Zhang G., „An Energy-Efficient Cross-Layer MAC for Wireless Sensor Networks“, 2011 International Conference on Electronic and Mechanical Engineering and Information Technology, August 2011, pp. 683-686.

[4] Attiah A., Akbas. M. I., Chatterjee M., Turgut D., „EE-MAC: Energy-Efficient Sensor MAC Layer Protocol“, 9th IEEE International Workshop on Peformance and Management of Wireless and Mobile Networks 2013, October 2013, pp. 116-119.

[5] Singhal D., Garimella R. M., „Energy-Efficient Cognitive Cross-Layer MAC Protocol“, 2015 International Conference on Advances in Computing, Communications and informatics (ICACCI), August 2015, pp. 987-992.

[6] Ehsan S., Hamdaoui B., "A Survey on Energy-Efficient Routing Techniques with QoS Assurances for Wireless Multimedia Sensor Networks", IEEE 2012 Communications Surveys & Tutorials (Vol. 12, Iss. 2), 2012, pp. 265-278.

[7] Heinzelman W., Chandrakasan A., Balakrishnan H., "Energy-Efficient Communication Protocol for Wireless Microsensor Networks", Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, January 2000, pp. 10.

[8] Sehgal N. K., Challa R. K., "Performance Evaluation of S-MAC Protocol", Proc. of the International Conference on Wireless Networks and Embedded Systems, October 2008, pp. 45-48.

[9] Puri T., Challa R. K., Sehgal N. K., "Energy Efficient QoS Aware MAC Layer Time Slot Allocation Scheme for WBASN", 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), August 2015, pp. 966-972.

[10] Buratti C., Verdone R., "Performance Analysis of IEEE 802.15.4 Non Beacon-Enabled Mode", IEEE Transactions on Vehicular Technology (Vol. 58, Iss. 7), September 2009, pp. 3480-3493.

[11] Buratti C., Verdone R., "Performance Analysis of IEEE 802.15.4 Beacon-Enabled Mode", ", IEEE Transactions on Vehicular Technology (Vol. 59, Iss. 4), May 2010, pp. 2031-2045.

[12] Kim C., Talipov E., Ahn B., "Reverse AODV (R-AODV) Routing Protocol in Mobile Ad hoc Networks", The 2006 IFIP International Conference on Embedded and Ubiquitous Computing (EUC'06), LNCS 4097, August 2006., pp. 522-531.

[13] Suhonen J., Hämäläinen T. D., Hännikäinen M., "Availability and End-to-end Reliability in Low Duty Cycle Multihop Wireless Sensor Networks", Sensors 2009 (Vol. 9, Iss. 3), 2009, pp. 2088–2116.

[14] Krogmann M., Tian T., Stromberg G., Heidrich M., Heumer M., "Impact of Link Quality Estimation Errors on Routing Metrics for Wireless Sensor Networks", 2009 5th International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), December 2009, pp. 397-402.

[15] Lal D., Herrmann F., Manjeshvar A., Uysal-Biyikoglu E., Keshavarzian A., "Measurement and Characterization of Link Quality Metrics in Energy Constrained Wireless Sensor Networks", IEEE Global Telecommunications Conference 2003 (GLOBECOM '03. IEEE, Vol. 1), December 2003, pp. 446-452.

[16] Millan P., Molina C., Medina E., Vega D., Meseguer R., Braem B., Blondia C., "Tracking and Predicting Link Quality in Wireless Community Networks", 2014 IEEE 10th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), October 2014, pp. 239-244.

[17] Omer S., Vesilo R., Dutkiewicz E., Zhang Q., "An LQI Based Dual-Channel Routing Protocol for Wireless Body Area Networks", 2015 International Telecommunication Networks and Applications Conference (ITNAC), November 2015, pp. 320-325.

[18] Samanta A., Bera S., Misra S., "Link-Quality-Aware Resource Allocation With Load Balance in Wireless Body Area Networks", IEEE Systems Journal (Vol. PP, Iss. 99), August 2015, pp. 1-8.

[19] Doddavenkatappa M., Chan M. C., Leong B., "Improving Link Quality by Exploiting Channel Diversity in Wireless Sensor Networks", 2011 IEEE 32nd Real-Time Systems Symposium (RTSS), December 2011, pp. 159-169.

[20] Holger K., Wilig A., „Protocols and Architectures for Wireless Sensor Networks“, John Wiley & Sons, 2005.

[21] Feng Xia, Azizur Rahim, „MAC Protocols for Cyber-Physical Systems“, Springer, 2015.

Long Term Evolution as a Precondition for Internet of Postal Things

Dr.sc. Amel Kosovac
Faculty of Traffic and Communications
University of Sarajevo
amelkosovac@gmail.com

Aida Veispahić
UAS Technikum Wien
Wien, Österreich
ti14m036@technikum-wien.at

Mr.sc Mirza Berković
Faculty of Traffic and Communications
University of Sarajevo
mirza.berkovic@gmail.com

***Abstract* – Wireless technology is incessantly used, from handheld devices up to industrial environment that further leads to global interconnectivity. LTE enables real time communication with high reliability, robustness and higher security. HD voice transmission, Machine-to-Machine communications and Internet of Things are only some of domains provided by LTE network nowadays. Such technology delivers connectivity known as Internet of Everything, which will make the world more suitable place. LTE can also be key enabler for Internet of Postal (IoPT) which is attracting growing attention on a global world market and in IT world.**

Keywords – LTE, LTE-Advanced, Machine-to-Machine, Internet of Things, Internet of Postal

I. INTRODUCTION

People use the cellular network in everyday life, whether for checking e-mails, sending messages, playing games or for other tasks such as video conferences, video calls, online banking, education purposes. Whereas smart devices became ubiquitous in the world usage, a more stable and faster network is required. On this stage LTE and its successors came as a response which will change the usage of cellular network in the way as it never had before.

Nowadays, LTE has been implemented from 318 operators in 111 countries [1]. LTE and foreseen 5G network are also one of the main enabling technologies and future of Machine-to-Machine (M2M) communications, Internet of Things (IoT) and Internet of Postal (IoPT) which all together lead to Internet of Everything by 2020.

These technologies will enable industrial growth, vehicular applications, M2M technologies, IoT, IoPT and facilitate various aspects of everyday life. For readers to fathom this paper, an inquiry is set: what will this developments bring forth to the end users?

II. CURRENT MARKET STATUS

LTE (Long Term Evolution) is a cellular network technology which assures higher data rates, high spectral efficiency, and flexibility in frequency and bandwidth, which are demanded from the vast usage of data services. Comparing to its predecessors, 2G and 3G networks respectively, the LTE is purely IP based network, based on packet switching. Therefore some additional frequency spectrum needed for more ubiquitous and fast network is achieved. This technology gives an evolution in network architecture, but it is complete new system behind it.
It has been forecasted that in 2015 cellular subscriptions will exceed worldwide population, what is the indication how far a deployment of LTE went.
It is saying that there are countries which have more mobile subscribers than population. Regardly, this jump in portable devices usage issued network development, where LTE and LTE-Advanced stood on the scene.

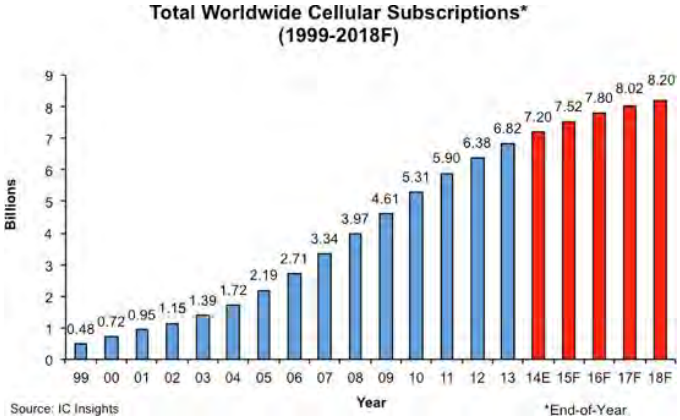


Figure 1 – LTE market state analyses [2]

One of the reason why LTE attracts so much attention today is fact that it is the enabler for IoT solutions. By some preditions by 2020 aroud 30 billion connected devices will be deployed on Internet. In order to fully support tracking services for logistics and transport, to be used in smart metering, cities, homes, health or in industry IoT solutions need fast, optimized, flexible technology that is enabled through LTE.

III. LTE NETWORK

LTE represents a complete revolution in the mobile network, for the new architecture is used and several advancements had been made according to its predecessors. These advancements are mentioned later in the LTE-Advanced chapter. Thus, this transitional phase creates enormous difficulties to telecom operators, mostly economical ones. One of the important is WiMAX (Worldwide Interoperability for Microwave Access) to LTE migration that opens access to a larger market, more network vendors and roaming opportunities, for which users have high demands. Nevertheless, this transition will not pass harmless, for the legacy network integration with latter, a Radio Access Network (RAN) upgrade will require some improvements to match technologies. Apart from challenges and difficulties, LTE and WiMAX have very similar RAN and Core network that will make this transition facilitate.

A. NETWORK IMPLEMENTATION QUESTIONS

Major obstacles of the LTE implementation are related to [3]:

- Technical implementations which are in relation with division and use of shared radio spectrum.
- Business and financial implications are connected with voice transmission. Circuit switched networks are used for the voice transmission and still, they are main source of incomes in the mobile networks across the world. For that, operators are under considerable pressure to develop commercial business models for LTE and 4G data services that will cover their costs and provide a reasonable return of investments.
- Security had always been a main key in the mobile communications. While providing fast and reliable data traffic to end users, operators need to provide impenetrable security for that case.

B. LTE BENEFITS

Practicality, additional researches and development of LTE should further improve the characteristics of 4G network. It is therefore the most efficient and fastest wireless mobile connection which is now commercially used. All sent and received data is transmitted as IP packets independently of the type of service.

Compared to its predecessors, the biggest advantages which LTE offers are [4]:

- Reduced costs per bit transferred information,
- High-speed transmission with a short delay means 100 Mbit/s in download and 50 Mbit/s in upload with a significant reduction in waiting time (latency) due to performance improvements of protocols in the higher layers (e.g. TCP),
- Improved provision of a range of services - ability to retain of existing database of end-users with smooth replacement from old to new services,
- Simplified architecture,
- Improved coverage - high transmission rates by covering wide area,
- Increased system capacity - which means increasing the number of users who can access to high - speed Internet,
- Higher energy efficiency, lower energy consumption of terminals and infrastructure,
- Smooth switching over heterogeneous networks,
- Interoperability with existing wireless standards,
- A fully IP-oriented, packet-switched network (all-IP).

IV. LTE - ADVANCED

LTE-Advanced system is considered like the next big step in the evolution of LTE. It is improved version of LTE 3GPP Release 10 [5], which supports:

- Flow rate of 1000 Mbit/s (1Gbit/s) in downlink and is 500 Mbit/s in uplink,
- Aggregation of the LTE Release 8 channels of 20 MHz,
- The frequency spectrum aggregation,
- 4 × 4 MIMO in uplink and 8 × 8 MIMO in the downlink.

One of the basic requirements of LTE-Advanced is that this standard should be the evolutionary upgrade of LTE. Primarily, it refers to the possibility of implementing LTE-Advanced system in the frequency range with already implemented LTE system, i.e. spectrum sharing without mutual interference. A direct consequence of this is related to the future of LTE-Advanced. Therefore cells must be able to deliver an LTE terminal from Release 8. In addition, there must be an option to upgrade already installed LTE infrastructure equipment with the LTE-Advanced one, within an affordable price.

The main features of LTE-Advanced standards are:

- Achieve speeds up to 1Gb/sec in downlink and 500 Mb/sec in the uplink,
- Terminal and infrastructure implementation with an affordable prices,

Major improvements to system performance is in terms of throughput and cell users. The key features of LTE-A are modulation techniques (OFDM for Downlink and SC-FDMA for Uplink) and antenna system, Multiple Input – Multiple Output (MIMO). Aforementioned key technologies are the basic precondition for LTE. Within this it is possible to avoid mutual interference which usually occurs between close spaced carriers. MIMO, spatial multiplexing and beamforming enable LTE-A to achieve this high transmission speeds.

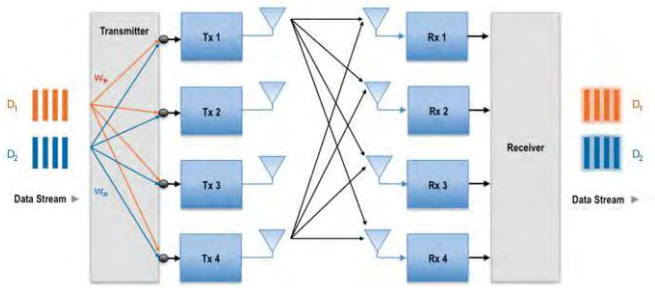


Figure 2 - MIMO 4x4 (Spatial Multiplexing) [6]

MIMO provides a way of utilizing the multiple signal paths that exist between a transmitter and receiver to significantly improve the data throughput available on a given channel with its defined bandwidth. By using multiple antennas at the transmitter and receiver (transceiver) along with some complex digital signal processing, such as m-QAM modulation, MIMO technology enables the system to set up multiple data streams on the same channel, thereby increasing the data capacity of a channel.

V. LTE AS A PART OF INTERNET OF EVERYTHING

Incessant technology growth caused ubiquitous devices connected to Internet. From industrial machines, vehicular applications, home appliances, embedded systems and other utilities, one can have real-time information from various fields. Apart from that some of these technologies could make communication processes to transfer different information and being even applied without human intervention.

Two domains are considered in the aspect of Internet of Everything (IoE):

- M2M communications: LTE is of big importance in this field. All communications which traverses through long distances demand high-reliable/low cost

option, such as bandwidth, quality of service (QoS) and low-latency.

- LTE satisfies more of demands for deployment in M2M domain, where some examples are mobile routers, remote surveillance systems, facial recognition systems, digital signage, vending, point-of-sales and tele-health terminals [7].
- Internet of Things (IoT): Apart from M2M, Internet connected devices have led to Internet of Things, that is driven by demands of industries, organizations, operators and customers. It is ubiquitous challenge to make all things connected regarding to architecture integration, but nowadays engineers make this possible. IoT enables convenient usage of our smartphones, such as controlling home appliances, air conditions, lightning and various insatiable demands through LTE network.
- Internet of Postal Things (IoPT): Beside M2M and IoT, it is important to mention and IoPT. IoPT can be considered as a new method that will transform not only the postal infrastructure, but also non-postal sectors. Providing not just information to postal vehicles, mailboxes, carriers, IoPT can also send collected environment data, e.g. to the local governments, thanks to sensor equipment. It will for sure improve overall performance, increase the customer services, has a huge impact on new products creation and services and provide support and optimize decision-making processes.



Figure 3 - Cloud computing and LTE as complementary technologies [7]

VI. THE INTERNET OF POSTAL - IoPT

Smart devices equipped with sensors could collect the data about the temperature, location, performance, motion, tracking, offer better management of mailpieces and in general make a big improvement of system processes and insure development of „smart“ products and services. Applications applied and developed for postal services unified under the IoPT are [8]:

- Transportation and logistics – smart sensors on vehicles could affect with better routing proposals, shorter delivery time, vehicle operations calculations (fuel, tire pressure, engine performance, brake conditions), location tracking.
- Enhanced mail and parcel services – using the smart sensor devices delivery and visibility of the mailpiece could be improved. Use of smart mailboxes could generate new revenue opportunities. Sensors could collect data about the temperature, mail delivery or pick-up time. Through app will be also possible to administer smart box locking mechanism, improve security and reduce operational cost.
- Smart post office – IoPT here can affect with less time waiting in the lines, reducing the number and duration of tasks, easily identification of parcels and their availabilty for pick-up, 3D calculations (weight and dimensions) of mailpieces.
- Smarter mail processing center – IoPT sensors can locate the objects in storages, provide information about items clerks need to process, prevent damages or idle time, replace manually track for shipping.
- Neighborhood services – using postal equipment, vehicles and know-how support for „smart-city“ inititives can be provided with value-added services especially for disabled persons.

Implementation of IoPT will need strategic adoption of existing technologies which will provide a careful prioritization of applications and processes. After involvement of comatible technologies, posts should work on digitization of processes and services and implementation of open information platform with included security, privacy, regulatory and interoperability requirements.

VIII. LTE AS AN ENABLER FOR IoT AND ITS LIMITATIONS IN COMPARISON TO 5G

2G was a standrad showed up in early 90s. It changed the way people used to talk before. Here for the first time it is possible to speak about digital cellular networks. This cellular generation was promising a lot but in a reality it could not achieve higher transfer data speed more then 14.4kbps. It means that with the 2G for the same amount of bandwidth in a wireless spectrum it is not possible to transfer much of data, achieve many simultaneous phone calls or management operations like with the newer technologies. [9] In order to establish data connection it implied to burn all available monthly minutes and pay enormously high bills. [10] Without high data speed which is crucial for the IoT enviroment and non spectrum-efficiency that newer generations can offer, 2G could not even be taken in a consideration for IoT.

In between these transition standards like GRPS, UMTS and EDGE took the role and generally brought some improvements like sound quality, higher security and capacity. With some of them it was possible to establish and have data connection everywhere whenever our phone is appropriate configured. [10] They offered higher data speeds but IoT predicts more the 50 billion connected devices online. If the network can not provide the support for IoT/M2M evoulution it is better to get ready for the growing demands and higher standards and generations.

With 4G it is already possible to speak about „all IP“ solutions. 4G also provide widely available coverage, it is faster and generaly it is more then enough to support almost all IoT/M2M demands. It provides scalable and efficient networks, low latency, better services which are able to support IoT applications and devices. It is high reliable and implementation means simple and less expensive infrastructure reorganisation what is not the case with older cellular technologies. 4G promises creation of new business-to-business (B2B) market opportunities. [11] Also 4G

technologies/LTE network can provide low data costs. It means that the air interface can be splited into several narrow band channels of different bandwidth. This allows bandwidth to be assigned in a very flexible way that could make it ideal for M2M and IoT applications.[12]

Nevertheless to the benefits that 4G brings, it still has some limitations which are planned to overcome with 5G networks. Currently the biggest issue with 4G is that roaming between Telecom operator is so poor and modules can not be easily moved from one to another operators. LTE is still not fully deployed and need for dual-mode IoT/M2M modules can not operate simultaneously in all technology modes. This extremely increase the prices of devices/hardware.[9]

5G as a next step in IT world is promising fully mobility and connectivity. It will support more then 1 billion connections with extremely high data speeds with latency below then 1s. It is fully designed to support all variants called Machine type communication (MTC) for IoT. It will be fully built around previous generations technologies do there will be no any problems in cooperation between operators like it is now situation with 4G and hard transition from one to another. 5G will serve well in a scenarios where there are a lot of connected devices and different traffic types. It will easily broadcast HD video streaming using ultra-high links and still shows good results in data transfer with low-data-rate speeds for sensor networks. [13]

VII. CONCLUSION

With wireless connected network reality might look very different by 2020 and beyond. Automatically controlled devices and machines will be present everywhere, for the sake of education, healthcare, safety, vehicular applications and many more. IoPT is still in its initialization phase but it is right time to give a chance for interconnection with new-generation networks just because of optimization, efficient respond to customers, modernisation of post services and business processes, generating new revenues and better QoE. It is on us to govern this step in human evolution as an intelligent cooperation with the technology that enables versatile and adaptable usage.

VIII. REFERENCES

[1] Global mobile Suppliers Association, Over 50 LTE networks launched this year, total is 318 in 111 countries, [Online] Available: http://www.gsacom.com/news/gsa_410.php, [Last visit: 14.12.2015]

[2] IC Insights, Worldwide Cellphone Subscriptions Forecast to Exceed Worldwide Population in 2015, [Online] Available: http://www.icinsights.com/news/bulletins/Worldwide-Cellphone-Subscriptions-Forecast-To-Exceed-Worldwide-Population-In-2015, [Last visit: 15.12.2015]

[3] Desmond Khoo, Challenges in implementing LTE, [Online] Available: http://desmondkhoolte.blogspot.co.at/2012/04/challenges-in-implementing-lte.html; [Last visit: 15.12.2014]

[4] Himzo Bajric, lecture notes: "4G novo"; Fakultet za saobracaj i komunikacije, Sarajevo, 2013

[5] Jeanette Wannstrom, LTE-Advanced, [Online] Available:http://www.3gpp.org/technologies/keywords-acronyms/97-lte-advanced, [Last visit: 11.12.2015]

[6] Behrooz Rezvani, Pushing 802.11n Wi-Fi Technology So It Can Deliver Multiple Simultaneous Streams, [Online] Available: http://www.wirelessdesignmag.com/blogs/2012/02/pushing-80211n-wi-fi-technology-so-it-can-deliver-multiple-simultaneous-streams [Last visit: 10.12.2015]

[7] Charles Sturman, Ezra Stein, Carl Fenger, LTE - Delivering on the mobile broadband promise, [Online] Available: http://m2mworldnews.com/2013/06/01/67554-lte-delivering-on-the-mobile-broadband-promise [Last visit: 12.12.2015]

[8] The Internet of Postal things could help Posts generate operational efficiencies, improve customer experience, and develop new services and business models, Postaltehnology

[9] Aeris, 2G, 3G, 4G ... OMG!What G Is Right for IoT / M2M?, [Online] Available: http://cdn2.hubspot.net/hub/246745/file-226631113-pdf/ [Last visit: 25.04.2016]

[10] Engadget, 2G, 3G, 4G, and everything in between: an Engadget wireless primer, [Online] Available: http://www.engadget.com/2011/01/17/2g-3g-4g-and-everything-in-between-an-engadget-wireless-prim/ [Last visit: 25.04.2016]

[11] Andrew Lund, The IoT Showdown and the Opportunity for 4G LTE, [Online] Available: http://www.telecomengine.com/article/iot-showdown-and-opportunity-4g-lte [Last visit: 25.04.2016]

[12] M2M Connectivity, 4G/LTE for the Internet of Things (IoT), [Online] Available: https://blog.m2mconnectivity.com.au/tag/2g/ [Last visit: 25.04.2016]

[13] Majeed Ahmad, 4G and 5G wireless: How they are alike and how they differ, [Online] Available: http://www.androidauthority.com/4g-and-5g-wireless-how-they-are-alike-and-how-they-differ-615709/ [Last visit: 25.04.2016]

Advanced Sensing and Internet of Things in Smart Cities

Dijana Capeska Bogatinoska *, Reza Malekian **#, Jasna Trengoska *, William Asiama Nyako*
* Faculty of Computer Science and Engineering, University of Information Science and Technology "St. Paul the Apostle" Ohrid Republic of Macedonia
** Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria, South Africa, 0002
dijana.c.bogatinoska@uist.edu.mk, reza.malekian@ieee.org, jasna.trengoska@cse.uist.edu.mk, william.a.nyako@uist.edu.mk
Corresponding author: reza.malekian@ieee.org

Abstract - Internet of Things is largely perceived as a collection of interconnected "things" within smart cities. However, it can also be observed as a collection of data, shared between different devices and advanced sensors, then analyzed and used in different aspects of the smart city. In this paper, we introduce a similarity between a layered network architecture and "layered smart cities architecture" used for collection and sharing of big data with the aid of a cloud infrastructure in the smart city. The Internet of Things is discussed in the viewpoint of specific applications and technologies, i.e. layers in smart cities, such as: smart homes and automation, energy efficiency, e-health, intelligent transportation, waste and water management, smart agriculture, security and emergency services, application services and connected smartphones, etc.

Keywords: Internet of Things, data sharing, smart cities, smart homes, smart energy, smart water management, intelligent transportations, smart agriculture, E-health

I. INTRODUCTION

A recent British government report defines the Internet of Things (IoT) as a complex "ecosystem" used to connect anything, anyone, any service, any business, and any context using any network, anytime and anywhere. The Internet of Things describes a world in which everyday objects are connected to a network so that data can be shared.

It has been estimated by Cisco that there are currently more than 14,8 billion objects connected to the Internet, with the likelihood to reach almost 15 billion by 2020 [1]. Such a massive amount of information from sensors and devices could be used to enhance existing data and knowledge, and resulting insights that drive business and operational transformation. So, the primary value that Internet of Things creates is a direct result of the data that can be captured from connected things [1].

The large-scale and complex data sets generated by IoT devices encounter a number of challenges including: effective design and development of the methods and architectures to gather data from a wide variety of smart devices, to process, manage, control and monitor the data, to turn the device data into floods of valuable information and to effectively identify actionable insights.

Cloud computing systems play a big role in the IoT. Internet cloud services give the users the ability to access the information remotely. The interaction between the cloud and sensors, which involves many levels of communication, enables gathering of vast amount of data to the cloud and analyzing them using cloud-based applications.

In this paper we introduce a similarity between a layered network architecture and a layered smart city architecture for collection and sharing of big data. The key point of this concept utilizes a cloud infrastructure that is used for sharing, transferring and accessing data within the layers of the smart city. With the aid of this cloud infrastructure for data sharing, interconnected smart city layers can exchange data for the purpose of improving the overall efficiency of the smart city. Such smart city layers include intelligent transportation, energy efficiency, e-health, waste and water management, smart farming, interconnected devices, and application services. The highest and most refined layer of this smart city concept is the smart home and the automation services in smart homes. All the other layers, besides their own optimization based on data sharing, are also used to provide relevant responses that can be implemented in the smart home. All of this is made possible with the use of the cloud. Of course, such concept raises certain security considerations that are discussed in this paper.

The remainder of the paper is organized as follows. Section II gives an overview of related work. In Section III, we discuss IoT in the viewpoint of specific applications. In Section IV we describe the proposed similarity between a network layered architecture and a smart city architecture for sensor data collection and sharing. Finally, Section V concludes the paper and proposes future work.

II. Related Work

The similarity between a layered network architecture and the layered architecture of smart cities is based on a few different models of network architectures that are used as analogy to the layers in smart cities. To begin with, we take a look at the traditional Open System Interconnection (OSI) model, which serves as the universal standard for creating networks [2], [3]. This concept represents that basis upon which we build the layered smart city architecture.

In the direction of maintaining an "active storage" for a network architecture, in [4] the concept of active networks has been introduced. Active networks offer the possibility to dynamically inject executable programs, i.e. pieces of code to the network nodes, which increases the overall network capability of performing computations to such called "capsules" that traverse the network. These capsules carry fragments of programs that the nodes execute. The node organization in active networks can be seen in Figure1.

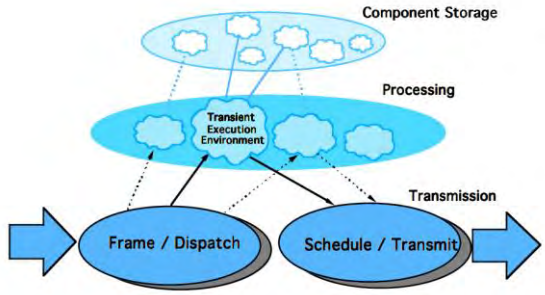


Figure 1: Active Node Organization in Active Networks, source: [4]

With the increased demand for security over the internet, certain improvements have been made to the traditional approach to a network architecture with the introduction of the concepts of trustworthiness and controllability. This architecture also increases the awareness for the quality of service of the network. This is a very important consideration for the design of the layered architecture of the smart city.

III. Internet of Things in the Viewpoint of Specific Applications and Technologies in Smart Cities

A. *Smart homes; Smart automation*

A smart home is a dwelling or a living environment in which all devices and appliances are connected through some central control unit, as a core of the home automation system, usually using incongruent network of sensors. The concept involves monitoring and control of the home devices and systems, providing a range of services including home security services, home healthcare services, comfortable living support services, energy management services, etc. Monitoring is made possible using smart home sensors and actuators for different purposes, see Figure 2. These sensors and actuators use various communication protocols to exchange information between devices. The full control of the devices is provided through the user interface, available from anywhere at any time, using PCs or smartphones through cloud based services.

The Smart home sensors generate a vast amount of data, which are delivered to data analytics and decision support tools to allow intelligent decision making and forward the obtained information as feedback to control smart devices and to carry out specific actions. This big data stream provides insight into the functioning of the devices, making better use of resources, and it enables better diagnostics and possibility to find and fix problems remotely and immediately. By conducting data analysis some patterns can be found, which leads to smarter life.

One main advantage of smart homes, especialy smart home automation systems, is the opportunity for improved life for the old or disabled people, along with longer independent life [5]. Authors in [5] have developed an interface for home appliances with the aim to support elderly and/or disabled in carrying out different tasks. The main characteristics of the proposed interface are *security* – security issues related with the residents' health, *intelligence* – to act depending on the situation, *cooperation* – to collaborate and exchange information with other home devices and *scalability* – for further extensibility.



Figure 2: Smart home and automation, source: http://specializedav.com/tag/smart-home-automation/

Smart homes transform the way we live, bringing a lot of benefits such as money savings, improved energy efficiency and environmental impact, improved security and comfort of homes, and ensure better quality of life.

B. *Energy efficiency*

Smart home automation and smart technology make our homes more energy efficient. Buildings are the major users of the electrical energy. Reduction of the energy consumption in the buildings is possible through

intelligent management and by making the home devices energy aware, smarter, intelligent and communicative. The smart devices are energy self-aware and share the energy consumption information with other smart devices.

The smart home communicates with smart grid and enables residents to manage electricity usage. By measuring home's electricity consumption more frequently to a smart meter, the home can provide the resident with much better information to manage their electricity bills. Smart meters provide the smart grid interface between the smart home and the energy provider, allowing automated transmission of information. This energy consumption data can be run through a home energy management system, which allows to track the energy consumption in details.

C. E-health

Health care is the most concerning part of human life. Current care models focus primarily on acute care. Significant advancements in technology are making it possible to deliver health care services in new ways. Using the technology, people are now able to take care of themselves at home. A digital medical home is the future of health care [6].

IoT-driven systems could help shift health care from medical treatment and acute care to prevention. Using in-home smart healthcare monitoring systems that rely on a number of different sensing systems, it is possible to detect subtle changes in health and to provide medical assistance. This is especially important for seniors and health care consumers with chronic, but manageable conditions.

Connected sensors, either mounted on the body or scattered around the home or living environment, collect a stream of health data in near real-time, which bring valuable information about a person's health. These devices can exchange data between them and with the emergency centers, hospitals and other medical institutions through Internet.

One possible scenario is to use passive infrared motion sensors to track every move of the patient, and to compute a motion density, as a measure of overall activity level. Other parameters, like walking speed, stride times, stride length can be measured, as well. The subtle changes in the sensor data patterns may correlate with changes in health.
The bed sensors are quite key to being able to detect very early signs of health changes. They are used to capture pulse and respiration as well as restlessness in bed.

IoT-based health paradigm can be focused on a variety of specific problems, like providing the daily dynamic health information for the patients, personalization of preferences, personalized diet prescription, auto reminder to take some medicines, auto detector of accidents, auto detector in case of emergency, or promoting understanding of how a therapy is helping to improve a patient's parameters.

D. Intelligent transportation

The increasing demand for mobility in smart cities implies looking at intelligent transportation systems [7, 8] that would not only be cost-effective, but would also be environment-friendly, safe for passengers and pedestrians, fast and fuel-efficient. After the expansion of mobility brought by the industrial revolution, people are put in the context of transportation, technology and economic development [9]. Many solutions have been brought to the traditional transportation systems throughout the years, which sometimes go on the account of the environment, the quality of air, or the economic costs in exchange for improved speed or efficiency.

It is very important to stress the need of road measurements and data in order to obtain accurate information about the city traffic. For this, various types of sensors can be used, such as static road sensors, mobile sensors placed on vehicles, or a hybrid of the two. Optimization of transportation and traffic in the city is rarely something the can be generalized – in most cases the methods used would depend on the traffic and road situation in each city. One good example of this are cities in India, where the methods used for traffic optimization can hardly be applicable for European cities [10].

E. Waste and water management

Waste management becomes an increasing problem in urban living. It has many aspects, among which are socio-economic and environmental aspects. Another important aspect lies in the consumption habits of habitants of modern cities. One important factor in waste management is sustainability, in the first place environmental sustainability. A more traditional approach to environmentally sustainable waste management in cities is the geospatial approach [11].

In smart cities, waste management is often viewed as an IoT service and is sometimes closely connected with intelligent transportation. Actually, the combination between Intelligent Transportation Systems (ITS) and Waste Collection (WC) provides a novel and advanced platform in smart cities. This system makes use of surveillance systems, sensors and cameras in order to handle difficult accessibility in a more efficient way, which is an important aspect in a smart city. In the core of this system lies a cloud-based Decision Support System (DSS), which combined with dynamic routing models and surveillance cameras is able to offer services to different types of stakeholders in the smart city. This has a significant positive impact on cost-effectiveness in the smart city.

F. Smart Agriculture

One approach towards the improvement of farming and agriculture in smart cities is the usage of applications with smartphone-based sensors to improve the productivity of farmers and to manage the tasks in each farming cycle [12]. Smart sensor technology can be used for monitoring environmental conditions in agriculture fields and shaping predictions for farmers to be able to take measures accordingly [13]. Monitoring the agricultural environment is important for factors such as humidity, temperature, moisture, etc.

Besides monitoring the outside environmental factors, in smart agriculture it is very important to be able to minimize the use of chemicals, pesticides and fertilizers by optimally using the natural resources from soil. In that direction, bio-sensors can be used to provide data containing measurements about the nutrient content in soil. Together with the agriculture environmental sensors, this network of sensors manipulates data and makes important conclusions that are then sent to the farmers via SMS [14]. The result of such system of networked sensors aided by data logging and notification center is primarily in obtaining a better quality and increased amount of food grain, while minimizing the use of chemicals, as well as with optimal consumption of water and electricity.

Similarly, a multidisciplinary model for smart agriculture using IoT, smart sensors, cloud and mobile computing and big data analysis is proposed in [15]. As we can conclude, the socio-economic factor in agriculture is also included in this smart system through the user interface platforms that connect the farmers to relevant information about market demands and production needs on one side, but also to loans and supporting programs on the other side.

G. Security

One factor that comes up when multiple devices start to communicate together is security. Smart devices and smart homes come together to form a smart city, but there is a large question of how secure all of this is. The recent increased use of smart devices has also been accompanied by more reports of devices leaking user's data. Smart TVs, fridges and even kettles have all been reported as leaking data during the years of 2013 to 2015.

For smart cities to function effectively, protocols have to be established for how data can be exchanged and transferred between devices. Various different protocols already exist, for example Thread. Thread is a protocol that is being worked on by Google, ARM and Samsung – all major players in the future of smart devices. Thread is designed to allow devices to talk with each other in a secure manner and with a high efficiency. Eventually, a protocol has to be developed that will allow all manner of devices to talk and communicate with each other. Without the secure transfer of this data in the city, users' data may be leaked and abused. A quick mention can be made of Ukraine's power grid hack that occurred in mid-January 2016. This is the first occurrence of a national power grid being hacked, but with the spate of smart devices, this is something that will continue to occur in the lack of secure data exchange protocols.

IV. Proposed Layered Smart City Architecture using Sensor Data Collection and Sharing

In this section we propose a similar model to the well-known layered network architecture (overviewed in Section II) in the context of layered smart city architecture. The layered smart city architecture is based on sharing big data between various advanced sensors implemented in different layers of the smart city. The sharing is done by utilizing a cloud infrastructure that is accessible by the different smart city layers. This approach emphasizes the importance of sharing of data in the smart city, as one of the core factors that make the city smart.

The layered architecture of the smart city is a bottom-up architecture where the Smart Homes and Automation Layer is considered as the top level of the smart city. The bottom line is that in different ways, all the other lower-level layers work between themselves in order to improve different aspects of smart home efficiency. Improving people's habitats' efficiency and sustainability results in overall improvement of those parameters in the smart city, because all layers work towards that goal.

An important consideration for this concept is the security of data that is shared and stored on the cloud. This cloud infrastructure is the core of this concept since the entire data is stored and processed there. Therefore, it is crucial for the smart city that all of its data is well secured and protected. We propose that all data transactions go through a secure connection with increased level of protection.

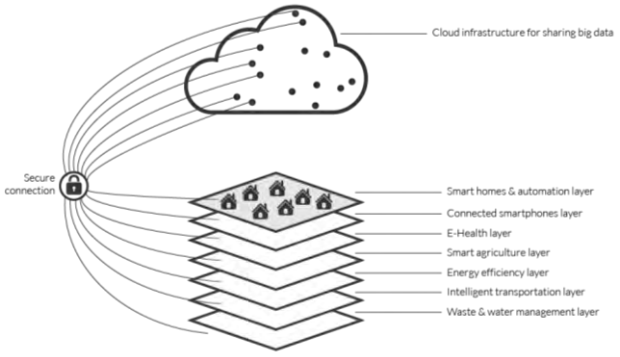


Figure 3: Proposed Smart City Layered Architecture

Below we discuss the interconnection and relationship of all smart city layers in this architecture, starting from the bottom layer.

The Waste & Water Management Layer provides a potential solution to the household waste management issues. As an alternative to the geospatial approach for environmental sustainability discussed in Section II, we propose an intelligent technology available to households in smart cities that will serve as a convertor of waste into green energy. This produced energy could then be used as a small alternative to using grid-based electricity for some less-consuming household devices. The convenience of this solution does not only lie in the production of green energy from household waste as a step towards energy-sustainable households, but also solves the problem of waste collection, transportation and management in the environment. In the case of water management in smart cities, known in some cases as smart water, we propose the concept of big data in the cloud in combination with some sensing technology. The concept consists in obtaining measurements about the water flow, pressure, current usage, etc., and then it makes smart predictions that would help to distribute exactly the necessary amount of water to various regions in the smart city, in different times of the day.

The Intelligent Transportation Layer in this case is focused on public transportation. Here we propose intelligent public vehicles for smart cities that would be based on the concept of big data shared via cloud, in order to improve the efficiency of public transport. We propose intelligent public vehicles that would use advanced sensor technology to prevent traffic accidents by making environment-based predictions (based on the proximity and speed of other vehicles and pedestrians, obtained with the use of sensors). Furthermore, these public transport vehicles would have access to a cloud of shared data and traffic management systems, assuring a shortest path to destination based on traffic predictions and determining the least busy routes.

The Energy Efficiency Layer has an important application in smart homes. Finding out about the electricity usage habits, also known as Habitual Average of electricity consumption in smart homes, is an important step towards the optimization of energy usage in the household [16]. By analyzing the Habitual Average of the usage of different devices in the home, we not only take in consideration the standard average based on data measurements, but also the habits of switching different devices on/off by different members of the household. In this way, a Smart ECO (Electricity Consumption Optimization) model can be developed and applied in smart homes, so that electricity usage is reduced. Combined with the conversion of waste into a small amount of green energy, it ensures a more efficient usage of electricity in the home.

In the Smart Agriculture Layer, sensing technology can be used to measure soil temperature and conditions, water needs of plants, soil composition, etc. A smart watering system along with a signalizing sensor that indicates when a plant needs more water, would optimize the water consumption for agriculture purposes on one side, and on the other side it would improve the growth and production of agro-products. As it can be noticed, this layer communicates with the Waste & Water Management Layer on the level of smart water usage for watering plants only when they need more water.

The E-Health Layer proves another advantage of exchanging data between different layers, for example, data exchange between the Intelligent Transportation Layer and the E-Health Layer, could result in increased efficiency of the smart city as an outcome from having access to data. Therefore, e.g. hospitals can make predictions about increase in visits by patients to the hospital, based on monitoring data obtained from transportation patterns through specific roads in the city, obtained from the Intelligent Transportation Layer. In this way, hospitals can make decisions about the time of the day when they should increase the number of workers in order to be able to serve patients more efficiently.

The Connected Smartphones Layer is very important since in smart cities, more and more control is given to smart devices that are able to control other devices in the household, office, etc., aided by smart application services and programs. The concept of connected smartphones that send data to the cloud is important in the smart city in order to process and conclude important things about usage habits, devices that are controlled, social activity of the users, etc. These are all very important data in a connected smart city, and can be largely applied and used within almost any of the other smart city layers.

At the highest level in the layered smart city architecture, stands the Smart Homes & Automation Layer. All the functions of the previously discussed layers, besides their interchangeability towards improved efficiency, work together towards the optimization of functions and services within the smart home. Smart people in a smart city come out of the smart homes concept – and this means improved quality of life, life efficiency and effectiveness, etc. This includes all the technological and socio-economic factors that are encountered in the everyday life of people. Smart people are those who eat healthily produced food in a sustainable green environment, use better health services, spend less time on transportation because they utilize smart transport, and pay lower bills because of optimized water and electricity consumption, where even their household waste is used to support this concept.

V. Conclusion and Future Work

It is important to stress that this proposed architecture of layered smart cities is possible thanks to the concept of sharing big data between different smart city layers. In this way, some patterns can be extracted from different data points that would otherwise not be possible to obtain, if data from every layer was to be isolated from the other layers. This interchangeability of shared data across all smart layers increases the overall productivity and efficiency in the smart city, which is not possible to this extent when layers do not share data with other layers. With this approach, we can look at the smart city as an enormous collection of big data, which processed and combined in different ways, ensures smart life of smart people.

ACKNOWLEDGMENT

This work is supported by Tertiary Education Support Programme (TESP), ESKOM, South Africa for a project entitled "Energy Consumption Optimization, an Internet of Things Approach".

References

[1] "Internet of Things Connections Counter", CiscoSystems, 2015.

[2] V. Chinmay, G. Rishabh, "A review paper on OSI Model - a seven layered architecture of OSI Model", International Journal of Innovative Research in Technology IJIRT, vol. 1, issue 12, 2015.

[3] S. Kumar, S. Dalal, V. Dixit, "The OSI Model: overview of the seven layers of computer networks", International Journal of Computer Science and Information Technology Research, Vol. 2, Issue 3, pp.461-466, July - September 2014.

[4] D. L. Tennenhouse and D. J. Wetherall, "Towards an active network architecture", http://www.tns.lcs.mit.edu/. An earlier version of this paper was delivered during the keynote session of Multimedia Computing and Networking, San Jose, CA, January 1996.

[5] R. Delgado, A., Robinet, A. McGinn, J. Grout, V. Picking, "Assistive human - machine interfaces for smart homes", in: 3rd Collaborative Research Symposium on Security, E - Learning, Internet and Networking (SEIN), pp. 147--157. Information Security & Network Research Group, Plymouth 2007.

[6] R. R. Stough and G. Yang, "Intelligent transportation systems", Transportation Engineering and Planning Vol. II, Encyclopedia of Life Support Systems (EOLSS).

[7] J. Wannenburg, R. Malekian, Body Sensor Network for Mobile Health Monitoring, a Diagnosis and Anticipating System, Vol.15, No.12, pp. 6839-6852, 2015.

[8] Ye Ning, Wang Zhong-qin, R. Malekian, Wang Ru-chuan, A. H. Abdullah, "Design of accurate vehicle location system using RFID", Elektronika ir Elektrotechnika, Vol.19, No.8, pp.105-110, 2013.

[9] Ning Ye, Yan Zhu, Ru-chuan Wang, Reza Malekian, Qiao-min Lin, "An efficient authentication and access control scheme for perception layer of Internet of Things", Applied Mathematics and Information Sciences, Vol. 8. No.4, pp.1617-1624, 2014.

[10] R. Sen and B. Raman, "Intelligent transport systems for Indian cities", https://www.usenix.org/system/files/conference/nsdr12/nsdr12-final2.pdf.

[11] S. A. Shah andM. A. Wani, "Geospatial based approach for enhancing environment sustainability of Srinagar city - a study on solid waste disposal", International Journal of u- and e- Service, Science and Technology, Vol.7, No.3, pp.289-302, 2-14.

[12] S. Pongnumkul, P. Chaovalit, N. Surasvadi, "Applications of smartphone-based sensors in agriculture: a systematic review of research", Hindawi Publishing Corporation Journal of Sensors, Vol. 2015, Article ID: 195308, http://dx.doi.org/10.1155/2015/195308.

[13] S. S. Mathurkar and D. S. Chaudhari, "A review on smart sensors based monitoring system for agriculture", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol.2, Issue-4, March, 2013.

[14] Dr. K. S. Ravi, K.Tapaswi, B.Lokesh, G.S. Krishna, K. S. Ravi et al., "Smart sensor system for agricultural chronology", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol.2(6), pp.2650-2658, 2011.

[15] H. Channe, S. Kothari, D. Kadam, H. Channe et al., "Multidisciplinary model for smart agriculture using Internet-of-Things (IoT), sensors, cloud-computing, mobile-computing & big-data analysis", Int.J.Computer Technology & Applications, Vol.6 (3), pp.374-382.

[16] R. Malekian, D. Capeska Bogatinoska, A. Karadimce, J. Trengoska, W. Asiama Nyako, "A novel Smart ECO model for energy consumption optimization", Elektronika ir Elektrotechnika, Vol. 21, No. 6, 2015.

Security challenges of the Internet of Things

Mario Weber, MsC
Croatian Regulatory Authority for Network Industries
Mihanoviceva 9, HR-10110 Zagreb, Croatia; e-mail: mario.weber@hakom.hr
Marija Boban, PhD
University of Split Faculty of Law
Domovinskog rata 8, 21000 Splilt; e-mail: marija.boban@pravst.hr

Abstract — **Internet of things (furthermore IoT) by definition refers to a networked interconnection of devices in everyday use that are often equipped with ubiquitous intelligence. It is based on processing of large quantities of data in order to provide useful information/service and enables an efficient regulatory policy in the area of IoT. One of the key challenges for the realization of the IoT includes security challenges, especially in the area of privacy and confidentiality among management of heterogeneities and limitations of network capacities. Those challenges will be based on information security management systems as well as on legal foundations. When considering the legal framework of security and privacy of the IoT, it has to be determined which model of regulation should be applied. Thereby, no traditional government regulation is actually appropriate for a global system such as IoT. The most important foundation consequently will be the regulatory foundations of European union (on the territory of EU) and, consequently building the model of self-regulation based partly on state and mostly on law international agreements which are to be considered as tools to govern the IoT. This paper gives an insight into the most important security challenges related to implementation of IoT and building appropriate regulatory framework. Adoption of the regulatory framework is crucial for the development of IoT services due to adequate legal security companies will get.**
Key words: security, IoT, challenges, privacy, data, confidentiality, regulation

INTRODUCTION

The Internet, or how we like to refer to it, the World Wide Web, is no longer merely a network of computers but it also connects different devices that are able to send and receive information about the status of devices connected to the network. *The Internet of Things* (hereinafter: IoT) is a concept allowing for the networking of different things and objects from everyday life and their everyday communication over the Internet, without human interaction, with a view to improving the conditions and way of life. The term Internet of Things was employed for the first time by Kevin Ashton, the director of MIT's Auto-ID centre, in 1999 in his presentation for Procter & Gamble, but it took a few years for the term to enter into more frequent use. IoT was formally introduced as a term in 2006 by the International Telecommunication Union (ITU) in its Internet report [1]. With the development of IoT technology, every device in our surroundings will be able to communicate with another device and send information to that device or control it, depending on the collected information.

The Internet of Things is a network connecting objects from different environments into a single large network based on the Internet Protocol, and it is a basis for the development of the so-called smart environments, such as smart houses, factories or even cities. As opposed to the former vertical development of individual areas and technologies (from technologies for connecting mobile devices, wireless sensory networks, processing of large quantities of data), the addition of a new middleware permits the realization of the concept of the Internet of Things. The middleware enables the connection of different technologies and platforms and it is a platform for the development of advanced services without the need to possess knowledge about every technology separately with a view to developing new services [2]. The networking of a large number of objects in our surroundings will result in a larger quantity of generated data that must be stored, processed and presented in an acceptable form. There are many applications for services in the IoT field, from user applications that will provide added value only to the user who defined them to entire branches of science that are provided simplified sharing of resources, collecting of data from a large number of sensors, statistical analysis of such data and conclusions in relation to data in real time [3].

The potential for economic exploitation and application of IoT was recognized by the Economist magazine as a result of a survey conducted in June 2013 when opinions of 779 business experts were collected to define the business index of IoT. Results have shown that three quarters of companies have been actively researching or using IoT, and 96% stated that they would start using IoT in one form or another in the following 3 years. The assessment of potential growth of the IoT market in the next few years varies depending on number of connected devices and on the economic growth of the market. Therefore, for example, Gartner estimated that 30 billion devices will be connected to the Internet in 2020 with a single IP address which would bring additional USD 1.9 billion (10^{12}) for the world economy [4]. At the same time, Cisco and Ericsson forecast that this number would increase to 50 billion [5] [6] objects connected to the Internet, which will result in earnings for the world economy amounting to USD 14 billion. For comparison sake, this amount exceeds total GDP of 17 EU Member States in 2011 [7]. IDC also envisaged in October 2013 that there will be a total of 212 billion "objects" by 2020, which would equal to EUR 8.9 billion per year with the annual growth rate of 7.9% [8].

This paper provides an overview of the situation in IoT with a special emphasis on regulatory challenges that will emerge in the upcoming years in this area. It gives history of the development of IoT technology through different European and international bodies and organizations, such as ITU and the European Research Cluster with projects

related to IoT. It also provides a definition of IoT and related concepts, an overview of two reference models for IoT architecture and the most important protocols for the Internet of Things. In addition to that, an overview is provided of the current situation in IoT in Europe including situation in Croatia and in the USA, including challenges and possible problems for the realization of IoT.

I. DEFINITION OF THE INTERNET OF THINGS

For the time being there is no generally accepted definition of the Internet of Things. The definition itself is still unclear and it is frequently a subject of philosophical discussions. Majority of authors in this area use their own definition which, to a greater or lesser extent, differs from other definitions. Some of the definitions that most frequently occur in literature are provided below.

As a rule, IoT may be defined as a network of objects that are able to detect and exchange information among each other, but it differs from the Internet in several other aspects [1].

A more widely accepted definition is ITU's definition from 2005, which is very general and reads as follows: IoT is a global infrastructure for the Information Society, enabling advanced services by interconnecting (physical and virtual) things based on, existing and evolving, interoperable information and communication technologies [1].

The RFID group defines IoT as the worldwide network of interconnected objects uniquely addressable based on standard communication protocols [9].

According to the *IoT European Research Cluster* (IERC)[1] IoT is an integral part of the future Internet, and it is defined as a dynamic global infrastructure that has the ability of self-configuration, that is, a network based on standard and interoperable communications protocols, and a network in which physical and virtual "things" have identity, physical characteristics and virtual personality and also use intelligent interfaces while being at the same time secretly integrated into the information network [10].

IERC also provides its own definition of "things" as active participants in business, information and social processes where they are enabled to interact and communicate among themselves and with the environment. This interaction and communication take place in the form of the exchange of data and information collected from the environment. While reacting autonomously to the real/physical world events and influencing it by running processes, "things" trigger actions and create services with or without direct human intervention [10].

"Things" may also be defined as real/physical or digital/virtual subjects that exist and move in time and space and may be identified. Things are often identified in accordance with the assigned identification number, name and/or location address.

IoT vs M2M

IoT and M2M systems largely overlap but, considering their definitions, they may not be regarded as identical. A shared characteristic of IoT and M2M is remote access to devices. However, there are some essential differences between them. A traditional M2M refers to communication between two or more devices by means of a mobile or fixed network. This is vertical point-to-point communication. M2M applications usually consist of a hardware module embedded into a device on the user side which communicates through a mobile or fixed network with the corresponding application, often on the side of the service provider. The objective is to reduce management and maintenance costs.

M2M connects "things" with a computer. IoT connects a computer with "things" (machines, devices, sensors, products etc.), systems (business applications, support systems, analytical systems, data storages, control systems etc.) and people (consumers, employees, partners and buyers). IoT is based on the IP network for horizontal connection of objects/devices to a cloud or user platform. IoT usually combines sensory data with "big data", analytic or company applications for the improvement of productivity, increase of production and share on the market, improvement of services, offer of SaaS (Software as a Service) etc.

As opposed to M2M which only supports devices, IoT solutions support passive sensors as well, small strength sensors and cheap devices that may not be adjusted to M2M hardware modules. IoT devices communicate through standard IP networks. IoT is usually connected to a cloud which makes it a scalable and flexible solution as opposed to M2M communications which are frequently oriented towards installation of SIM cards or drawing a fixed line. IoT is more oriented towards software solution or IP network while M2M communication is primarily oriented towards an embedded hardware and mobile network. It must be mentioned that M2M with Internet Protocol is a part of IoT. IoT is a broader concept than M2M because it may be integrated into the overall company business solution while M2M is more oriented towards maintenance.

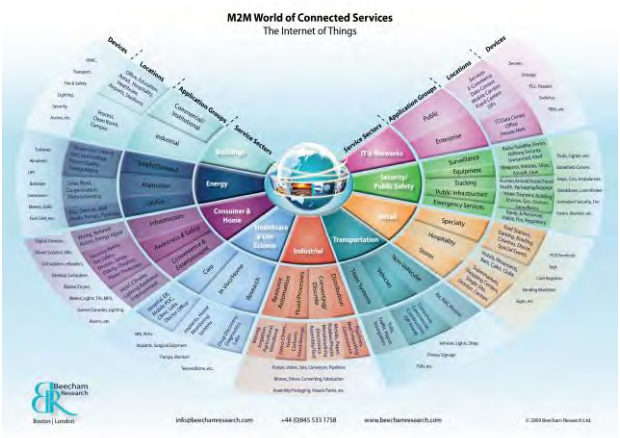


Figure 1. M2M/IoT Sector Map (source: www.beechamresearch.com)

SCOPE OF IoT

Although IoT enables a wide spectrum of applications in everyday life, its current application is limited and significant changes are expected in the future. Key enabling technologies include ubiquitous connectivity, smart devices and possibility for integration of smart objects into different applications. There are many different connected things ranging from machines to vehicles, devices, light fixtures and buildings. They are not limited only to material objects, for example, smart places and sensing of the environment are very important

[1] http://www.internet-of-things-research.eu/

for many applications. The scope of application is very wide, including upgraded traffic safety and traffic management, transformation of the power grid towards advanced grids on the basis of new requirements such as energy efficiency, micro production of electricity, electrical vehicles and awareness of users about energy consumption.

The scope of application of IoT may be divided into the following areas: personal and household use, economy, service programmes and mobile area. The above-mentioned division is presented in Figure 2.



Figure 2. A graphic presentation of IoT end users and areas of application depending on the type of collected data [9]

Personal and household IoT

Information collected by sensors is used only by the person who owns the network. Wi-Fi is often used as the core network enabling high permeability and supporting the transmission of video signal and high frequency of sampling which favours the transmission of audio signal.

One of the best examples of application of IoT in this area is e-health care. Data collected by sensors placed on body are sent to one of the smart devices in the house which further transmits them to a server. A smart phone is a good example of a data-collecting device because it contains several interfaces the most suitable of which for this purpose is Bluetooth. The collected data may be used by doctors to supervise patients even when they are at home, which contributes to the reduction of hospitalization costs.

Control of household devices, such as air-conditioning devices, refrigerators, washing machines etc. will enable better management of the house (smart house) and more efficient energy use [9].

Economic IoT

The economic application in the work environment refers to the "network of things". Data collected by such network are used only by the network user and they may be selectively published.

Sensors have always been a constituent part of factories in relation to safety, automation and other processes necessary for continuous operation. Sensors will be replaced in the future by wireless systems which will ensure flexible changes of settings, that is, simpler changes of settings whenever necessary. The new system is merely an IoT sub network which is restricted to one factory.

Service programmes

Data collected from the network are most frequently used for the optimization of processes. Examples of such use include smart meters. Data collected by means of smart meters enable providers of utility services to manage their resources with a view to achieve greater optimization of costs and profits. Such systems consist of very expensive networks used for the supervision of key infrastructure and efficient resource management. Mobile network, Wi-Fi network or satellite communication may be used as a key network. Concrete application of achievements in this area will significantly improve the existing monitoring systems that will be more efficient in the monitoring of certain targets, noticing of suspicious activity and monitoring of unauthorized access.
Ensuring the quality of water or monitoring of agricultural land are only some of the areas in which IoT has already had many benefits, and it is certain that benefits from use of IoT technology will be even greater in the future [9].

Mobile IoT

Noise pollution is mostly caused by city traffic. Furthermore, traffic also contributes to the reduction of air quality and to emission of greenhouse gases. Traffic jams contribute to increased costs of economic and social activities in majority of cities.

All of the above-mentioned problems may be reduced by continuous collecting and processing of traffic data. In relation to traffic via long distance wireless sensors networks, IoT enables constant monitoring of time of travel, from points of origin to destination points, of air pollution and noise pollution. This kind of an IoT system will most likely replace current systems for collecting traffic data and support the development of algorithms for traffic management, including higher object control systems. The results of processing of collected data will be presented to passengers who will have a continuous insight into situation in traffic [9].

II. BUILDING THE SECURITY AND REGULATORY FRAMEWORK OF THE INTERNET OF THINGS

IoT is a complex system that might bring about numerous innovations and new rules of behaviour. Scientists, politicians, CEOs of large companies in Europe and the USA are currently discussing potential rules of behaviour. Regulation must take into account the complexity of the IoT system which is why the problem should be considered from several forward-looking levels. Special attention should be paid to results of market regulation, that is, to its impact on innovations and adjustment and speed of spreading of the IoT in society and business environments. Below is an overview of discussions in the private sector and in the public about key IoT problems that include privacy, security, ethics, data confidentiality, competition, economic development and freedom of innovation.

Europe

Due to its demanding objectives, Europe has achieved the greatest progress in the IoT segment and its development

is further supported by regulations in several sectors. Guidelines important for the for the IoT sector are defined in the following documents:

1. "The Energy Services Directive" – Directive 2006/32/ECon energy end-use efficiency of various devices and energy services In accordance with this Directive, it is expected that 80% of households will have installed smart energy meters by 2020.
2. "eCall Directive"– by 2015 the eCall service should be available in all EU Member States. Until then all new vehicles need to have installed devices that will automatically dial 112 in case of an accident and provide key information, such as GPS coordinates, data about the driver and the similar.
3. "The EU 531/2012 Roaming Regulation" – the Roaming Directive introduces important changes that may assist EU market, such as the introduction of an aggregator as a wholesale service provider.

The discussion on the importance of IoT in Europe started in 2006 and continued in 2007 when the concept of IoT was officially accepted in the Commission Communication on RFID [15].

In the report on "Future Networks and Internet" from November 2008, it was recognised that IoT has high potential for the development of new services but at the same time presents risk in terms of protection of privacy of individuals [16]. The need for discussion about architecture and management of the IOT has been recognized, and all the Member States have been invited to promote decentralised management in order to ensure confidentiality, safety, privacy and ethical management of data exchanged in the IoT [16].

The European Commission accepted the strategic action plan for IoT in 14 points [17] in 2009 which represented a collection of principles for management of the IoT, the assessment of possible risks, the financing of joint investments, participation in international dialogue, and integration of IoT into 4 research and development projects of a public-private partnership. In June of 2010 the EC established a multi-stakeholder working group consisting of people of different profiles including scientists, technicians, representatives of the industry, legal and social sciences as an advisory body of the European Commission dealing with main issues including management mechanisms, ownership of data, privacy, security, standards and international cooperation.

The EC conducted a public consultation in IoT management in 2012. The questionnaire collected opinions on the appropriate approach to management in order to stimulate faster development of IOT, while at the same time ensuring adequate protection of EU citizens. The results have demonstrated a joint position that IoT will contribute to significant economic and social progress, in particular in the areas of health, independent life and support for disabled persons. In all other areas, it illustrated differences in opinions between citizens and associations of citizens that desire a stricter regulation compared to the industry. According to the results, 77% of respondents supported the development of indicators for impact assessment in the field of data protection, including requirements related to the consent of users to collect data, data anonymization, restricted use and data retention and privacy. An interesting fact is that the majority of replies indicated that IoT will bring about many ethical issues related to the collected data but also that the development of IoT should not result in social injustice. The European Union expert group concluded in December 2012 that there is a significant disagreement between the business community (in the first place the industry) and the public on the majority of problems, in particular privacy, security and need for faster development of competitiveness in an IoT environment [18]. On the basis of this conclusion, the EC entrusted a European company named RAND with the development of guidelines for IoT management. Its report was published in August of 2013 stating that IoT is developing quickly and represents a challenge for traditional business, market, management and social models. Economic, social, political, legal and technological Internet management is based on a presumption of rational selection, market strength and efficient self-organization adjusted to systems under human control. Since interactive autonomous IT systems are significantly different from this paradigm [19], RAND recommends the application of a "soft law" approach that will include standards, supervision and ethical character and will at the same time ensure freedom for the industry to adjust to requirements in a simpler manner. This should ensure space for the development of IoT in the EU.

The United States of America

As opposed to Europe which stimulates the development of IoT with is regulatory activities, discussions on IoT in the United States are sporadic and scattered. The majority of debates take place within individual federal agencies interested only in some segments of IoT. Thus, for example, in 2012, the National Institute of Standards and Technology published the Framework and Roadmap for Smart Grid Interoperability Standards [20] which covered only one segment of IoT. The State Department investigated in 2013 the possibilities for using the IoT for monitoring weapons of mass destruction, and the Department of Homeland Security has been continuously installing sensory systems for detection of chemical and biological threats [21].

The first serious discussion was initiated in April 2013 by the Federal Trade Commission by asking for comments on the influence of IoT on privacy and security [22]. Only 27 replies were received, out of which more than 60% from the industry or their associations. Consequently, comments were against regulation. The majority of comments were focused on self-regulation and the removal of government influence on the development of standards of privacy and security that would endanger the development and innovations for IoT. The industry believes that the development of certificates for security and privacy could ensure sufficient protection of users.

In November of 2013 the FTC held a public workshop on IoT in order to investigate problems related to security and privacy due to increasing numbers of connected

devices [23]. At this workshop, the President of the FTC pointed out the key challenges, one of the main being the loss of collected and stored personal data. The three basic principles he singled out included data transparency, simple selection of control of personal data and the privacy model [29]. The round table further opened a discussion on additional control issues such as the use of IoT for government supervision of citizens. Several options for the protection of privacy and safety of users were proposed, such as self-regulation, the preparation of certified approval certificates, implementation of ethical codes, the implementation of the existing law on users and the development of new regulation of IoT. The conclusion of the work group was that regulation will depend on whether companies earn revenue exclusively from IoT products and services or they plan to profit from user data by offering data analysis services.

III. SECURITY CHALLENGES RELATED TO IMPLEMENTATION IOT

IoT is an area in which research is in full swing. Following the basic research on technologies used in the area of IoT, it is necessary to introduce standards for architecture, platforms and communication between individual components. The development of standards and platforms for IoT is a basis for the development of advanced services.

There are many possible directions for research because the mature phase of development of IT brings new challenges related to market regulation, payment, security and improvement of performances and efficiency of work of platforms. So far, research focused on the establishment of functional platforms that will enable the provision of an increasing number of services, but, with the increase of the number of users, it will necessary to provide support to the quality of service. The open issues in the area of data processing include standardization of query language, definition of key evaluation (and comparison) parameters for various components and adjustment of algorithms for processing data flow in cloud computing and variable characteristics of entry data.

Energy efficiency will represent a significant challenge in the development of communication protocols and devices. Power sources and autonomous operation will be of great importance, in particular for sensory modules that will for the most part be battery-powered. The best indicators are today's sensors which are mostly battery-powered and constantly broadcast data by using Wi-Fi or Bluetooth technology. Key challenges and possible problems [26] that need to be considered and solved before mass application of the IoT include the following:

1. Security privacy and confidentiality - the security domain has the following challenges: (a) securing IoT architecture, (b) proactive identification and protection of IoT from arbitrary attacks (e.g. DoS and DDoS attacks) and abuse and (c) proactive identification and protection of IoT from malware. In the area of user privacy, special challenges include the following: (a) control of personal data (privacy of data), (b) need to improve privacy technologies and the relevant laws [28].and (c) standards [27]., methods and tools for managing the identity of users and objects. In the area of confidentiality, some of the specific challenges include the following: (a) the need for simpler exchange of critical, protected and confidential information and (b) confidentiality must be a constituent part of IoT design.
2. Standardisation: Management of heterogeneity - management of heterogeneous applications, environments and devices represents a great challenge as well as the standardization of heterogeneous technologies, devices, applications and interfaces. The lack of a single connection standard (a wide selection of wired and wireless and "proprietary" solutions) will represent a significant challenge to connectivity at the global level.
3. In security challenges great role play limitations of network capacity – the convergence of devices that arises from the IoT stimulates greater demand for a certain degree of expected QoS of the associated network infrastructure. New mobile applications that provide certain services may demand more frequent sending of small blocks of data (sessions) necessary for upgrading and synchronizing. The frequency of the above-mentioned sessions will have a great impact on delay and permeability of the network itself. This part of infrastructure must be securely delivered in order to ensure secure data flow.
4. And last but not least: the management of a large quantities of data and processing of large quantities of data to ensure useful information/service as well as to ensure confidentiality and integrity of data as a part of information/service (IoT except for "raw" data from the source may generate a large quantity of metadata or temporary data necessary for the realization of services (e.g. sensory reading may be stored in a relational and semantics database). Characteristics are the following: quantity (size of data), diversity (heterogeneity of processed data, e.g. tables, pictures, e-mail) and dynamics (characteristics of data, intensity and frequency of arrivals, manner of processing). Platforms for IoT must enable efficient processing of data in real time resulting from user demands because the user wishes to be instantly informed of interesting events (e.g. selection of the most favourable travel route).

In addition to the above-mentioned main challenges, there are also some other challenges such as: (a) market regulation, (b) designing of a more efficient architecture for networking of sensors and storage of collected data, (c) development of mechanisms for the processing of the flow of data gathered in sensory networks, (d) transition to IPv6 (large number of addresses, possibility for auto configuration and improved security parameters), (e) power sources of devices/sensors (devices are powered by electricity produced from the environment such as vibrations, light and air flow) and (f) decreased cost of IoT components.

CONCLUSION

IoT represents a new, interesting direction in the development of the Internet. It refers to unique identification of objects and their virtual representation in the structure of the Internet. Such objects may communicate with each other, provide information about itself and accept data collected by other objects. The development of IoT depends on the dynamics of innovations in numerous technical fields, from wireless sensors to nanotechnology. Capacities, such as the monitoring of changes in the environment or communication between devices, represent high priority in the development of IoT.

One of the key challenges for the realization of the IoT include security, privacy and confidentiality, management of heterogeneities, limitations of network capacities, management and processing of large quantities of data in order to provide useful information /service and enable an efficient regulatory policy in the area of IoT. Since, protection of privacy is one of the key constitutional rights of citizens' it is very imortant to note that IoT will have to be in compliance with the new European regulatory frameworks for data and privacy protection, as well as with all legal requirements into a single group of rules in the EU and inclucing also revised measures for data transparency and safety issues.

REFERENCES:

I. „ITU Internet Reports 2005: The Internet of Things“, International Telecommunication Union

II. Antonić, „Platforme za obradu podataka u stvarnom vremenu u području Internet objekata“, Report for doctoral qualification exam, 2013.

III. John Soldatos, Nikos Kefalakis, Manfred Hauswirth, Martin Serrano, Jean-Paul Calbimonte, Mehdi Riahi, Karl Aberer, Prem Prakash Jayaraman4, Arkady Zaslavsky, Ivana Podnar Žarko, Lea Skorin-Kapov and Reinhard Herzog. OpenIoT: Open Source Internet-of-Things in the Cloud, Invited paper, to appear in Lecture Notes in Computer Science, vol. 9001, 2015.

IV. „Forecast: The Internet of Things, Worldwide, 2013.“, Gartner, 2013.

V. D.Evans, “ The Internet of Things How the Next Evolution of the Internet Is Changing Everything”, White Paper, Cisco, April 2011.

VI. Miguel Blockstrand, Tomas Holm, Lars-Örjan Kling, Robert Skog and Berndt Wallin, „Operator opportunities in the internet of things“, Ericsson review, 2011

VII. Rooney, B. “Internet of Things Poses Big Questions.” Wall Street Journal Online, July 3, 2013.

VIII. International Data Corporation (IDC) Press Release. “The Internet of Things Is Poised to Change Everything, Says IDC.” October, 2013,

IX. Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, Marimuthu Palaniswami, „Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions“, Future Generation Computer Systems, 2013.

X. CERP-IoT, Vision and Challenges for Realising the Internet of Things, March 2010

XI. Toby Jaffey, „MQTT and COAP, IoT Protocols“, Eclipse Newsletter, February 2014.

XII. Paul Duffy, „Beyond MQTT: A Cisco view on IoT Protocols“, Cisco Blog, April 2013.

XIII. XMPP Standards Foundation, http://xmpp.org/about-xmpp/technology-overview/

XIV. Stan Schneider, „Understanding The Protocols Behind The Internet Of Things“, Electronic Design, October 2013.

XV. G. Santucci, “1.1 The Internet of Things: Between the Revolution of the Internet and the Metamorphosis of Objects”, Vision and Challenges for Realising the Internet of Things, March 2010.

XVI. European Commission. “Communication on future networks and the internet.” Brussels, September, 2008.

XVII. European Commission, „Internet of Things: A 14-point Action Plan“, Brisel, lipanj 2009.

XVIII. T. Wachtel, “10th Meeting of the Internet of Things Expert Group Meeting Minutes” Brussels, November 2012.

XIX. RAND Europe, “Europe’s policy options for a dynamic and trustworthy development of the Internet of Things.”, SMART 2012/0053, August 2013

XX. National Institute for Standards and Technology (NIST), “NIST Releases Final Smart Grid 'Framework 2.0' Document”, February 2012.

XXI. Department of Homeland Security, http://www.dhs.gov/st-snapshot-detect-protect, rujan 2013

XXII. Federal Trade Commission, “Public Comments: # 484; FTC Seeks Input on Privacy and Security Implications of the Internet of Things; FTC Project No. P135405; Commission Staff to Conduct Workshop on November 21, 2013 in Washington, DC“, July 2013

XXIII. Federal Trade Commission, Internet of Things Workshop, Washington DC, http://ftc.gov/bcp/workshops/internet-of-things/, November, 2013

XXIV. http://www.hakom.hr/default.aspx?id=116&subID=3366

XXV. ECC Report 153 on Numbering and Addressing in Machine-to-Machine (M2M) Communications – Nov. 2010 (http://www.cept.org/files/5424/documents/ECC%20REP%20153%20-%20Numbering%20and%20Addressing%20in%20Machine-to-Machine%20%28M2M%29%20Communications.pdf)

XXVI. Rafiullah Khan, Sarmad Ullah Khan, Rifaqat Zaheer and Shahid Khan, “Future Internet: The Internet of Things Architecture, Possible Applications and Key Challenges”, 10th International Conference on Frontiers of Information Technology, 2012.

XXVII. Boban, M., Information security and the right to privacy in digital economy– the case Of Republic of Croatia // Zbornik radova 37. međunarodnog skupa za za informacijsku i komunikacijsku tehnologiju, elektroniku i mikroelektroniku - MIPRO 2014. Zagreb : MIPRO, 2014. 1687-1692

XXVIII. Boban, M., Upravljanje sigurnosnim rizicima i krizno upravljanje u mrežnoj komunikaciji // .Zagreb, 2014. 549-572

XXIX. Boban, Marija, Pravo na privatnost i pravo na pristup informacijama u suvremenom informacijskom društvu. // Zbornik radova Pravnog fakulteta u Splitu. 49 (2012) , 3 (105); 575-598

Promoting Health for Chronic Conditions: a Novel Approach that integrates Clinical and Personal Decision Support

I. Lasorsa*, M. Ajčević*, P. D'Antrassi*, G. Carlini **, A. Accardo* and S. Marceglia*
* Department of Engineering and Architecture, University of Trieste, Trieste, Italy
** R&D Department, TBS Group, Area Science Park, Trieste, Italy
irene.lasorsa@phd.units.it

Abstract – Direct and indirect economic costs related to chronic diseases are increasing in Europe due to the aging of population. One of the most challenging goals is to improve the quality of life of patients affected by chronic conditions, and enhance their self-management. In this paper, we propose a novel architecture of a scalable solution, based on mobile tools, aimed to keep patients with chronic diseases away from acute episodes, to improve their quality of life and, consequently, to reduce their economic impact. Our solution aims to provide patients with a personalized tool for improving self-management, and it supports both patients and clinicians in decision-making through the implementation of two different Decision Support Systems. Moreover, the proposed architecture takes into account the interoperability and, particularly, the compliance with data transfer protocols (e.g., BT4/LE, ANT+, ISO/IEEE 11073) to ensure integration with existing devices, and with the semantic web approaches and standards related to the content and structure of the information (e.g., HL7, ICD-10 and openEHR) to ensure correct sharing of information with hospital information systems, and classification of patient behaviors (Coelition). The solution will be implemented and validated in future study.

I. INTRODUCTION

In the European region, the population aged 65 and over is projected to rise from 15% in 2000 to 23.5% by 2030 [1]. This trend is considered as one of the reasons for the growing burden of chronic conditions and diseases. Chronic diseases, such as coronary heart disease, chronic respiratory diseases and diabetes, are the leading cause of mortality and morbidity in Europe [2]. The home environment and the informal caregivers are the most common care solution for chronic patients. The interaction among patients/caregivers and healthcare providers needs to be strengthened to enhance adherence to the clinical pathway and monitor health outcomes.

Despite the increasing burden of chronic illnesses, healthcare is still largely built around an acute, episodic model of care that is ill-equipped to meet the requirements of those with chronic health problems [3]. Mobile technologies can support a healthcare redesign based on disease management programs and integrated care models [4], [5], [6], [7].

Mobile technologies may act as a complementary tool to provide support and motivation to regular patient's self-monitoring of health parameters (e.g., blood pressure, heart rate, blood glucose) [8]. The resulting empowerment may help care recipients to achieve personal health objectives, modify lifestyle patterns and/or high-risk behaviors for an optimal management of their chronic condition.

At present, over 35,000 mobile health (mHealth) apps are available for iOS and Android [9], but, despite the widespread diffusion, a lot of challenges are still to be faced, such as apps security, data safety, regulatory framework, interoperability, certification and quality, and integration in existing healthcare models [10], [11].

In the last decades, Decision Support Systems (DSSs), defined as any electronic or non-electronic system designed to aid directly in clinical decision making [12], gained increasing attention thanks to the promise to use characteristics of individual patients to generate patient-specific assessments to support clinicians' decision-making.

On the other hand, even though mobile technologies can ground the development of patients' self-management systems, the development and use of DSSs for supporting patients' decision-making is still lacking.

In this paper, we preliminarily present the architecture of a comprehensive, mobile-based, patient-oriented solution aimed to diminish acute episodes of patients with chronic diseases in order to reduce their economic impact on the healthcare system and to improve their quality of life. These goals can be reached through the development of a new model of relation between patient and disease, based on a mobile technology that integrates Clinical DSS (CDSS) for physicians and Lifestyle DSS (LDSS) for patients. The model is primarily focused on self-management of care, instead of handling the mere information exchange between patients and healthcare professionals.

II. REQUIREMENTS

Clinical physician-oriented DSSs are largely used and implemented while the new challenge is to create and integrate a patient-oriented DSS as an innovative tool

designed with the aim of providing patients with more knowledge about their conditions and the disease they are affected by, and guide them towards a healthy lifestyle.

The majority of the mHealth existing solutions (e.g., *glooko*, *bloodpressurecompanion*, *moticon*, *Metria*™ *Wearable Sensor*, *nuubo*®) provide acquisition and processing of pathology related to biomedical data and health status. However only few of them (e.g., *mySugr*, *IndiGO*™) provide integration with clinical evidence and guidelines, and prediction of health status.

The most common mHealth solution state-of-art limitations concern:

- Lack of interoperability and scalability which also result in high initial costs for citizens [13];
- Lack of standards: many systems work only on proprietary platforms, from sensors to software (e.g., iPhone) and integrations with open platform such as Android are limited [14], [15];
- Cultural challenge: patients are unaccustomed to use technologies for clinical purposes, as they usually prefer to contact doctors directly [16];
- Need to improve mobile interface user-friendliness and user experience [17];
- Need to increase the mutual information flow between healthcare professionals and patients rather than keeping with one direction flow of data from patients to clinicians [16];
- Need to guarantee privacy, integrity and confidentiality of the system information.

The novelty of our proposal mainly concerns the requirements identified as key features of our system, that overcome the aforementioned limitations of the solutions currently available in the literature and market. Particularly, we defined the following requirements:

- Open architecture;
- Coexistence of clinical and lifestyle decision support systems;
- Compliance to the existing standards and regulations;
- Scalability and versatility of the system customizable to patients affected by different chronic diseases.

III. ARCHITECTURE

The proposed architecture is based on three layers (i.e., patient self-management, multiservice storage system and clinician supervision), as shown in Fig. 1.

In order to allow the customization for different needs of different user groups (e.g., pathology, age) the use of loosely coupled set of heterogeneous modules were used. These modules are organized as apps, services and processes running either in foreground or background according to the specific implemented functionality. Regarding the scalability, our proposal is an architectural paradigm able to sustain a system designed to seamlessly handle changes and reconfigurations both in its front-end and in its back-end (Fig.1). Hence, introducing new sensors or apps in the input space or rules in the back-end that lead to new users' functionalities (in the form of apps and tools) available at the output space, requires only minimal intervention. In order to comply with the privacy requirements, no sensitive clinical data are stored on the mobile device: personal, sensors and therapy information are stored in a remote DB that ensures privacy and provides a highly concentrated source of data, shared by all the mobile SW modules. Data integrity is foreseen through algorithms that can assure that authenticate users are handling unchanged and original information (cryptographic hash functions, such as Message Digest 5 or Secure Hash Algorithm 1). To ensure confidentiality of data exchanged through the network from unauthorized/authorized agents, our solution is based on an encryption/decryption hybrid approach using symmetric and asymmetric algorithms to guarantee data confidentiality between the communication processes.

The patient is engaged in the self-management starting from data acquisition using biomedical wireless sensors and also through input of self-reported data.

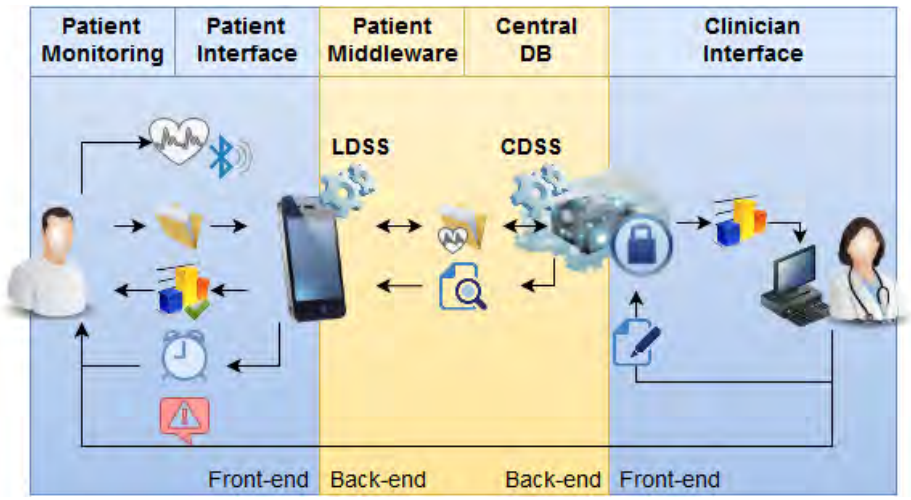


Figure 1. Overview of the proposed architecture

The collected data are processed firstly in the mobile device where they are validated, sent and, subsequently, stored in a central DB. The stored data (i.e., acquired new data, historical data, therapy, etc.) are extracted to feed a LDSS inference engine in order to provide feedback, recommendations and therapy reminders to the patient. Furthermore, interfaced with central DB, a CDSS provides a diagnostic and prognostic support to the clinician accessing the platform through a dedicated portal. The physician can directly alert the patient in emergency cases or change the prescribed therapy.

The access to the central DB requires basic authentication, SSL channel and information cryptography. For the content and structure of the information semantic, web approaches and standards (e.g., HL7, ICD-10, and OpenEHR) are used to ensure correct sharing of information with hospital information system, and classification of patient behaviors (Coelition).

The components of the architecture are briefly detailed in the following sessions.

A. Patient Monitoring

The first layer of the platform consists of a sensing device network which feeds the system with patient biomedical data needed for patient monitoring. Heterogeneous pathologies require different modalities/sensor types (e.g., weighing scale, blood pressure monitor). Consequently, the mobile application considers the compliance with transfer and data protocols (e.g., BT4/LE, ANT+, ISO/IEEE11073) to ensure integration with existing wearable devices.

B. Patient Interface

The patient interacts with the mobile application through its user-friendly interface, combining various input and output modalities (such as seeing/vision, hearing/audition, haptic/tactile). The main functionalities of the interface are:

- displaying questions to the patient;
- receiving feedback from the patient;
- providing push messages (i.e., alerts, lifestyle suggestions, reminders, achieved goals);
- enhance patients' involvement, engagement and awareness through smart design and interaction, and social networking;
- displaying biomedical parameters.

C. LDSS: Patient's Lifestyle DSS

The LDSS correlates processed and validated data, coming from the central DB (from questionnaires, biomedical data, etc.) and lifestyle suggestions. This DSS provides patients only with recommendation related to healthy behaviors. It supports patients for staying away from acute episodes of their chronic condition through lifestyle suggestions.

The core of LDSS is represented by an inference engine based on principles of probabilistic graphical models. For each specific condition addressed (e.g., Diabetes, Coronary Heart Disease CHD, Chronic Obstructive Pulmonary Disease COPD and Chronic Hypertension CHTN) a dedicated probabilistic model is implemented considering guidelines and existing predictive models.

The LDSS uses information stored on the central DB in order to provide customized advices related to the management of the patient's conditions. Particularly, the LDSS sends periodic advices concerning one or more of the following:

- identification of undesired trends in measurements;
- advice on healthy lifestyle including physical activities and suggested exercises;
- dietary advice;
- reminders regarding medications.

The LDSS knowledge base is implemented using the Unified Medical Language System (UMLS) to represent concepts and relationships.

D. Patient Middleware

The previous modules are orchestrated around a middleware layer, conceived as a bus acting as the glue of the mutually interacting app(s), services and processes of the patient side architecture. The middleware offers all modules abstraction mechanisms to provide a generic API, suitable for the most common needs for customer extensions or third party apps and process integration, so that other mobile embeddable devices can be integrated directly onto the patient framework. Its main purpose is to direct and integrate processes and services. Table 1 shows the core capabilities provided by the middleware bus layers and it is also capable to proxy third party devices operative paradigms into our execution model for input data management and data conversion.

E. Central DB

The central DB is a database located in a separate and secure data center that provides a highly concentrated source of data and reports of randomized controlled patient's lifestyle. Records contain the biomedical information and preferences of the patient collected by the LDSS and the related therapy and suggestions provided by the clinician through the CDSS.

According to the Service-Oriented Architecture (SOA), the external database comes with a web service on server side in order to manage data flows with the central DB. An Input Data Manager Tool categorizes data in functional structures and stores them in the mobile DB of the device as RAW data. The latter are collected and processed by DSS in order to extract parameters and write them on the DB. Thus, the DB represents the source of data for all the tools integrated in middleware.

One of the key features in the pre-processing stage is the pre-validation of the biomedical data: if the measurements are out of predetermined safe ranges, defined according to different and specific situations, the system asks the patient to repeat the measurement.

F. CDSS: Clinical DSS

The CDSS provides automated advices to healthcare professionals by constantly monitoring and interpreting the patients' data. This goal is achieved by using an intelligent diagnostic system based on existing probabilistic models (e.g., Archimedes [18], EAGLE [19], ADA guidelines [20], Seattle model [21], MAGGIC [22], ESC guidelines) that, according to the peculiarities of each patient, identifies trends and indicates possible undesired conditions or worsening of patient's health.

The CDSS provides advices intended for healthcare professionals. It is responsibility of the professional caregiver to interpret the output generated by the CDSS and decide whether to take further actions or ignore the alert provided by the CDSS. If the caregiver decides to take further actions, this is done by sending feedback (i.e., push messages) through the ICT network to the LDSS or contacting directly patients through SMS or phone calls.

G. Clinician Interface

The Clinician Interface foresees different User Profiles (Operators, Nurse, General Practitioner, and Specialist) that follow different rules in order to remotely monitor the patient by mean of several tools, such as:

- Alarm Monitor
- Electronic Health Record
- Signal Visualization
- Diagnostic Tools (advanced signal processing)
- Notifications/Lifestyle Advice

The Clinician Interface is developed in accordance to Medical Device Directive compliance.

TABLE I. CORE CAPABILITIES OF MIDDLEWARE LAYER

Capability	Description
Invocation	support for synchronous and asynchronous communication protocols, service mapping (locating and binding)
Routing	addressability, content-based routing, rules-based routing, policy-based routing
Mediation & Protocol Conversion	adapters, protocol transformation, service mapping, as we are expecting to integrate different components, with likely different origins and thus semantics
Messaging	message processing, message transformation and message enhancement
Process choreography	implementation of complex business processes
Service orchestration	coordination of multiple implementation services exposed as a single, aggregate service
Complex event processing	event interpretation, correlation, pattern matching
Quality of service	security (encryption and signing), reliable delivery, transaction management
Management	monitoring, audit, logging, metering, admin console
Protocol Conversion	comprehensive support for topical communication protocols service standards
Security	a standardized security model to authorize, authenticate and audit use of the ESB
Transformation	facilitation of the transformation of data formats and values, including transformation services (often via XSLT or Query) between the formats of the sending application and the receiving application
Validation	validation against schemas for sending and receiving messages
Abstraction	the provision of a unified abstraction across multiple layers
Routing and Transformation	routing or transforming messages conditionally, based on a non-centralized policy (without the need for a central rules engine)
Queuing and staging	queuing, holding messages if applications/network connections temporarily become unavailable or work at different speeds
Commodity Services	provisioning of commonly used functionality as shared services depending on context

IV. DISCUSSION

Here we proposed a novel architecture based on mHealth technologies for improving quality of life of patients affected by chronic conditions, and empowering their self-management. The architecture's requirements are thought to overcome the main limitations of the solutions currently available in the literature and on the market. The modular architecture allows the technology to be customizable, scalable and applicable to different kinds of chronicity. The proposed architecture is based on the standards that guarantee interoperability with other medical devices and users' privacy. Furthermore, the integration of two different Decision Support Systems, one dedicated to patients and one to clinicians, helps both patients and clinicians in decision-making for a better management of chronic condition oriented to decrease acute episodes.

Our system may improve and enhance the communication and information flow between patients and medical professionals through the direct involvement of clinicians in case of need.

However, in order to further boost patients' wellbeing and engagement, the system may support both patients' compliance with a good self-management and a positive mental status. Indeed, even though the DSS can improve the compliance of patients with their clinical pathway, an additional technique (i.e., gamification), could be added for strengthening the link between users and their good health. Moreover, an additional approach for enhancing a positive psychological status (i.e., mindfulness) should be taken into consideration in order to improve patients' quality of life.

The proposed solution will be implemented and validated in future work, also adding further features such as gamification and mindfulness.

REFERENCES

[1] J. Pomerleau, C. Knai , and E. Nolte, "The burden of chronic disease in Europe," in Caring for people with chronic disease. A health system perspective, E. Nolte and M. McKee, Eds. Maidenhead: McGraw-Hill Education, 2008, pp. 15-42.

[2] R. Busse R, Tackling chronic disease in Europe: strategies, interventions and challenges, No. 20, WHO Regional Office Europe, 2010.

[3] M. M. Harbers and P. W. Achterberg, Europeans of retirement age: chronic diseases and economic activity, Bilthoven: RIVM, 2012.

[4] R. Istepanian , S. Laxminarayan, and C. S. Pattichis "Introduction to mobile m-health systems," in M-health: Emerging Mobile Health Systems, Springer, Berlin Heidelberg, 2006, p. 3.

[5] M. Kay , J. Santos, and M. Takane mHealth: New horizons for health through mobile technologies, World Health Organization, 2011, pp. 66-71.

[6] R. Luanrattana, K. T. Win, J. Fulcher, and D. Iverson, "Mobile Technology Use in Medical Education," Journal of Medical Systems, vol. 36, pp. 113-122, 2012.

[7] Y. SC, "Mobile applications and 4G wireless networks: A framework for analysis," Campus-Wide Information Systems, vol. 29, pp. 344-357, 2012.

[8] S. Marceglia, S. Bonacina, V. Zaccaria, C. Pagliari, and F. Pinciroli, "How might the iPad change healthcare?," J R Soc Med, vol. 105, pp. 233-41, 2012.

[9] A. Sunyaev, T. Dehling, P. L. Taylor, and K. D. Mandl, "Availability and quality of mobile health app privacy policies," Journal of the American Medical Informatics Association, amiajnl-2013, 2014.

[10] The PLOS Medicine Editors, "A reality checkpoint for mobile health: three challenges to overcome," PLoS Med, vol. 10, e1001395, 2013.

[11] P. Quinn, A.K. Habbig, E. Mantovani, and P. De Hert, "The data protection and medical device frameworks-obstacles to the deployment of mHealth across Europe?," European journal of health law, vol. 20, pp. 185-204, 2013.

[12] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success," Bmj, vol 330, pp. 765, 2005.

[13] Mobile Health Alliance (2013) at http://mhealthalliance.org.

[14] D. Lou, X. Chen, Z. Zhao,Y. Xuan, Z. Xu, H. Jin, et al., "A Wireless Health Monitoring System based on Android Operating System," IERI Procedia, vol. 4, pp. 208-215, 2013.

[15] B. Martínez-Pérez, I. De la Torre-Díez, and M. López-Coronado, "Mobile Health Applications for the Most Prevalent Conditions by the World Health Organization: Review and Analysis," Journal of medical Internet research, vol 15, e120, 2013.

[16] R. Elangovan, and S. Arulchelvan, "A study on the role of mobile phone communication in tuberculosis DOTS treatment," Indian journal of community medicine, vol. 38, pp. 229-233, 2013.

[17] O.Vélez, P. Okyere, S. A. Kanter, and S. Bakken, "A usability study of a mobile health application for rural Ghanaian Midwives," Journal of Midwifery & Women's Health, vol. 59, pp. 184-191, 2014.

[18] D. M. Eddy, and L. Schlessinger, "Archimedes: a trial-validated model of diabetes," Diabetes Care, vol. 26, pp. 3093-3101, 2003.

[19] E. Mueller, S. Maxion-Bergemann, D. Gultyaev, S. Walzer, N. Freemantle, C. Mathieu, et al., "Development and validation of the Economic Assessment of Glycemic Control and Long-Term Effects of diabetes (EAGLE) model," Diabetes technology & therapeutics, vol. 8, pp. 219-236, 2006.

[20] American Diabetes Association, "Standards of medical care in diabetes-2016 Abridged for primary care providers," Clinical Diabetes, vol. 34, pp. 3-21, 2016.

[21] W.C. Levy, D. Mozaffarian, D.T. Linker, S.C. Sutradhar, S.D. Anker, A.B. Cropp, et al., "The seattle heart failure model: Prediction of survival in heart failure," Circulation, vol. 113, pp. 1424-1433, 2006.

[22] S.J. Pocock, C.A. Ariti, J.J. McMurray, A. Maggioni, L. Køber, I.B. Squire, et al., "Predicting survival in heart failure: a risk score based on 39372 patients from 30 studies," European heart journal, ehs337, 2012.

A taxonomy of localization techniques based on multidimensional scaling

Biljana Risteska Stojkoska
Faculty of Computer Science and Engineering (FCSE)
University "Ss. Cyril and Methodius"
Skopje, Macedonia
biljana.stojkoska@finki.ukim.mk

***Abstract*—Localization in Wireless Sensor Networks (WSNs) has been a challenging problem in the last decade. The most explored approaches for this purpose are based on multidimensional scaling (MDS) technique. The first algorithm that introduced MDS for nodes localization in sensor networks is well known as MDS-MAP. Since its appearance in 2003, many variations of MDS-MAP have been proposed in the literature. This paper aims to provide a comprehensive survey of the localization techniques that are based on MDS. We classify MDS-based algorithms according to different taxonomy features and different evaluation metrics.**

Keywords—taxonomy; localization; wireless sensor networks; multidimensional scaling; positioning

I. INTRODUCTION

Wireless Sensor Networks (WSNs) have been a hot research area in the last decade. The potential application of WSN include different military and civilian application for monitoring (environmental, healthcare, structural health, smart home or smart grid)[1].

In the recent years, the concept of scattered devices dedicated to one-purpose application has been replaced with Internet of Things (IoT) and machine-to-machine (M2M) communication for general-purpose application [2]. However, the basic principles and challenges associated with WSNs remain attractive for the research communities. Moreover, with the extended range of potential WSN applications, these challenges are being reinvented and redesigned.

Localization in WSN refers to the process of discovering the locations of the nodes in the network. Many of the algorithms for WSN localization proposed in the literature are based on multidimensional scaling technique (MDS). This approach is very robust even if the network has only small portion of anchor nodes (nodes with a priori known location). Since 2003, when this approach was firstly presented on INFOCOM as MDS-MAP [3], new modifications and improvements were constantly developed by the researchers worldwide.

Multidimensional scaling (MDS) is a set of techniques used for reducing the dimensionality of the data (objects). MDS visualizes the results in order to show hidden structures in the data [4]. MDS algorithm uses the distances between each pair of object as an input and generates 2D-points or 3D-points as an output.

Multidimensional scaling as a technique for WSN localization consists of the following three steps:

- Step 1: Calculate the distances between every pair of nodes and generate a distance matrix that serves as an input to the step 2.
- Step 2: Apply multidimensional scaling (classical or non-metric) to the distance matrix. The first largest eigenvalues and eigenvectors give a relative map with relative location for each node.
- Step 3: Transform the relative map into absolute map using sufficient number of anchor nodes.

The basic pipeline of MDS algorithm for nodes localization is shown on Fig. 1. The input to the algorithm is available distance measurements between the nodes. The

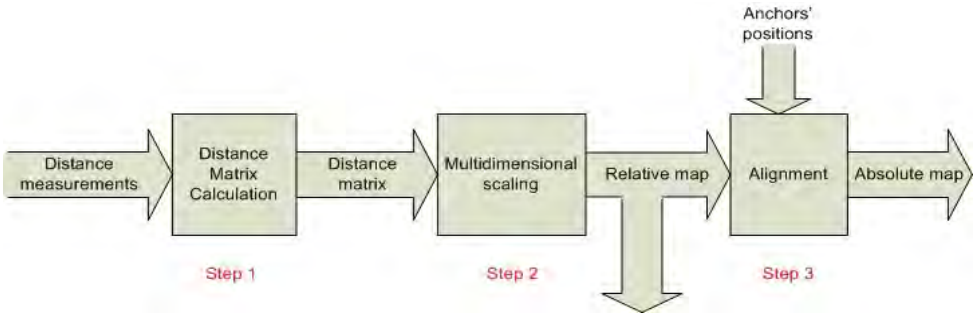


Fig. 1. Pipeline of MDS-based localization techniques.

output can be obtained after step 2 as a relative map of the network, or after step 3 as an absolute map of the network.

In this paper, we present taxonomy of localization techniques based on multidimensional scaling algorithm. This taxonomy can be easily mapped to other localization approaches that use similar nonlinear dimensionality reduction techniques. For example, self organizing maps (SOMs) [5][6], manifold learning algorithm [7] or semidefinite programming [8][9] can follow same or similar taxonomy like the one presented in this paper. However, most of these approaches are represented with limited number of publications compared to MDS with respect to solving WSN localization problem.

The rest of this paper is organized as follows. In the next section, we present the taxonomy features used to distinguish different localization technique based on MDS. In the third section, we elaborate the metrics used for evaluation of the MDS-based techniques. Section four surveys different MDS techniques and qualitatively compares them regarding different taxonomy features and evaluation metrics. This paper is concluded in section five.

II. TAXONOMY FEATURES OF MDS-BASED TECHNIQUES

We identify five taxonomy features of localization techniques based on multidimensional scaling (Fig.2):

A. Computational model

WSNs are characterized with limited resources (CPU, memory, energy, etc.). Hence, choosing the most suitable computational model is not a trivial task. We define the computation model with respect to nodes' tasks in MDS computational pipeline. We indentified three main computational models.

1) Centralized approach

In centralized localization approaches, all distance measurements between neighboring nodes are collected at one central point (base station or sink). The sink node has three main tasks:

- to construct the distance matrix
- to apply MDS on the distance matrix
- to convert the relative map into absolute map (if anchors' positions are available)

which correspond to the three steps of the algorithm.

The main drawback of this model is its communicational overhead, as all distance measurements need to be propagated to the sink node. For a large scale networks, this requires retransmission through many intermediate nodes, which leads to additional energy consumption. On the other side, the step 2 of the algorithm will benefit from this model, as MDS will tend to minimize the overall positioning error.

2) Fully distributed approach

In distributed approaches, the distance measurements are not propagated to the sink node. Instead, they are broadcasted only in the near proximity. Each node is responsible to collect the distance measurements from its one-hop or two-hop neighbors. Then, the node uses these measurements to construct the distance matrix and to perform MDS in order to obtain the relative map of its own neighborhood. The relative maps from all nodes can later be propagated to the sink node, where they can be merged into one global map. In this approach, step 1 and step 2 are performed at the nodes' side, while step 3 (if applicable) is performed at the sink side.

As a special case of fully distributed approach is self-localization using location assistants (LA). Location-unaware nodes measure a set of distances to LAs, which are stationary or mobile nodes (e.g. airplane) with stronger radio signal. LAs broadcast their absolute location. Having radio signal strengths and LAs coordinates, location-unaware nodes perform MDS to obtain its own location [10].

3) Cluster-based approach

This approach is also known as hierarchical. Here, the network is divided into clusters which represent group of nodes in close proximity. One node in the cluster (known as cluster-head) is responsible to collect the distance measurements from all members of that cluster, to construct the distance matrix and to obtain relative map of its own cluster.

This approach is very similar with the fully distributed, except that only the cluster heads are responsible to run the algorithm. It is especially suitable for irregularly shaped networks (C-shape, O-shape, H-shape, etc.). If applied on networks with regular shape, this approach performs worse than centralized approach, due to cumulative error from the merging process.

B. Dimensionality of the network

We define three types of network regarding network dimensionality: two-dimensional (2D), three-dimensional (3D) and surface networks. Two-dimensional network is suitable for flat indoor and outdoor environments (flat agricultural field, one-floor apartment, etc.). Most of the outdoor networks fit into surface type (multi-floor shopping mall, hill, valley, etc.), while underwater networks can be considered as pure three-dimensional networks.

When applying MDS pipeline, step 2 and step 3 differ for different networks' dimensionality (Table 1).

TABLE I. NETWORK DIMENSIONALITY REGARDING MDS PIPELINE

Network dimensionality	Number of eigenvalues (step 2)	Minimal number of anchors needed for alignment (step 3)
2D	first 2 largest	3
3D	first 3 largest	4
surface	first 3 largest	4

C. Deployment environment

Regarding the environment where the nodes are deployed, we consider indoor and outdoor networks. The main difference between these types is availability of anchor nodes. In outdoor environment, nodes equipped with Global Positioning System (GPS) can become anchor nodes. Since GPS signals are not available in indoor environments, the role of anchors belongs to static nodes that are fixed with the

buildings' infrastructure and their location is usually recorded manually. Underwater network is considered as outdoor; still, GPS signals are not available.

Other important difference is the radio propagation model, which is dependent on environmental condition (temperature, humidity, etc) and is highly variant in outdoor environment.

D. Anchors

Depending on the WSN application, in some cases there is no need to obtain the absolute location of the nodes. Users are interested in node location with respect to other nodes in the network. In such cases, step 3 of the pipeline is not needed.

If anchor nodes are available, the relative map can be easily translated into absolute map using geometric transformations (translation, rotation, and reflection). At least three (four) anchors are needed for 2D (3D) networks.

Some algorithms based on MDS allow random placement of the anchors, while other have more strict requirements about the anchors' position, i.e. allow only anchors placed at the edges of the deployment area.

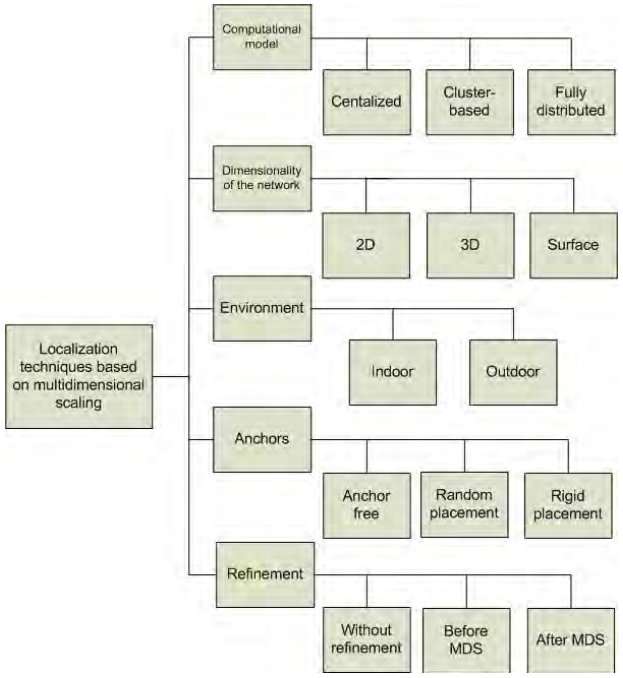


Fig. 2. Taxonomy of MDS-based localization techniques

E. Refinement

MDS is very accurate technique for dimensionality reduction. If the correct distance matrix is given as input, MDS algorithm will reconstruct the map of the network without error. But, calculating distance matrix for networks where only distances between neighboring nodes are known is not a trivial task. This problem in MDS-MAP is solved by applying Dijkstra's or Floyd's all pairs shortest path algorithm [11]. Dijkstra's algorithm is a graph search algorithm that solves the single-source shortest path problem. In WSN localization problem, the sensor network is represented as a graph with non-negative edge path costs, while the real, Eucledian distance between two non-neighboring nodes is replaced with the distance calculated using Dijkstra (or other) algorithm. But the assumption that Dijkstra distance between two nodes correlates with their Euclidean distance is hardly true. This approximation produces an error, i.e., the positions obtained after alignments usually differ from the real positions. The error is bigger when the nodes are in multi-hop communication range, which is a common case in obstructed environments. It is usually caused by the presence of obstacles or terrain irregularities that can obstruct the line-of-sight between nodes or cause signal reflections.

Refinement process aims to reduce this error using different optimization techniques. Refinement can be applied in step 1 and in step 3 from the MDS pipeline:

1) Refinement before MDS

This refinement aims to minimize the inter-node distance error between non-neighboring nodes in the distance matrix. Variations of MDS-MAP use other approaches for distance matrics calculation, that are not based on Dijkstra's or Floyd's all pairs shortest path algorithm. Those techniques apply additional geometric constrants and produce better distance matrices compared to Dijkstra's or Floyd's algorithms.

2) Refinement after MDS

This refinement tends to minimize inter anchor distances error. The MDS-based techniques generate the coordinates of all nodes, including the anchor nodes. Moreover, after the alignment, the predicted positions for all nodes are obtained, including the anchor nodes. Although predicted coordinates of the anchors are usually discarded because their exact positions are known in advance, they can be used to refine other predicted positions. Through refinement, the predicted anchor positions are shifted toward real anchors positions. The refinement algorithms are usually iterative, and this shifting is actually performed for each node, toward a position for which the localization error is smaller.

III. EVALUATION METRICS

In this section, we are describing different metrics used to evaluate the accuracy and the quality of the localization techniques.

A. Localization error

The main objective of localization is to predict the nodes location as accurate as possible. In order to evaluate the prediction error, different expressions of the error are adopted in the literature. The difference between the real and the predicted positions is known as estimation error, localization error or positioning error. For WSN that consists of W unknown nodes, where (x_i, y_i, z_i) is the real position and (x_i', y_i', z_i') is the predicted position of i-th node, the localization error can be expressed as:

1) Average localization error(ALE)

$$ALE = \frac{1}{W}\sum_{i=1}^{W}\sqrt{(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2} \quad (1)$$

2) Root-mean-square error (RMSE)

$$RMSE = \sqrt{\frac{1}{W}\sum_{i=1}^{W}(x_i - x_i')^2 + (y_i - y_i')^2 + (z_i - z_i')^2} \quad (2)$$

Localization techniques can be evaluated by comparing the localization accuracy obtained using (1) or (2) with the corresponding Cramer-Rao lower bound (CRLB), which is a lower bound on the variance of the estimator for locations. CRLB is used to assess the optimum achievable localization accuracy which can be attained with the available measurement set. CRLB is computed using the Likelihood function in which the covariance of the measurement vector enters as a factor [12].

B. Radio range error

Different techniques can be used for inter-node distance measurements, like Radio Signal Strength Indicator (RSSI), Time of Arrival (ToA) or Time Difference of Arrival (TDoA) [13].

Among them, RSSI is the most popular since it doesn't require any additional hardware. On the other side, RSSI is not that accurate, due to the different reasons, like different propagation model or device calibration. When modeling the radio range, we need to consider radio range error which is at least 20%, but more often is greater [13].

C. Communication cost

WSN has limited energy resources. It is known that data transmission spends much more energy compared with data processing. Hence, it is very important in WSN to reduce the number of transmissions and to apply local processing whenever possible.

Centralized approaches for WSN localization require all distance measurements to be transmitted to the sink node. In multi-hop network, this requirement can lead to battery drain.

D. Computation cost

Although local computations are more preferable for energy saving, they are not always fusible due to CPU limitations. WSN nodes are usually equipped with limited CPU and memory capacities, which cannot handle exhaustive computations, or at least not in expected time span.

Considering communication cost and computation cost, the ideal algorithm for localization should maintain a tradeoff between energy saving and required localization accuracy.

E. Network density

MDS-based techniques require connected network, which means that there is a path between every pair of nodes in the network. Moreover, greater network density leads to better localization accuracy. This is rather expected knowing that the main factor that affects the localization performance is the distance matrix. More one-hop neighbors will lead to more accurate distance matrix, which will decrease the overall localization error. The most common parameter used to represent the network density is network connectivity, which is the average number of one-hop neighbors. Therefore, radio range is a crucial parameter that directly affects network connectivity. By extending the radio range, one can increase the network connectivity, or the network density.

However, a dense network does not necessarily guarantee high accuracy, especially in networks with irregular topologies.

F. Anchor location and anchor density

Number of anchors and their location affect the localization accuracy. As expected, more anchor nodes will improve the performance of the MDS-based technique, but increasing the number of anchors does not have a crucial influence on the localization error. This is especially notable for networks with high connectivity levels.

G. Irregular vs regular network topology

Most of the techniques in the literature are evaluated on networks with regular topologies (evenly distributed nodes and anchors). To achieve this regularity, simulations are usually performed on grid, hex or random topology, although the later can barely be considered as evenly distributed.

However, real wireless networks have irregular shape due to presence of obstacles in the environment where WSN is deployed. In order to obtain the performances of the algorithm, simulations and real experiment should also consider irregular network topologies. Most simulation scenarios consider C-shape or H-shape networks to investigate the performances for irregular networks.

Localization error in the case of networks with irregular topology is generally greater than in the case of regular network topologies. Better results can be obtained if network is divided into sub-segments, which can be considered as sub-networks with regular topology.

IV. QUALITATIVE COMPARISON OF MDS-BASED TECHNIQUES

In this section, we are going to describe the most important MDS-based techniques for nodes localization in WSN. Among dozens, we have chosen the most representative, which were inspiration for new improvements. Table II provides a qualitative comparison of these techniques regarding different taxonomy features, localization error and computation cost. The later is denoted as *O(n)*, where *n* is the total number of nodes in the network.

Hereafter, we are going to present them in chronological order.

MDS-MAP is the first approach based on MDS [3]. It follows the basic MDS pipeline without any refinements. In step 1, MDS-MAP uses Dijkstra algorithm for distance matrix calculation.

MDS-MAP(P, R) is a distributed localization algorithm, where each node creates a local map within its two-hop neighbors using classical MDS algorithm [14]. Each local map is then refined with least-squares minimization.

MDS-MAP(P, O) is distributed algorithm and can be considered as an extension of MDS-MAP(P, R). The modification is the use of the ordinal MDS (instead of classical MDS) during the estimation phase [15].

TABLE II. QUALITATIVE COMPARISION OF DIFFERENT MDS BASED TECHNIGUES

			Taxonomy feature					Evaluation metric	
Author	**Year**	**Techniques' acronym**	Computational model	Anchor placement	Dimen-sionalty	Refinement before MDS	Refinement after the alignment	Localization Error	*O(n)*
Shang [3]	2003	MDS-MAP	Centralized	random	2D	NA	NA	ALE	$O(n^3)$
Shang [14]	2004	MDS-MAP (P,R)	Fully distributed	random	2D	NA	least squares	ALE	*O(n)*
Vivekanandan [15]	2006	MDS-MAP (P, O)	Fully distributed	random	2D	iterative monotone regression	least squares	ALE	*O(n)*
Stojkoska [16]	2008	CB-MDS	Cluster-based	random	2D	NA	NA	ALE	*O(n)*
Chaurasiya [17]	2013	NDEA	Centralized	random	3D	iterative distance estimation	NA	ALE, RMSE	$O(n^3)$
Stojkoska [18][19][20]	2013	IMDS	Centralized	random	2D, 3D, surface	heuristic approach based on averaging	NA	ALE	$O(n^3)$
Saeed [21]	2014	MDS-LM	Centralized	random	2D	NA	Levenberg–Marquardt	RMSE, CRLB	$O(n^3)$
Stojkoska & Saeed [22]	2015	MHL-M	Centralized	random	3D	heuristic approach based on averaging	Levenberg–Marquardt	RMSE, CRLB	$O(n^3)$
Fan [23]	2015	D3D-MDS	Cluster-based	edge	3D	NA	least squares	ALE	*O(n)*

CB-MDS is a cluster-based approach where MDS is performed at cluster level [16]. At the end, all cluster maps are merged in one global relative map, which is centrally align into absolute map. This approach has very acceptable error on the case of networks with irregular topologies.

HDEA is a novel distance estimation approach [17] that uses iterative algorithm for distance matrix calculation. In order to estimate the distance between two nodes, this approach uses group of three common neighboring nodes. This requires high network connectivity in order to produce accurate inter node distance estimation.

IMDS is an improved version of MDS-MAP, where Dijkstra algorithm from step 1 is replaced with heuristic algorithm for inter-node distance estimation based on averaging [18-20]. This algorithm has implementations for 2D, 3D and surface networks.

MDS-LM is a centralized approach that uses Levenberg–Marquardt optimization in step 3 [21].

MHL-M [22] combines IMDS [20] and MDS-LM [21]. Computational cost is the same as IMDS, but the accuracy is much better. For high connectivity, this approach achieves Cramer-Rao Lower Bound (CRLB).

D3D-MDS [23] is a 3D version of CB-MDS [16]. Hence, its benefits are not sufficiently explored on networks with irregular topologies. For regular networks, IMDS [18-20] performances are much better than D3D-MDS, which is rather expected. Cluster-based approaches should be favored only for irregular topologies.

Most of the algorithms based on MDS generally assume that RSSI is used to generate the measurement set. However, there are many approaches in the literature that assume angle of arrival (AoA) measurement set. For example, in [24], the authors propose an algorithm (RAST) that is very similar to MDS-MAP, but instead of MDS uses Singular Value Decomposition (SVD) approach for dimensionality reduction. Cluster-based solution of RAST is proposed in [25]. These approaches follow the taxonomy proposed in this paper. Regarding the evaluation metrics, instead of radio range error, these approaches should consider standard deviation of AoA measurements.

Approaches overviewed in this paper are evaluated only through simulations. In order to provide an exact quantitative comparison of these algorithms, they need to be re-implemented and evaluated under same simulation setup. In the simulation scenarios, algorithms are evaluated considering large scale networks, from 100 up to 1000 wireless nodes. Therefore, there is lack of experimental evaluation of these algorithms in the literature, and, if there are any, they are performed on small scale networks, where nodes are in close proximity to each other. With the development of new technologies, problem of WSN localization has overlapped with the problem of smartphones' localization, especially in indoor environments, where GPS signals are not available. In the near future, it is expected that new crowdsensing and/or crowdsourcing approaches can be used to experimentally verify most of the algorithms for localization, as well as those based on MDS.

V. CONCLUSION

Multidimensional scaling is a very explored concept for localization in Wireless Sensor Networks. Although initially proposed more than ten years ago, it is still attractive among researchers.

This paper is a valuable introduction for the newcomers to the field of localization in WSN, especially for those that aim to investigate and improve MDS-based techniques.

It can also serve as a tutorial for those that are implementing solutions for localization, in order to help them choose the most suitable MDS approach for their particular problem regarding different network characteristics and hardware limitations.

REFERENCES

[1] B. Risteska Stojkoska, A. Popovska Avramova, and P. Chatzimisios, "Application of wireless sensor networks for indoor temperature regulation," *International Journal of Distributed Sensor Networks* 2014 (2014).

[2] L. Mainetti, L. Patrono, and A. Vilei, "Evolution of wireless sensor networks towards the internet of things: A survey," In *Software, Telecommunications and Computer Networks (SoftCOM), 2011 19th International Conference on*, pp. 1-6.

[3] Y. Shang, W. Ruml, Y. Zhang & P. J. M. Fromherz, Localization from mere connectivity, in Proceedings of the 4th ACM International Symposiumon Mobile Ad Hoc Networking, Annapolis, Maryland, USA, pp. 201–212, (2003).

[4] T. Cox and M. Cox, Multidimensional Scaling, Chapman & Hall, London, UK, 1994.

[5] G. Giorgetti, S. K. S. Gupta, and G. Manes, "Wireless localization using self-organizing maps," in Proceedings of the 6th International Conference on Information Processing in Sensor Networks, ser. IPSN '07. New York, NY, USA: ACM, 2007, pp. 293–302.

[6] J. Hu and G. Lee, "Distributed localization of wireless sensor networks using self-organizing maps," in IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, 2008, pp. 284–289.

[7] N. Patwari and A. O. Hero III. "Manifold learning algorithms for localization in wireless sensor networks." *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. Vol. 3. IEEE, 2004.

[8] P. Biswas and Y. Ye. "Semidefinite programming for ad hoc wireless sensor network localization." *Proceedings of the 3rd international symposium on Information processing in sensor networks*. ACM, 2004.

[9] B. Stojkoska, I. Ivanoska, and D. Davcev. "Wireless sensor networks localization methods: Multidimensional scaling vs. semidefinite programming approach." *ICT Innovations 2009*. Springer Berlin Heidelberg, 2010. 145-155.

[10] X. Zhou, L. Zhang, and Q. Cheng, "Landscape-3D: a robust localization scheme for sensor networks over complex 3D terrains," in Proceedings of the 31st Annual IEEE Conference on Local Computer Networks (LCN '06), pp. 239–246.

[11] R. W. Floyd, "Algorithm 97: shortest path," Communications of the ACM, vol. 5, no. 6, p. 345, 1962.

[12] K. Regina, J. Hörst, and W. Koch. "Accuracy analysis for TDOA localization in sensor networks." *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*. IEEE, 2011.

[13] J. Zhang and L. Zhang, "Research on distance measurement based on RSSI of ZigBee," in ISECS International Colloquium on Computing, Communication, Control, and Management (CCCM '09), vol. 3, pp. 210–212.

[14] Y. Shang, W. Ruml, Improved MDS-based localization, in Proceedings of the IEEE INFOCOM Twenty-third Annual Joint Conference of the Computer and Communications society, vol.4, (2004).

[15] V.Vivekanandan, and V.W.Wong, 2006. Ordinal MDS-based localisation for wireless sensor networks. *International Journal of Sensor Networks*, *1*(3-4), pp.169-178.

[16] B.Stojkoska, D. Davcev, and A. Kulakov. "Cluster-based MDS algorithm for nodes localization in wireless sensor networks with irregular topologies," In *Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology*, ACM, 2008, pp. 384-389.

[17] V.K.Chaurasiya, N.Jain, and G.C.Nandi, 2014. A novel distance estimation approach for 3D localization in wireless sensor network using multi dimensional scaling. *Information Fusion*, *15*, pp.5-18.

[18] B. Risteska Stojkoska, D. Davcev, MDS-based Algorithm for Nodes Localization in 3D Surface Sensor Networks. In Proceedings of the 7th International Conference on Sensor Technologies and Applications (SENSORCOMM'13) 2013, pp. 44-50.

[19] B. Risteska Stojkoska and V. Kirandziska. "Improved MDS-based algorithm for nodes localization in wireless sensor networks," In *EUROCON, 2013 IEEE*, pp. 608-613.

[20] B. Risteska Stojkoska, Nodes Localization in 3D Wireless Sensor Networks Based on Multidimensional Scaling Algorithm, *International Scholarly Research Notices* 2014 (2014).

[21] N. Saeed & H. Nam, MDS-LM for Wireless Sensor Networks Localization, in Proceedings of the IEEE 79th Vehicular Technology Conference, Seoul (Korea), pp. 1–6, (May 2014).

[22] N. Saeed, and B. Risteska Stojkoska, 'Robust localisation algorithm for large scale 3D wireless sensor networks', to appear in Int. J. Ad Hoc and Ubiquitous Computing.

[23] J. Fan, B. Zhang, and G. Dai, "D3D-MDS: A Distributed 3D Localization Scheme for an Irregular Wireless Sensor Network Using Multidimensional Scaling," International Journal of Distributed Sensor Networks, vol. 2015.

[24] J.N. Ash, and L. C. Potter. "Robust system multiangulation using subspace methods." *Proceedings of the 6th international conference on Information processing in sensor networks*. ACM, 2007.

[25] A. Damir. "Distributed algorithm for anchor-free network localization using angle of arrival." *Industrial Electronics, 2008. ISIE 2008. IEEE International Symposium on*. IEEE, 2008.

Distributed real-time lift kinematic monitoring using COTS smartphones

N. Miškić-Pletenac* and K. Lenac*
* Faculty of Engineering/Department of Computer Engineering, Rijeka, Croatia
nino.mip@gmail.com, klenac@riteh.hr

Abstract - This paper presents a distributed system for real-time monitoring of lift movements and traffic analysis in a remote monitoring and predictive maintenance Internet of Things scenario. By harnessing processing and communication capabilities of today's smartphones and their integrated inertial sensors, the system allows for continuous tracking of the lift's position in the shaft without any interference with the lift operation nor interfacing to the existing lift controller. The paper describes the mathematical modeling on which the system is based, as well as the system's architecture and main components. It also presents and discusses the data gathered by the system in real lifts in multiple locations.

I. INTRODUCTION

Some of the key issues in vertical traffic planning is designing lifts and escalators with sufficient handling capacity for the building's traffic, with short passenger waiting and journey times in different hours of day, optimum use of core building space with a smooth flow of people and goods.

Having good traffic data on which to base the planning often results in improved quality of service, reduced energy consumption and increased safety. Therefore it is important to estimate lift and passenger dynamics in buildings to allow more accurate traffic analysis to be performed [1].

Efficient traffic planning depends on the characteristics and population distribution of a building so it is imperative to base the analysis on realistic traffic patterns for specific types of buildings.

Depending on type and usage of the building the traffic patterns vary significantly. For example in a residential building the traffic intensity is rather low and longer waiting times than those in commercial buildings may be acceptable. In airports and shopping centers lifts travel through just a few floors and are typically combined with escalators. Moreover lifts are designed with increased awareness for the needs of disabled people and with transportation of cargo in mind. In hospitals the traffic follows patterns all of their own depending on location of beds, services, etc. In all cases the traffic varies through the day following characteristic patterns. Traditionally some building types, like commercial buildings, have been studied more than others and patterns like the classical pattern depicted in Fig. 1 are often used for planning.

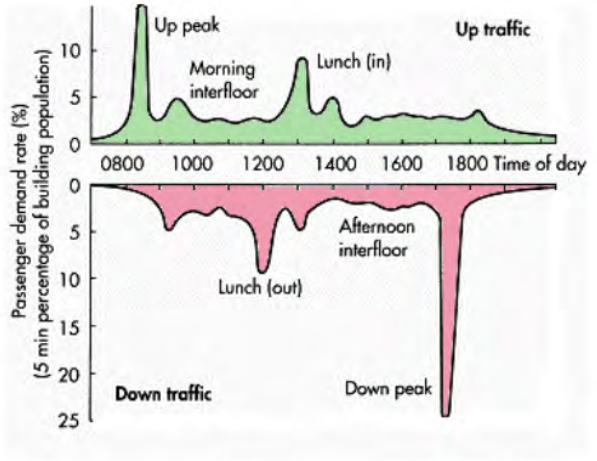


Figure 1. Passenger demand for an office building (from [2])

Traditionally these kinds of graphs have been obtained with real people observing the lift and passenger movements during the day, measuring the times, marking the number of persons entering and exiting the lift at each floor and then deriving traffic patterns by hand.

While connecting to lift controllers could automate the acquisition of lift kinematic data, it requires to obtain the permission to physically access the lift controller and to interface to it.

It is known that lift kinematics can be studied with an accelerometer by modeling the lift movements and processing acceleration data along the vertical direction. Example of a special sensor that has been designed and used for this purpose can be found in [4].

Today's devices like smartphones and tablets have embedded accelerometers sufficient for detecting and tracking lift movements independently of the orientation of the device. Moreover typical modern smartphone provides much more than sensors. In a compact device there is sufficient processing power to acquire and process the data in real time, ample storage for logging, several wireless communication channels to stream the data to a PC or servers in the cloud, and sufficient battery for days of uninterrupted operation without a power supply.

In this paper we described a system for real-time lift kinematic and traffic monitoring that harnesses most of said benefits. Moreover we describe the developed architecture suitable for an Internet-of-things scenario

which can be deployed for predictive maintenance of monitored lifts.

In section II used lift model is explained, in section III filtering of data from sensor, in section IV developed system and in section V experimental evaluation.

II. LIFT MODEL

Monitoring of lift movement can be realized in several ways. The most intuitive one is by using an altimeter, however, altimeter measurement errors can be rather high (up to 200 meters), which makes it unusable for this application. Since most of today's smartphones have accelerometers embedded, our work was focused on tracking motion by accelerometer. In this paper, measured acceleration is processed as a signal *a(t)*. Lift movement can be tracked in several different ways:

1. by detecting impulses of acceleration when lift started and stopped moving,
2. by integrating acceleration to get velocity,
3. by double integration of acceleration to get position.

Main advantage of integration is that it's universal i.e. one does not need to worry about what lift motion model is used by a particular lift. Main disadvantage of integration is accumulation of error. Integration is described with (1), (2) and (3).

$$a(t) = \frac{dv(t)}{dt} \tag{1}$$

$$v(t) = v_0 + \int_0^t a(\tau)\,\mathrm{d}\tau \tag{2}$$

$$v[n] = v_0 + \sum_{i=0}^{n} a[k] \times T \tag{3}$$

Figure 2 depicts the original (unfiltered) recording. Gray line on figure 3 shows the same recording processed with a high-pass (low-cut) filter to remove DC component of signal. Blue line shows the same recording that is also processed with a low-pass (high-cut) filter to remove noise. Figure 4 shows a different recording. Gray is signal with DC removed, blue is with noise removed. Purple shows an attempt of integration in order to get velocity from acceleration.

A. Lift types

Generally, there are two types of lifts regarding the nature of their movement:

1. lifts whose acceleration and deceleration are short,
2. lifts whose acceleration and deceleration are long and thus they accelerate through several floors.

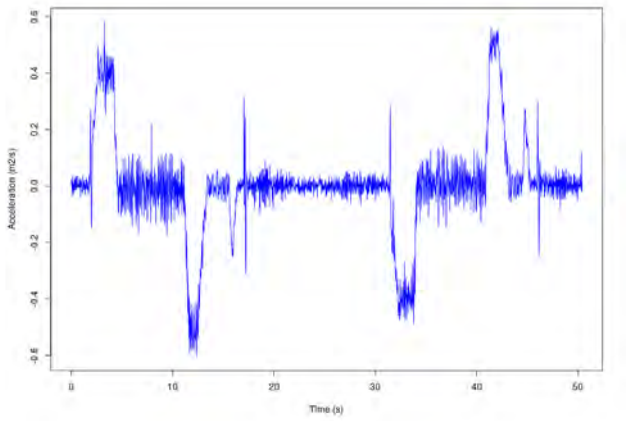


Figure 2. Plot of unfiltered recording of two lift trips

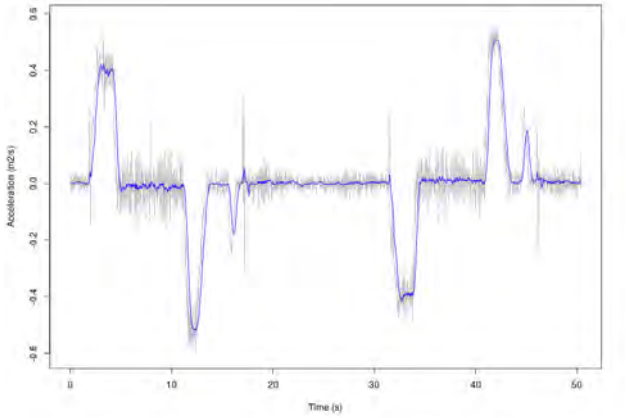


Figure 3. Plot of filtered recording of two lift trips

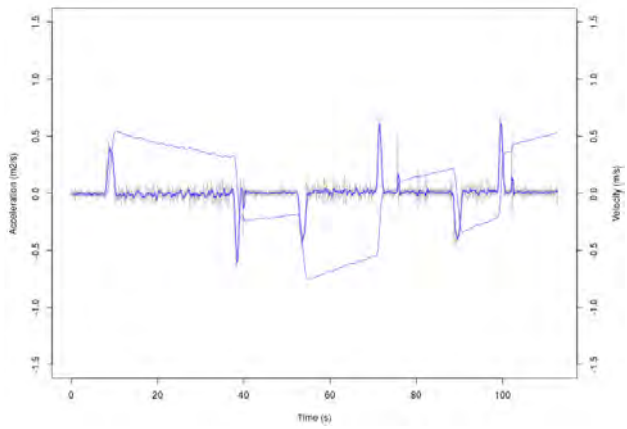


Figure 4. Example of integration with visible consequences of accumulating error

For simplicity in this paper we discuss the more common lift type (1) only.

B. Used model

As previously mentioned, used model can only be used for lifts with short acceleration/deceleration. Movement of these lifts consists of three phases:

1. acceleration,
2. movement with constant speed,
3. deceleration.

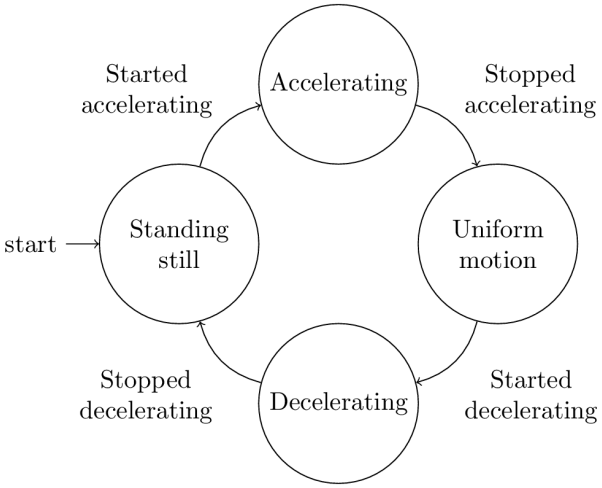


Figure 5. Lift movement detector state diagram

Since lift is most of the time moving with constant velocity, duration of lift trip is proportional to number of passed floors *k*:

$$\Delta t \mu \Delta k \tag{4}$$

$$\Delta t = t_k \times \Delta k \tag{5}$$

$$\Delta k = \frac{\Delta t}{t_k} \tag{6}$$

Lift motion model is presented with (6).

Lift states and transitions between them are depicted in figure 5.

III. FILTERING SENSOR DATA

Accelerometer sensors in smartphones are neither precise nor accurate, however it is possible to track lift movement with it, but sensor data needs to be properly filtered / prepared.

We can say that signal we get from sensor consists of several components:

- gravity
- acceleration of device/lift itself
- device orientation/position
- measurement error
- noise

For tracking lift movement we need to isolate only the acceleration of device/lift itself. This is done by a band-pass filter in order to extract medium frequencies which roughly correspond to needed part of the signal (acceleration of lift cabin). The used band-pass filter was made of one simple moving average which was used as low-pass (high-cut) filter and one exponential moving average which was used as high-pass (low-cut) filter. This can symbolically be presented as (7).

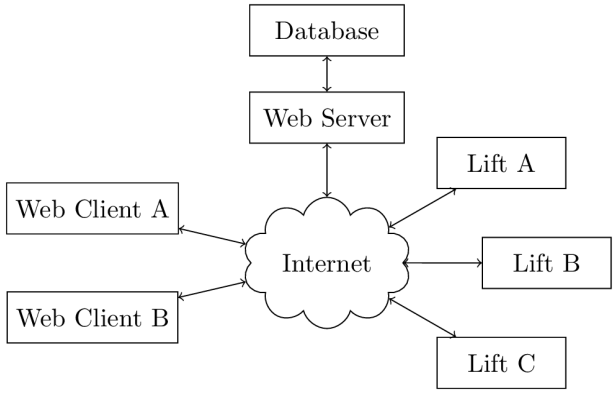


Figure 6. Overall system architecture

$$y[n] = SMA(x[n]) - EMA(x[n]) \tag{7}$$

Future work on this part of the system includes some more advanced filters like Kalman filter and correction of measurement error introduced by the accelerometer itself and/or its A/D converter.

IV. SYSTEM DESCRIPTION

The developed system consists of several parts that communicate via internet: Server, Web client (browser) and Android smartphones that are placed in lift cabins. Overall system architecture is shown on figure 6. Server consists of database, web portal for users and API interface for communication with Android application that is installed on smartphone that is in lift cabin.

A. *Android application*

Android application is installed on a smartphone that is placed in the lift cabin. This application has several responsibilities: tracking movement of lift by analyzing data from accelerometer and sending of collected data to server. For purpose of debugging, application also has a logging component. Remote control interface is also developed, for purpose of remote administration. Remote control interface is written as a classic TCP server. Since this uses nonstandard/unconventional port, problem detected while testing is that most of the unconventional



Figure 7. Android application main interface

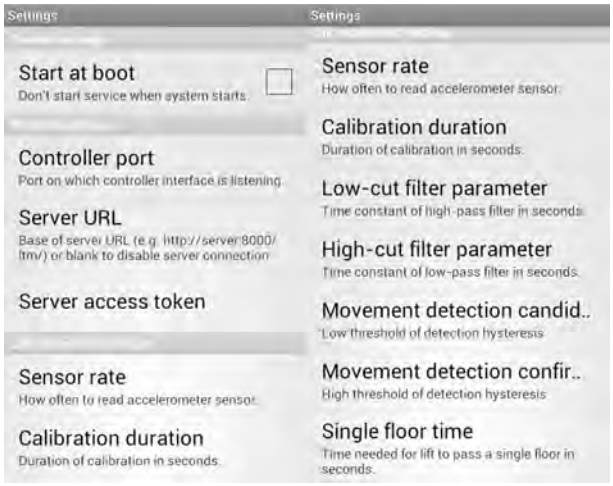


Figure 8. Application preferences

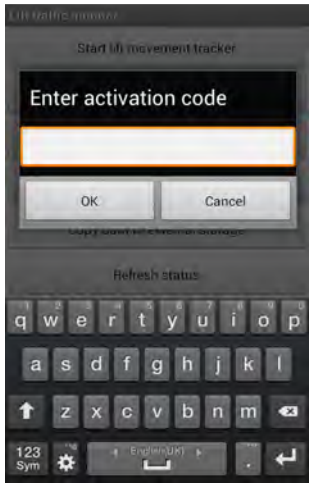


Figure 11. Lift activation dialog in Android application

ports are usually blocked for security reasons, the only universal solution is to use standard ports and protocols (such as HTTP/HTTPS).

User interface (as shown on figure 7) is rather crowded because system is in an early stage of development. Final interface should be minimal and wizard-like to guide user through process of preparing Android smartphone for installation in lift cabin.

The Android application also has settings/preferences in which parameters for tracking lift motion can be adjusted. This is shown on figure 8.

Application on smartphone in lift cabin is a pure Java Android application. Lift movement detecting and tracking logic is put in a separate library, which is used by the Android application.

For lift traffic monitoring, people counter is also necessary. Proof of concept of people counter was also developed for a smartphone mounted on lift cabin ceiling with a top-down looking camera. The frames acquired from the camera in real time are processed using the OpenCV library functions. Image processing pipeline consists of background subtraction, followed by Gaussian blur, followed by contour (blob) detection. Number of detected contours (blobs) presents number of people inside lift cabin. The usage of OpenCV library allowed for a rapid implementation of the image processing pipeline. Figures 9 and 10 depict top-mounted camera view.



Figure 9. Empty lift cabin as seen from top-mounted camera



Figure 10. Lift cabin with four passengers as seen from above

B. *Server*

Server is located somewhere in the cloud. Server application consists of web portal and API interface for communication with Android application. Web portal allows users to view their lifts and analyze gathered data.

Server is written in PHP using *Laravel* framework. *Twitter Bootstrap* and *jQuery* are used for web frontend.

C. *Communication*

Communication between Android application and server is done by HTTP POST requests. Future work includes usage of push technology such as *server-sent events* or *WebSockets*. Currently, communication protocol that is used for communication between server and Android application consists of four POST HTTP requests. For start, Android application has to be bound with a logical lift on server. This process is called activation of lift.

Logical lift is modeled as a finite state machine (FSM), and can be in one of four states:

- Inactive – means that logical lift is not associated with a physical lift
- Active – means that logical lift is bound to a physical lift
- Pending activation – means that an activation code for this logical lift is generated and the lift is waiting to be bound with a physical lift
- Active with pending reactivation – means that lift is currently bound with a physical lift, but is also waiting to be bound with a new physical lift (this can be used for switching smartphone that is monitoring a specific lift).



Figure 12. Lift deactivation dialog in Android application

Activation of lift is done:

1. User logs onto *Lift Traffic Monitor* web portal
2. User creates a new lift
3. User requests activation of lift and receives an activation code which is a four-digit hexadecimal number (e.g. ab12).
4. User installs LTM application to smartphone that he will attach to lift cabin roof
5. User enters the activation code in application on smartphone
6. Application connects to server and finishes activation process

Figure 11 depicts lift activation dialog.

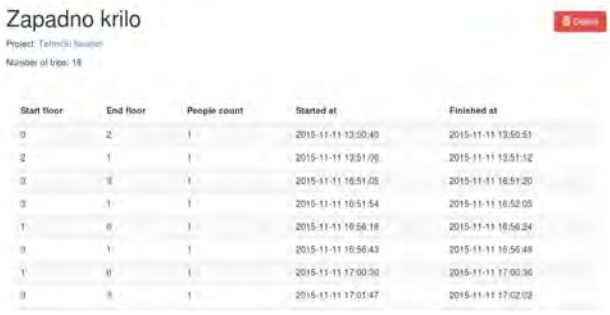


Figure 13. Extract of lift trips collected on location 1



Figure 14. Extract of lift trips collected on location 2

Deactivation of lift is also possible, but only from Android application. Figure 12 depicts lift deactivation confirm dialog in application.

Reactivation of lift is possible in case the smartphone that is monitoring a specific lift needs to be replaced if it gets stolen, broken or otherwise non-functional.

V. CASE STUDY

For the purpose of testing, the system was put in operation for several hours on two locations. Results were satisfactory. On first location detection was 100% correct, and on the second 64%.

The system was installed on two locations:

1. Faculty of engineering in Rijeka, Croatia
2. Elementary school gym in Buje, Croatia

Testing on location (1) was started on Wednesday, November 11th 2015. around 5pm and it run until 7:30pm that day. Device *Samsung Galaxy S2* was hidden in the roof of the lift cabin. This lift can go from 0 to +3. Total of 18 lift trips was collected. There were no error in detection, so detection accuracy for this lift is 100%. Extract of collected data is shown on figure 13.

Testing on location (2) was started on Thursday, November 12th 2015. around 2:30pm. It was running until 7:30pm. Because detection is not 100% accurate, lift tracking had to be restarted several times. Device *Vodafone 858 Smart* was placed inside the lift cabin. This lift can go from -2 to 0. Total of 36 lift trips was collected. Among them, 13 was one-floor trip and 10 was two-floor trip, the others are incorrect due to unanticipated very strong lift vibrations, so detection accuracy for this lift is 64%. Extract of collected data is shown on figure 14.

For testing, server application was installed on AWS (Amazon Web Services) *EC2* node. Its software configuration was: *Amazon Linux* as OS, *MySQL* as *RDBMS* and *nginx* HTTP server with *php-fpm*.

Note that a fully functional and accurate people counter is not yet developed, so people count was set to 1 while testing.

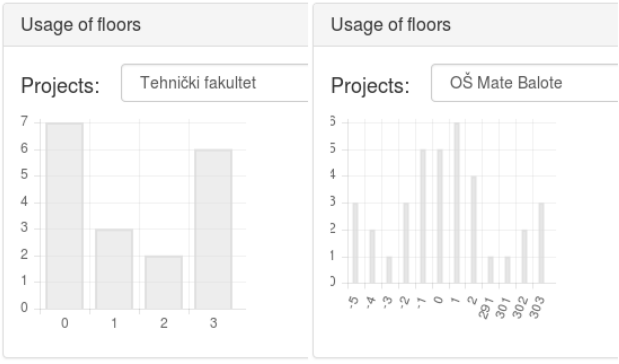


Figure 15. Histograms of lift usage by destination floor

VI. CONCLUSION

In this paper we described a system for simultaneous remote real-time monitoring of multiple lifts using commercial easily available smartphones. We presented the mathematical lift model and described the system's architecture and main components. The operation of the system was validated in two real locations and the results were reported. In addition of tracking the lift movements the system also implements a preliminary people counting solution enabling simultaneous and real time counting of people using the same device.

REFERENCES

[1] J.-M. Kuusinen, J. Sorsa, M.-L. Siikonen and H. Ehtamo , “A study on the arrival process of lift passengers in a multi-storey office building ” Building Services Engineering Research and Technology, 33 (4), pp. 437–449, 2012.

[2] G. Barney, L. Al-Sharif, "Elevator Traffic Handbook: Theory and Practice", 2015

[3] G. Barney, "TRAFFIC DESIGN BENCHMARKS - STANDARD TEMPLATES" Elevation-A View of British Lift Industry 58, 2009

[4] Elevate Perform accelerometer (https://www.peters-research.com/)

[5] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems" Journal of Basic Engineering 82: 35. doi:10.1115/1.3662552., pp. 35–45, 1960.

Mobile devices as authentic and trustworthy sources in multi-agent systems

M.Sc. Vedran Vyroubal [*] , Ph.D. Adam Stančić[*], Ph.D. Ivan Grgurević[**]
[*] University of Applied Sciences Karlovac / Engineering Dept., Karlovac, Croatia
[**] University of Zagreb / Faculty of Transport and Traffic Sciences, Zagreb, Croatia
vedran.vyroubal@vuka.hr, adam.stancic@vuka.hr, igrgurevic@fpz.hr

Abstract – Current state of the art of mobile devices enables user to collect measured data of various physical phenomena. Such devices can be viewed as participating agents in a multi-agent system (MAS). By utilizing public key infrastructure (PKI) these devices can be promoted to an authentic and trustworthy source of data which can be used for forensic purposes. Broader picture of circumstances of an event of interest can be achieved by concurrently collecting still images or video files, and sensor measurements from multiple users, who's mobile devices participate as agents in a MAS. Participation of a user is voluntary, conditioned on the close proximity of the user to the event of interest ("*geo-fencing*"). The collected data is processed by cryptographic and steganographic methods, for the purpose of guarantee that the collected data if authentic and trustworthy. By digitally signing, the collected data has legal power and can be used for forensic purposes and in legal proceedings.

I. INTRODUCTION

Moore's law and integration of sensors combined with the development of modern mobile platforms, enables the use of a mobile device for collection and transmission of data regarding various physical phenomena [1]. Data networks with high bandwidth and signal coverage provide the necessary infrastructure for collecting data regarding some incident, in both urban and rural areas [2]. Modern mobile devices are capable of capturing still images and videos in high quality. Combined with measurements from device sensors this data can be delivered to a monitoring system. Capabilities of mobile devices and integrated sensors are constantly being upgraded with each new model to arrive on market [3].

This paper shall present a possible procedure for utilizing users mobile device as device for collecting authentic and trustworthy data, such as road and traffic conditions and environment conditions. Specifically the presented case is about an incident in traffic, where participating user captures still images or video and collects sensor data. After the data has been collected it is digitally signed and sent to the monitoring system. User is thus unable to affect and modify the collected data. Access to the collected data is completely in the monitoring system's domain. Digital signature provides integrity and trustworthiness, thereby giving legal power to the collected data. Such data can later be used in forensic analysis and in legal proceedings. By utilizing the multi-agent paradigm, multiple users can concurrently collect data for a given incident, from various vantage points. Data from various vantage points provides a broader picture regarding the incident in question.

II. CURRENT RESEARCH

Utilizing the mobile device sensors for data collection is not a novel research topic. The device manufacturers primarily assert the possibility of utilizing device sensors for monitoring users health [4, 5]. Current research publications on the topic single out subjects regarding criminal forensics [6], monitoring of human psychosomatic activities [7], effect of the environment of human health [8] and subjects regarding security of data and privacy of users [9]. Published papers take into account issues concerning possible ethical and legal questions, which concern the area of privacy issues and civil rights. Possible privacy issues with the proposed system are the fact that the collected data can include information about a user who happened to be in the vicinity. In some case even the user is not aware that his device is collecting data about his location and communications [6, 9]. Issues regarding the privacy concerns are not in the focus of this paper, but the need for analysis and resolving of such concerns shall be stressed through the discussion.

Research regarding utilizing cameras and sensors of mobile devices for data collection purposes, are focused on vehicle tracking, driver behavior monitoring, control of incident situations and for notification of drivers about traffic conditions (example; danger ahead, speed limit, etc.) [10]. There are publications which give drivers instructions on how to use their mobile devices as a medium for data collection, but also the rules and regulations to which user must abide (example; it is not permissible to covertly record incident situations) [11].

Problems regarding providing the authenticity, integrity and trustworthiness of collected data, are rather poorly represented in current research publications. Mentioned problems are the focus of this paper.

This paper shall present a collaboration procedure of multiple users, in process of collecting, transforming and storing collected data, regarding traffic conditions.

III. AIM OF RESEARCH

The aim of the research is the presentation of the feasibility of the system for collection, processing and transmission of data regarding incidents, via mobile devices. By combining data from multiple sources, and users who are participants or witnesses of the situation, it is possible to form a detailed description of the origin, development and epilogue of the observed incident. The authenticity of the collected data will be secured by using public key infrastructure mechanisms and steganographic data integration, with a purpose to ensure the legal grounds and legal power in possible judicial proceedings.

IV. TECHNICAL CHARACTERISTICS OF MOBILE DEVICES

Manufacturers compete for market share with ever more frequent launches of new models of mobile devices, which by every new generation become more computationally powerful and with an increasing number and higher quality of integrated sensors [3], as it is shown in Table 1.

Table 2 shows a comparison of the technical characteristics of mobile devices in 2013 and 2015 based on a review of the technical specifications of major manufacturers [12]. The obvious is the significant increase in processor and memory capacity, image resolution and video format and the presence of fingerprint sensors and atmospheric pressure. It also manifests negative trends: low growth of battery capacity, batteries can no longer be manually replaced, lack of additional memory cards, etc. The use of processors with multiple cores, large screen, large number of sensors and recording in high resolution requires a larger amount of energy that must be provided by the irreplaceable battery.

All this has a direct impact on the use of mobile phone users as a means of data collection during the incident.

The problem may occur during the data collection and recording of image or video if the remaining battery capacity or the free space is insufficient. The above problem can be partially solved by pre-allocation of a certain amount of internal memory (or memory card if present) by the application, and by informing the user about the available battery level. As a measure of saving memory space a reduction of resolution or quality of the video can be used. The minimum resolution of the pictures and quality of the video can define the entity that will receive and analyze the collected data. It should be taken into account the possibility that the user will most likely try to capture a larger number of pictures or videos depicting the incident situation.

The collection and storage of data by the sensor does not require an intervention by the user, as does taking a picture with the camera. The amount of information collected by the integrated sensor can also be significant, and as such can have a direct impact on consumption of memory and battery of mobile devices. Table 2 shows a list of sensors that can be found on the mobile device [3].

From the above mentioned it is clear that mobile devices have powerful hardware capable of collecting and processing large amount of data. Software support has evolved over the years and in today's market there are three mayor mobile platforms: Apple iOS, Google Android and Microsoft Windows Mobile 10 (in previous versions of Windows Phone). Each platform provides development tools and more crucially a market for distributing applications to end-users over the Internet. The largest market share of mobile platform (data for the second quarter of 2015) belongs to Android 82.8%, followed by iOS with 13.9% and Windows Phone with 2.6% of the market [14]. Android devices count over 24 000 different models on the market [3], mainly thanks to openness of the development platform. At this point rest of the mobile platforms are very restrictive and closed. Each of these platforms can be used to develop the suggested data collection system.

Table 1: Mobile device features

	CPU	***GPU***	***Memory***	***SD card slot***
2013.	2–4 cores 1,3–1,7 GHz	2-4 cores 1,3–1,7 GHz	16 – 64 GB	Small number of devices
2015.	4–8 cores 1,5–2,0 GHz	Integr. in CPU 192 ALU[1]	16 – 128 GB	Very small number of devices
	Camera resolution	***FPS @ Full HD[2]***	***IP protection***	***Battery capacity***
2013.	4 – 13 MP[3]	24 – 30 FPS[4]	Single manufacturer	1500-2500 mAh
2015.	12 – 21 MP	30 – 60 FPS	Small number of manufacturers	1700-3000 mAh

Integrated sensors	
2013.	accelerometer, gyroscope, proximity detection, magnetometer
2015.	accelerometer, gyroscope, proximity detection, magnetometer atmospheric pressure

Access
WLAN / Bluetooth / GPS / GSM/ HSPA / LTE

1) ALU: (*Arithmetic/Logic Unit*) – number of arithmetic-logic units
2) Full HD: video resolution @ 1920 x 1080 pixels
3) MP: Million pixels in a still image
4) FPS: Frames Per Second

Table 2: Mobile device sensors

Sensors present in modern mobile devices		
Accelerometer	Atmospheric pressure	Bluetooth
Camera (CCD[1])	Telecomm. network	Rotation
Orientation	GPS[2]	Gravity
Gyroscope	Air humidity	IR[3]
Brightness	Magnetometer	Microphone
NFC[4]	Step counter	Proximity
Motion	Thermometer	WiFi[5]

1) CCD (*Charge-Coupled Device*)
2) GPS (*Global Positioning System*)
3) IR (*Infra-Red*)
4) NFC (*Near Field Communication*)
5) WiFi (IEEE 802.11 standard)

V. COLLECTING DATA ON INCIDENT

Application of user mobile device collects two kinds of information regarding the incident: analytical sensor readings and a snapshot of the location taken by the mobile phone's camera. The collected data provide further insight into the conditions at the site of an incident, and into the transport environment. The collected data can be subsequently compared with data from other data sources (e.g. roadway control system) as a method of verification.

The collected data can be stored in a structured text file to facilitate export into a database for further analysis and comparison with other relevant data. For simplicity's sake proposed structure and file format for storage is JSON ("*JavaScript Object Notation"*). The advantages of the above mentioned format is that it is independent of programming languages and platforms, and that its simple syntax enables rapid data processing [15]. The file contains data that clearly defines the sources of data, the time of collection, location and value of the sensor readings. Before sending the collected data to the control center, a process of digital signing of files is performed by leveraging public key infrastructure.

The collected videos or photos of an incident are stored on the mobile device in a pre-defined directory. Also data regarding location, time of the recording and on the parameters of the camera, is automatically stored in the structure of the picture file itself (EXIF, *Exchangeable Image File Format*). The application performs the digital signing of the records (images or videos). Even though the sensor data and image recordings are temporally correlated a steganographic mechanism can be used to further establish the link between the diverse collected data files. A hash value (SHA-1, or similar) of the JSON file, containing sensor readings, is integrated within each of the images, by use of the steganography mechanisms. For added security, it is possible to password protect the process of steganographic data integration.

In addition to the aforementioned analytical and visual data regarding the incident, it is possible to use additional control data, e.g. adding the identification information of the user, device, sessions and the like, for faster and easier control, search and processing of collected data. The company that manages the roadway can also publicly transmit data on the current state of the roads. Optionally collected data on traffic signs and signals, state of road and environment, visibility, current load and all other relevant traffic data, can further improve the process of analysis of the incident.

Data protection is performed on different levels and in every stage of processing: integration of control data, use of a password on methods of data compression and with steganographic integration of hash value of analytical data, and also by using cryptographic mechanisms and digital signature. The proposed system increases the ability to detect manipulation of the content of files in order to provide a basic condition for successful analysis of an incident situation - authentic and trustworthy collected data.

VI. PROPOSED SYSTEM FOR DATA COLLECTION ON INCIDENT

Concerning the heterogeneity of mobile devices with regard to their abilities and available mobile platforms, it presents a challenge to consider using such devices in a distributed data submission system. Even if we limit ourselves on only one mobile platform, the problem persist because of different capabilities of different models of mobile devices, especially if the devices are of different generation. The problem is particularly clear on the Android platform while slightly less on iOS platform, due to smaller number of device models available on the market. This research was limited on Android platform, due to its market share and platform openness.

System for submission of collected incident data, which we propose here, is both distributed and asynchronous. The system consists of a central node for the receiving data collected by the user and of a variable number of users mobile devices, that are logged into the system. Due to the nature of distributed systems, parameters that must be taken into account are delay, throughput and robustness of the data network.

The submission system presumes that the user must voluntarily install the appropriate software on their own mobile device, thereby confirming their intention to participate in such a system.

While considering what parameters must be taken into account for the submission system, in addition to the parameters stemming from the characteristics of the data network, it is also necessary to take into account inherent limitations of mobile devices. Also additional parameters can be custom data submission rules that may be prescribed by the central data receiving node. The rules can define what should be the quality of still images or video files, what sensors should be read and with what frequency or, for example, whether the users data is relevant to situation at hand. Factors that primarily influence the relevance of the collected data are physical proximity of the user to the incident ("*geo-fencing*") and a time-stamp. These rules can be dynamic in their nature and can be subject to change within a relatively short time intervals.

Because of all the aforementioned limitations and design parameters a suggested model for the submission system is based on multi-agent paradigm. Multi-agent systems (MAS) have in practice proved themselves to be a good paradigm for the development of complex systems whose components require a high degree of autonomy and flexibility [16]. One of the most important areas of application of multi-agent systems is in data management [17, 18].

VII. SYSTEM ARCHITECTURE

The proposed system is based on JADE-LEAP agent platform, which is developed in Java, primarily due to the fact that a version for the Android operating system is available. JADE-LEAP platform allows creation of any number of agent containers within which agents are executed. Architecture of JADE-LEAP platform is implemented in accordance with *FIPA Abstract Architecture standard* [19]. Running agents communicate

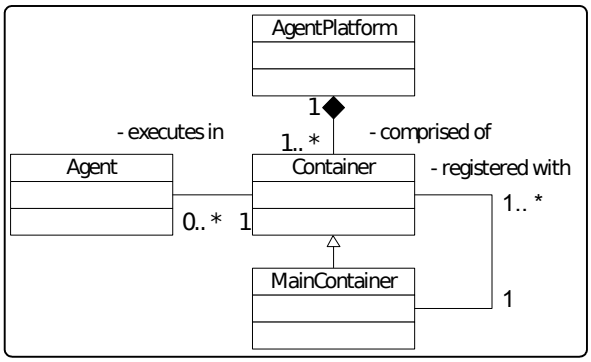


Figure 1: JADE-LEAP UML class diagram

with the other agents, within the same or different containers, by exchanging FIPA ACL messages [20]. Key components of the agent platform are messaging system (*Agent Communication Channel*, ACC), directory service (*Directory facilitator*, DF), and system management agents (*Agent Management System*, AMS). AMS and DF are also agents that run on agent platform and communicate with other agents by exchange FIPA ACL messages.

Each instance of the agent platform consists of at least one container, which is also called the "*main container*", whose main tasks are execution of AMS and DF agents, maintaining a list of other possible containers which are components of a single instance of an agent platform and maintaining a global list of agents that are executed within the current running instance the agent platform (Figure 1).

All containers of an agent platform can be run on the same physical computer or can be distributed across multiple nodes in the network, which by definition makes the agent platform a distributed system. JADE-LEAP provides two modes of operation for the container: a "*stand-alone container"* and "*split container"*. In first mode the container is running on the same physical device that is running JADE-LEAP execution environment. In a second mode an agent container is divided into a "*front*" and "*rear*" part. "*Rear*" part of the container can be executed on another device in the network.

Split mode containers are recommended for use on devices with very limited resources and volatile network connectivity. Split container requires substantially less memory than the full container, and therefore runs faster. Also significant feature is the fact that the communication between the front and rear parts of the container is optimized, which results in lower bandwidth requirements. The rear of the container takes care of communication with the rest of the agent platform. In a case of interruption in network link between the front and back part of the container, the back part of the container receives messages intended for the agent that runs in the front part of the container and stores them in a message queue for the respective agent. After the re-establishment of the link between the front and rear parts of the container all messages are forwarded to the agent. These features of the JADE-LEAP platform allows for the robustness of the proposed solution.

The proposed solution consists of one instance of the agent platform, and a single agent container for each physical user of the submission system (Figure 2). Within each of the users containers one agent is created and started. This agent represents a physical person. Inside the main container one or more *Inbox* agents is created and started, whose primary task is to store received collected data in the database. The proposed system

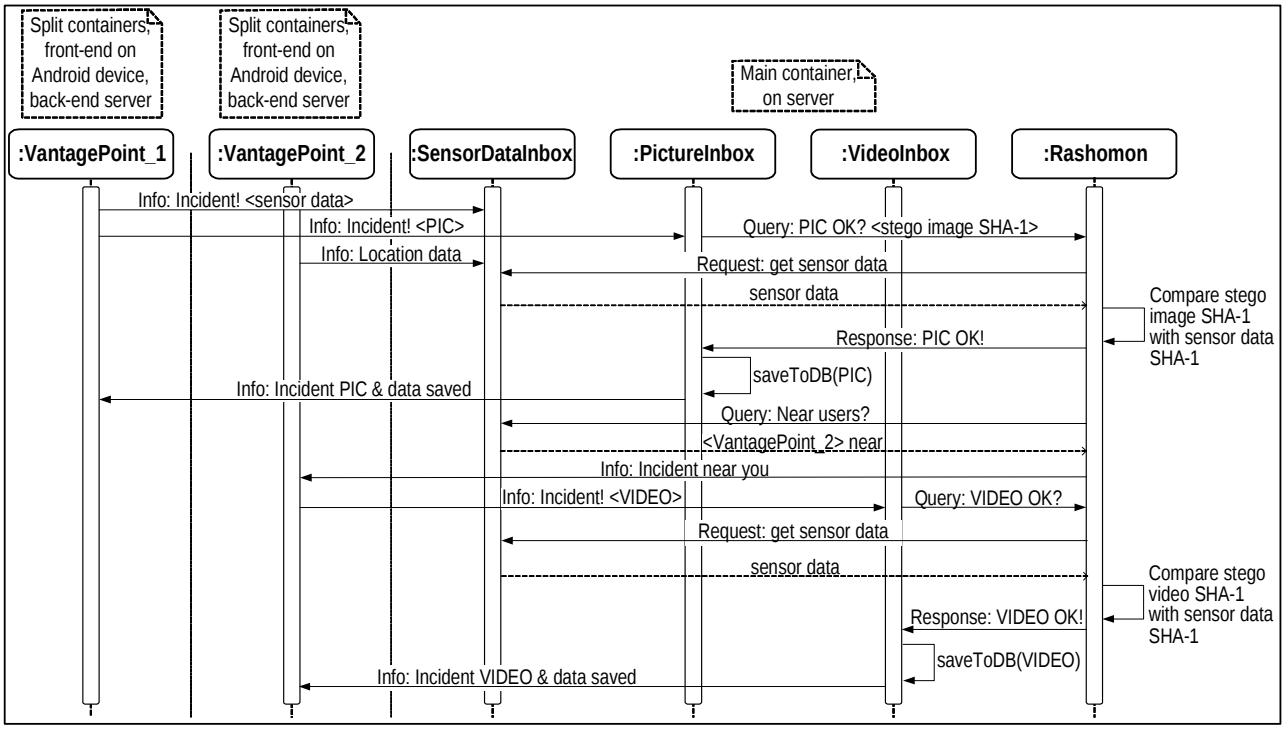


Figure 2: Submission system UML sequence diagram

consists of *PictureInbox*, *VideoInbox*, *SensorDataInbox* and *Rashomon* agents, and for each physical user one *VantagePoint* agent. All agents publish their capabilities to DF agent. The system can be expanded with additional functionalities by adding agents who specialize in a particular purpose. *VantagePoint* agent is executed within split container of any physical user. Its task is to collect data from mobile device sensors and camera and depending on the data context, send the data encapsulated in a FIPA ACL message to one of the above mentioned *PictureInbox*, *VideoInbox* or *SensorDataInbox* agents. These agents then query the *Rashmon* agent about relevance of the received data to the current incident. The decision whether the received data is relevant is based on several parameters; validity of digital signature, geo-location of the user, or if hash value of received sensor data matches the hash value steganographically embedded in the received image or video file.

If the received collected data is deemed relevant and trustworthy it is stored into the database by the *Rashomon* agent, for the purpose of forensic analysis later.

A typical scenario for an incident situation is a traffic accident. In such case external forces are acting on a user and his mobile device. *VantagePoint* agent polls the accelerometer of the mobile device. In accordance with pre-defined rules, the agent then concludes that it is necessary to initiate the process of data collection. After the corresponding *Inbox* agents receive sensor data, images and video via FIPA ACL messages, the *Rashomon* agent may deduce that other users are in the proximity of the incident. It is then possible to send a FIPA ACL message to agent running on those users mobile device, with a request for participation in data collection.

VIII. Privacy protection

The presented work is primarily focused on the process of collecting, processing and transfer of data collected on the incident to the submission system. Since the present method of data collection can involve other users, it is necessary to give an overview on the subject of privacy. According to the provisions of the Croatian Criminal Code in part which refers to criminal offenses against privacy, presented is an Article 144 which deals with procedures of unauthorized image recording [21]. The text explicitly states that the image recording in an apartment or in an enclosed space protected from view is criminal, while image recording in public areas is not mentioned explicitly or implicitly. It is inevitable that the user of the submission system, as well as potential participants in an incident shall have to provide certain personal information to the competent authorities. The question is whether the privacy of all participants and witnesses of an accident is violated, especially in the case of recorded image or video.

An image of the vehicle, or a person, who accidentally passed by the incident may compromise person's safety or be a cause of inconvenience in the area of private and professional life. It should be noted that the video surveillance systems are an integral part of modern transport infrastructure and that each user and vehicle on the road can be monitored and under surveillance in accordance with the purpose of such system, which further diminishes privacy rights.

IX. Conclusion

The proposed method is to be regarded as an aid in the process of analyzing the traffic incidents and road conditions. The degree of technological development does not prevent the implementation of the described solution in any way. Proposed system has various methods of confirming that the collected data is both authentic and trustworthy. If the data is manipulated in any way it can be detected. Utilizing the public key infrastructure the submission system becomes a reliable source of legally verifiable information with legal power. The larger amount of collected data relevant to the incident can aid forensic examinations and analysis, which in turn can be faster, more detailed and precise. It should be emphasized that is necessary to ensure privacy of all participants in the incident, during the collection, processing, analysis and especially in the presentation phase of the data collection process. In the development of the submission system a multi-agent paradigm was used.

The advantages of the presented system are ease of use, portability, scalability, utilization of the existing infrastructure for data transmission and the minimal additional investment in development and user training. Disadvantages are the heterogeneity of the system in terms of technical characteristics of equipment, varying number of samples collected and their quality.

Limitations of mobile devices and the ability of devices and users to react and trigger the process of data collection can represent a problem. In the event of an incident with potential damage to mobile device, presented data collection method's usefulness can be limited.

Future research should focus its attention on privacy issues of all participants in the incident. It should also be useful to consider the possibility of integration of existing sensors in the vehicles into the data collection system.

References

[1] C. Bonnington, In Less Than Two Years, a Smartphone Could Be Your Only Computer, URL: http://www.wired.com/2015/02/smartphone-only-computer/ (22.01.2016.)

[2] Hrvatski Telekom, 4G najbrži mobilni Internet, URL: https://www.hrvatskitelekom.hr/4g (22.01.2016.)

[3] Mobile Sensors Database, URL: https://opensignal.com/sensors/library (22.01.2016.)

[4] Apple Inc., An innovative new way to use your health and fitness information, URL: http://www.apple.com/ios/health/ (22.01.2016.)

[5] E.J. Topol, The Future of Medicine Is in Your Smartphone, URL: http://www.wsj.com/articles/the-future-of-medicine-is-in-your-smartphone-1420828632 (22.01.2016.)

[6] A. Mylonas, V. Meletiadis, L. Mitrou and D. Gritzalis, „Smartphone Sensor Data as Digital Evidence“, Computers & Security, Cybercrime in the Digital Economy, Volume 38, pp. 51–75, October 2013

[7] G. Chettya, M. White and F. Akthera, „Smart Phone Based Data Mining for Human Activity Recognition“, Procedia Computer Science, Proceedings of the International Conference on

Information and Communication Technologies, Volume 46, 2015, pp. 1181–1187, December 2014
[8] V. Patel, M. Nowostawski, G. Thomson, N. Wilson and H. Medlin, „Developing a free open-source smartphone application for studying tobacco use in the field (observing smoking in vehicles)“, Report for the Asthma and Respiratory Foundation of New Zealand, February 2012
[9] T. Jeske, „Floating car data from smartphones: What Google and Waze know about you and how hackers can control traffic“ Proceedings of the BlackHat Europe, March 2013
[10] L. Hannawald, M. Marschner and H. Liers, „The usage of smartphones for recording accidents and incidents from the critical situation up to the post-crash phase“, Verkehrsunfallforschung an der TU Dresden, Paper Number 13-0293, URL: http://www-nrd.nhtsa.dot.gov/pdf/esv/esv23/23ESV-000293.PDF (22.01.2016)
[11] Anderson, Hemmat & McQuinn LLC: How to Use Your Smartphone as Your Singular Tool for On-Scene Information Gathering, URL: http://www.andersonhemmat.com (22.01.2016)
[12] Technical speciation's: All mobile phone brands, URL: http://www.gsmarena.com/makers.php3 (22.01.2016)
[13] IDC Worldwide Smartphone OS Market Share, URL: http://www.idc.com/prodserv/smartphone-os-market-share.jsp (22.01.2016)
[14] Android Fragmentation Visualized (August 2015), URL: https://opensignal.com/reports/2015/08/android-fragmentation/ (22.01.2016)
[15] RFC 7159: The JavaScript Object Notation (JSON) Data Interchange Format, URL: https://tools.ietf.org/html/rfc7159 (22.01.2016)
[16] Nicholas R Jennings. 2000. On agent-based software engineering. Artificial intelligence 117, 2 (2000), pp. 277–296., 2000
[17] K. Decker and K. Sycara, Intelligent Adaptive Information Agents. Journal of Intelligent Information Systems, 9(3): pp. 239–260, 1997.
[18] N. Jennings, and M. Wooldridge, Applications of Intelligent Agents. In Agent Technology: Foundations, Applications, and Markets, pp. 3–28, Secaucus, NJ, Springer-Verlag, Berlin, 1998.
[19] FIPA:Abstract Architecture Specification, doc. SC00001, URL: http://www.fipa.org/specs/fipa00001/index.html (22.01.2016)
[20] FIPA: ACL Message Structure Specification, doc. SC00061, URL: http://www.fipa.org/specs/fipa00061/index.html (22.01.2016)
[21] Kazneni zakon RH: Kaznena djela protiv privatnosti,(NN 125/2011) URL: http://narodne-novine.nn.hr (22.01.2016)

SOFTWARE ENGINEERING

Comparative analysis of functional and object-oriented programming

Dino Alic*, Samir Omanovic* and Vaidas Giedrimas**
* University of Sarajevo/Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina
** Siauliai University/Institute of Informatics, Siauliai, Lithuania
alic.dino@gmail.com, samir.omanovic@etf.unsa.ba, vaigie@mi.su.lt

Abstract - The choice of the first programming language and the corresponding programming paradigm is an important part of the software development process. Knowing the advantages and constraints of individual programming paradigms is important as it can be crucial for successful software implementation. In this paper we conduct an empirical comparison of functional and object-oriented programming languages using analog examples in C#, F#, Haskell, and Java. Three algorithms were implemented: algorithm for solving N queens problem, algorithm for generating n-th left-truncatable prime and merge sort algorithm in C#, F#, Haskell and Java programming languages. An overview of programming languages efficiency is given by measuring two basic parameters: number of lines of code and program execution speed. Also, system resource usage is monitored during execution. Limited experiments showed that the programming language Java is faster than the other three languages whose performances were measured. Java was surprisingly fast on these problems that are more suitable for functional programming languages. Haskell was less memory intensive (up to two times less than Java) with similar execution times, while .NET languages were slower up to four times in comparison to Java. Object-oriented languages C# and Java had significantly more lines of code for all three algorithms when compared to functional programming language Haskell and the hybrid one F#.

I. INTRODUCTION

As with the many things in programming languages history, there is an always present debate about which programming paradigm is better and when a specific paradigm should be used [1] [2]. Even the classification of paradigms itself is often cause for a debate. However, programming paradigms are often classified in six fundamental paradigms: imperative, declarative, object-oriented, functional, symbolic, and logic programming paradigms. Declarative and symbolic paradigms are sometimes left out leaving only four fundamental paradigms [3].

It is important to note that most of the mainstream programming languages in wide use today support multiple paradigms (multi-paradigm programming languages [4]). For example, C++ has native support for seven programming paradigms with additional libraries extending the support for six additional programming paradigms.

In this paper we focus on two fundamental programming paradigms: object-oriented and functional paradigm. Object-oriented paradigm is in wide spread use in industry [5] today with functional paradigm used mainly by the scientific community.

However, there is a notable trend of object-oriented programming languages introducing some basic principles of functional programming in new versions. In 2008 lambda expressions were introduced in .NET framework version 3.5 [6], primarily to support LINQ (Language Integrated Query). Lambda expressions were also introduced in Java 8 [7], allowing the developers to express instances of single-method classes more compactly. Introduction of functional programming principles in object-oriented languages shows that functional and object-oriented programming languages are not mutually exclusive and that combining the two principles can be effective [1].

II. FUNCTIONAL VS OBJECT-ORIENTED PROGRAMMING

Despite the always present debate over choosing the programming paradigm, there is reasoning behind choosing the programming paradigm which will be used to solve the problem. Object-oriented languages are good when there is a fixed set of operations on things, and as the code evolves, primarily new things are added [8], like adding new classes which implement existing methods (existing classes are left alone). Functional languages are good when there is a fixed set of things, and as the code evolves, primarily new operations are added upon existing things [8], like adding new functions that work with existing data types (existing functions are left alone). This is known as an expression problem [9], a term coined by Philip Wadler.

It is important to note that programmer can do functional programming in object-oriented languages. To accomplish this all program object (data) must have constant state and all methods and operations must depend only on its arguments and the context available at definition time.

A. Object-Oriented Programming Paradigm

Object-oriented paradigm is based on the concept of *objects* – data structures containing data known as attributes and corresponding code known as *methods*. The basic idea of object-oriented languages is combining *data* and the *functions and methods that operate on that data* into a single component.

Object-oriented programming languages implement three basic principles of object-oriented approach [5]:

- *encapsulation* – limiting the access to the object from other pieces of code;
- *inheritance* – using the existing object or class as a base for new ones;
- *polymorphism* – provision of a single interface to entities of different types (object can take on several forms while the program is running).

Object-oriented programming is not primarily concerned with the details of program operation. Instead, it deals with the overall organization of the program.

B. Functional Programming Paradigm

Functional programming is at its core based on the mathematical apparatus. The base for all functional programming languages is *lambda calculus*. Lambda calculus has three basic rules [10]:

$$\begin{aligned} &(\alpha \text{ conversion}) && x.A \xrightarrow[\alpha]{} y.[y/x]A \\ &(\beta \text{ reduction}) && (x.A)B \xrightarrow[\beta]{} [B/x]A \\ &(\eta \text{ conversion}) && x.Ax \xrightarrow[\eta]{} A \quad \text{if } x \text{ not free in } A \end{aligned} \tag{1}$$

It can be argued that functional programming languages only implement the lambda calculus while adding constants and new data types. Lambda calculus treats functions as expressions which are gradually transformed until final solution is found – function actually defines the algorithm. Lambda calculus is in wide use in mathematics, philosophy, linguistics, and computer science.

In contrast to imperative programming, order of execution in functional programming languages is not defined by the programmer. Programmer only tells the program what needs to be done without explicitly defining how it should be done. Compiler/interpreter will determine order of execution. Functions in functional programming languages are evaluated only when the result they return is needed. This approach is known as *lazy evaluation* [11].

C. Differences Between Functional and Object-Oriented Programming

One of the fundamental differences between the two paradigms is the interrelationship between data and operations. Basic rule of object-oriented programming is that data and operations affecting them are tightly linked. Objects owns their data and control all actions (functions and methods) modifying them while hiding them from other objects using the interface. This implies that the abstraction model in object-oriented programming is the *data* itself. In functional programming data is tightly related with functions. It is possible do implement different operations over the same data structure. However, basic abstraction is a *function*, not the data structure. Functions hide their own implementation.

Recursion is used in both programming paradigms, but object-oriented programmers tend to avoid it. While recursion reduces the code length it can have a negative effect on code readability. In addition, iterative approach is more performance wise in comparison to recursive approach. In object-oriented programming there is always an iterative equivalent to any recursive algorithm. On the other hand, recursion is essential part of the functional programming since it is based on the lambda calculus where "everything is a function". The essential build blocks of object-oriented programs are *statements* while functional programs are composed mostly from *expressions*.

Another fundamental difference between the two programming paradigms is the *program state*. The main limitation in functional paradigm and purely functional programming languages is the absence of program state. Purely functional programming languages are strictly *stateless* and work with *immutable* data. Functional program can be viewed as independent on the machine meaning that there is no state as we do not have the notion of memory. Conversely, object-oriented programs work with *mutable* data, storing the data in variables and defining statements which can change program state. As a result of stateless approach in functional programming, program execution has no side effects. This implies that the order of statement execution is not important. The program will, for identical inputs, return identical outputs regardless of the order of statement execution. This is known as *referential transparency*, with *referential opacity* being the opposite. Referential opacity is a side effect of working with mutable data. Eliminating side effects can make it simpler to help understand as well as predict behavior of the program [12].

III. Programming Languages in Comparative Analysis

Four programming languages were chosen for this comparative analysis: Haskell, C#, F# and Java. Haskell is a purely functional programming language [13], C# and Java are object-oriented languages with built in support for additional programming paradigms [14] [15], while F# is a hybrid of functional and object-oriented programming [16]. Here are presented some of the most important characteristics of four programming languages used in the analysis.

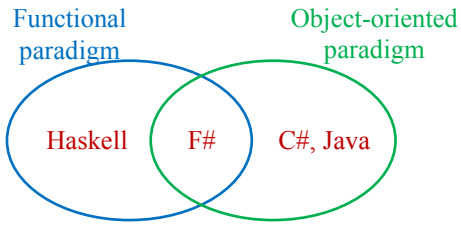


Figure 1. Principle programming paradigms in chosen programming languages

All chosen programming languages are *statically typed* which means that all data types are known at the compile time which leads to *type-safe code*. However, while Haskell and F# are statically type-checked, compiler automatically infers a precise type for all values. In Java,

manners of evading the static type system are controlled by the Java virtual machine's type system. C# is similar to Java in this respect but it allows disabling dynamic type checking by explicitly placing code segments in *an unsafe context*. F# and C# are based on .NET platform. All .NET languages are compiled to Common Intermediate Language (CIL) meaning that, in theory, programs implementing the same functionality written in F# and C# will be translated into the same CIL code.

TABLE I. OVERVIEW OF PROGRAMMING PARADIGMS SUPPORT IN PROGRAMMING LANGUAGES IN ANALYSIS

Programming language	Supported paradigms
C#	structural, imperative, object-oriented, event driven, functional, generic, reflective, concurrent
F#	functional, imperative, object-oriented, metaprogramming, concurrent
Java	structural, imperative, object-oriented, functional, generic, reflective, concurrent
Haskell	functional, lazy/non-strict, modular

All four programming languages are standardized, general purpose programming languages. Haskell and Java are cross – platform languages while C# and F# are designed for Windows platform only, sharing the CLI platform.

A. Haskell

Haskell is a purely functional programming language with no attributes of imperative programming languages. It is a *higher order* programming language – functions can receive functions as arguments and return functions as a result. By using this feature it is possible to define new function as a composition of two existing functions [17][18].

Conditional expressions can be nested and they must have an *else* branch. There are no loops in Haskell with recursion as a programming equivalent. Functions can be defined using lambda expressions which create a frame for each argument and function body which will calculate result based on the arguments provided but they do not name the function. In other words, lambda expressions are nameless functions in Haskell [19].

Algorithms in Haskell consist of *base cases* and recursive cases. Determining the base cases is crucial for generating the correct output and avoiding the infinite loops [17]. For example, to calculate factorial of an integer provided as an input, the base case is 0 when the result is generated immediately and recursive case is used for all other inputs.

B. F#

F# is a programming language supporting both object-oriented and functional programming paradigms [16]. Object-oriented approach allows the programmer to write code using classes and objects. On the other hand with functional approach programmer can tell the program what needs to be done without explicitly defining how it should be done. F# uses *greedy evaluation* techniques, a distinctive feature of imperative programming languages. When using greedy evaluation, an expression is evaluated as soon as it is linked to a variable.

In object-oriented programming all variables are mutable unless explicitly declared as static. In F# all variables are static unless explicitly declared as mutable. *Null* values are not supported and *NullReference* exceptions are not possible [20]. Another important difference between most object-oriented languages and F# is that F# programs do not force the programmer to declare data types. Instead, type is determined based on the value provided (C# has the keyword *var* providing this feature).

C. Java

Java is a general purpose programming language. Java applications are compiled to *bytecode* which can then run on Java virtual machine. Important advantage of using Java is that the same code can be run on all platforms that support Java without the need for recompiling the code. The existence of the intermediate had significant impact on the performance until the just in time compilation was implemented, partially resolving the performance hit. Just in time compilation is a form of dynamic compilation which allows the code compilation during run-time opposed to compile and run approach. However, this can cause a small delay when starting the application.

Java 8 added lambda expressions – a key concept in functional programming. Lambda expression in Java is a block of code that can be passed around and executed later multiple times (*deferred execution*). Java lambda expressions are especially useful in GUI applications [21]. According to TIOBE index Java is the most popular programming language in use today [22]. That is why comparison with other languages is important to find his strengths.

D. C#

C# is a general purpose object-oriented language. C# supports more data types than Java and implements most of the features that Java possesses, with *checked expressions* being the notable exception. C# 3.0 added support for lambda expressions and expression trees as an incremental step for steadily expanding support for functional programming in C# [23]. A lambda expression in C# is an anonymous function that can be used to create delegates or expression tree types. By using the lambda expressions programmer can write local functions and then pass them as an argument or return them as a result. Expression lambda is a lambda expression with an expression on the right side of the operator and is used to create expression trees. Expression trees are a code in a tree-like data structure, where each node is an expression. Expression trees are also used in the dynamic language runtime (DLR) to provide interoperability between dynamic languages and the .NET Framework.

IV. OWERVIEW OF EXECUTION ENVIRONMENT AND DEFINITION OF PROBLEMS FOR ANALYSIS

For the purpose of comparative analysis three different algorithms where implemented in all four chosen languages. The implemented algorithms are:

- **Algorithm for solving the N queens problem.** Chessboard queens can attack horizontally, vertically, and diagonally. The N-queens problem asks: "*How can N queens be placed on an NxN chessboard so that no two of them attack each other?*". N-queens problem complexity is roughly exponential. Naïve approach to solving the problem consists of generating all possible configurations of queens on board and finding those that satisfy the requirement. Backtracking algorithm significantly reduces the number of variations checked by eliminating variations without explicitly checking them as they will not be part of the solution. Backtracking approach in solving the problem was used in order to show limitations of functional programming as this approach requires mutable data to work with. However, purely functional programming languages only work with immutable data. Algorithm was used to solve the N queens problem for N = 12, 13 and 14.
- **Algorithm for generating n-th left truncatable prime.** Left-truncatable prime is prime numbers which, in a given base, contains no zeroes, and if the leading left digit is successively removed, then all resulting numbers are prime numbers. There are only 4260 known n-th left trucatable primes in base 10 with the largest currently being the 24-digit number 357 686 312 646 216 567 629 13 [24]. Algorithm performance was measured when generating 1300th, 1700th and 1800th left trucatable prime.
- **Merge sort algorithm.** Merge sort is a divide and conquer algorithm for sorting. It works by dividing the unsorted lists into sub-lists until each sub-list contains one element only. Then the merge phase commences. Sub-lists are repeatedly merged producing new sub-lists until there is only one sub-list remaining. This is the final sorted list. Text files containing 10 000, 30 000, 50 000 and 100 000 random numbers were generated. Program flow consists of importing the text file and then sorting the numbers found in it by using the merge sort algorithm.

Object-oriented approach was used for implementing algorithms in C# and Java, functional approach for implementing algorithms in Haskell and a combination of functional and object-oriented approaches when implementing algorithms in F#. Algorithms were not implemented to support any kind of parallelism in code. It is important to note that Glasgow Haskell Compiler (GHC) was used. GHC is itself written in Haskell based on a technique known as bootstrapping. The runtime system for Haskell (used to run programs) is written in C.

A. *N Queens Alghorithm with Backtracking*

Haskell program had the least number of lines of code with C# having the most. Java program has only two lines less than the C# version. C# had twice as much lines of code comparing to F# but F# had more almost three times as much lines of code compared to Haskell.

TABLE II. OVERVIEW OF THE EXECUTION ENVIRONMENT

Processor	Intel Core i5-3210M 2.5 GHz
RAM	6 GB DDR3-1600
GPU	Optimus - Intel HD Graphics 4000, 2 GB
GHC version	7.10.3
Java version	8.66
.NET version	4.6.1
Operating system	Windows 7 Ultimate Service pack 1, 64bit

When it comes to program run time Java program was the notably faster than all other programs. C# was four times slower in average when compared to Java. Haskell and F# programs were significantly slower in generating results in comparison to object-oriented Java and C#. This can be attributed to the inability to take full advantage of backtracking approach.

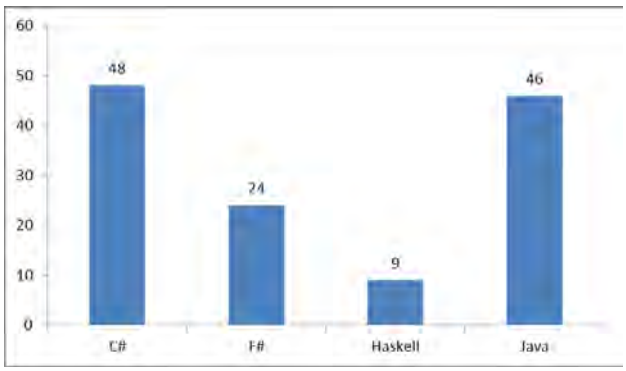


Figure 2. Number of lines of code in algorithm for solving the N queens problem

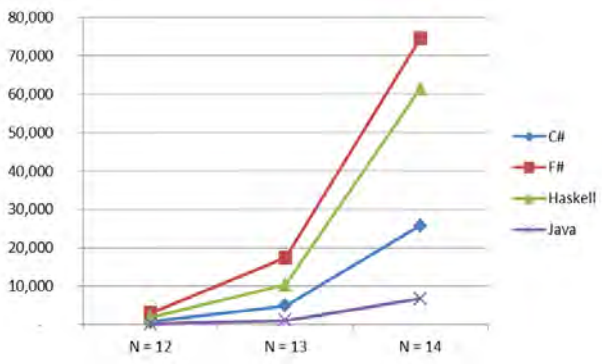


Figure 3. Average run time of N queens algorithm [seconds]

B. *Generating n-th left trucatable prime*

As with the N queens algorithm Haskell program had the least amount of number of lines of code. However, this time F# had only nine lines more than Haskell. Java and C# were almost tied again. Java and C# programs had almost three times as much lines of code when compared to Haskell and more than twice when compared to F#.

In comparison to N queens program, Haskell and C# changed places with Haskell program coming in second place. Once again, Java program was the fastest, C# and F# were behind the Java and Haskell with a notable difference.

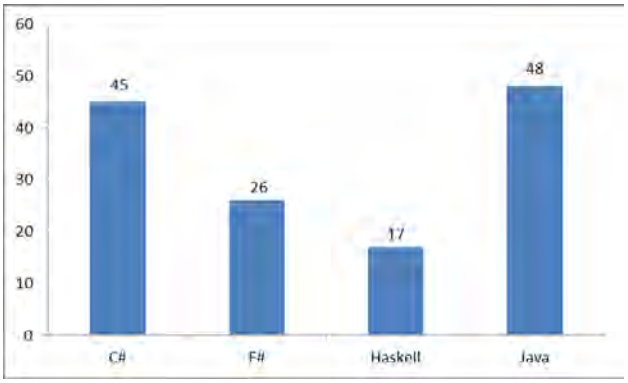


Figure 4. Number of lines of code in algorithm for generating n-th left trucatable prime

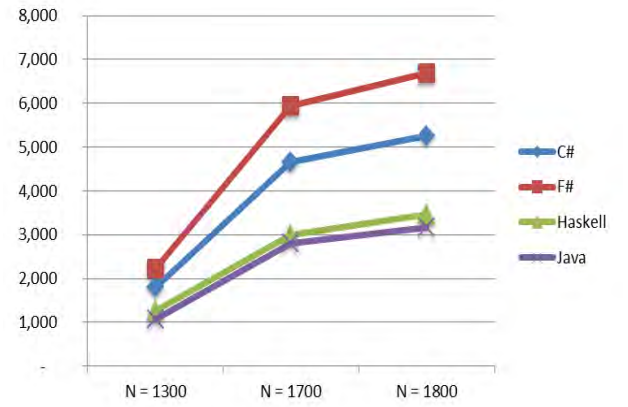


Figure 5. Average run time of algorithm for generating n-th truncatable prime [seconds]

C. *Merge sort*

As it was the case with the previous two algorithms, Haskell program had the least amount of lines of code. F# had somewhat more lines of code in comparison to Haskell but still notably less when compared to C# and Java. In merge sort algorithm a significant difference in number of code lines between Java and C# exists, which was not the case in the previously discussed algorithms.

Java program run time was fastest once again. However, Haskell program was second with an almost negligible difference which was reduced as the N increased. For N = 100 000, run time for Haskell and Java programs were almost identical. C# had the third longest execution time, notably longer than those of Java and Haskell. F# had slightly longer execution time than C#, coming in last.

Merge sort implementation in all programming languages was based on recursive approach, where we would expect that functional programming languages (purely functional in particular) show better performance. However, Java program was slightly faster than Haskel even in this algorithm.

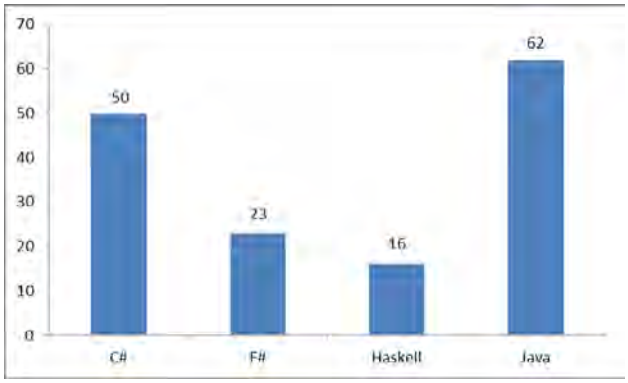


Figure 6. Number of lines of code in merge sort algorithm

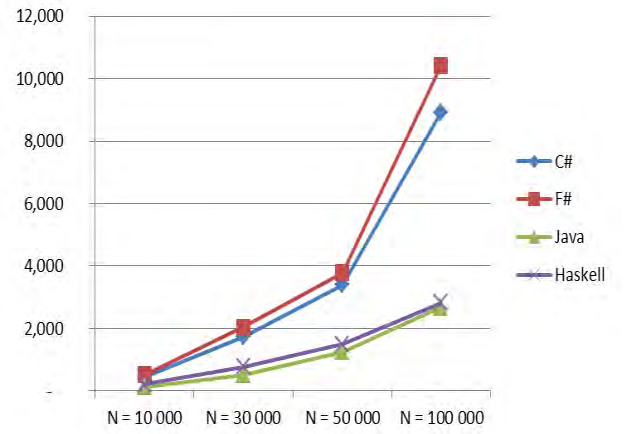


Figure 7. Average run time of merge sort algorithm [seconds]

V. DISCUSSION

Java programs run times were fastest in all three algorithms analyzed. Haskell programs had fewer lines of code in all three analyzed algorithms. Experiments showed that .NET programming languages were significantly slower solving the presented problems when compared to Haskell and Java.

Object-oriented languages were significantly faster when using the backtracking algorithm, being able to utilize all the advantages of backtracking approach.

Processor usage was almost the same for all programs with a negligible difference of +-1%. However, purely functional Haskell programs had a considerably lower memory load, being in average two times less memory intensive than all other programs. This can be explained with stateless approach being used in functional programming (working with immutable data only) which consumes much less memory in comparison to mutable data.

As a rule, functional programming languages tend to have less lines of code. This does not imply that they have better code readability as it will depend on the implementation and preferences of the programmer. Object-oriented languages, C# and Java, had only small

differences in number of lines of code between the two. Haskell had, in average, three times less program code in comparison to object-oriented languages. Hybrid programming language F#, combining both functional and object - oriented paradigms proved to be significantly slower than all others. This goes to show that combining the two programming paradigms might not result in a better solution (in this case we have quite an opposite situation). When combining these two paradigms a programmer should have in mind all the advantages and disadvantages of each programming paradigm and combine them only in an extent necessary to solve the actual problem.

It would be interesting to compare the two paradigms when using the multithreading and concurrent approach. Mutable variables can result in many issues in object-oriented languages. Working with purely functional languages and using the immutable data eliminates most of these issues. This way, programmer does not have to worry about side effects of multithreading and concurrent approach.

VI. CONLUSION

Functional paradigm is preferable choice for implementing algorithms described above. That is partially confirmed by the results obtained in the experiments, having in mind number of lines of code, memory usage and speed. On the other side, Java is surprisingly fast comparing to Haskell. C# implementation showed mainly expected results while F# showed disappointing results.

Experiments confirmed once more that it is hard to make general conclusion which programming paradigm is better, but it is obvious that they should be combined to increase execution efficiency. For the software engineering industry combination of paradigms is the best choice. Besides that, selection of the programming language is also important. Results support conclusion that Java is the best choice among languages used in these experiments, having in mind execution efficiency. The results also explain Java's popularity.

Comparation of programming paradigms and programming languages is always interesting topic if we want to estimate influence of technology changes to the observed concepts. In that sense our future work will be directed toward comparing Java and Haskell, as the best choice from these experiments, with some languages that are not pure functional programming languages but support functional programming (Swift programming language, for example) and having in mind some execution platform limitations (mobile phone, for example).

REFERENCES

[1] B. Meyer, "SOFTWARE ARCHITECTURE: OBJECT-ORIENTED VS FUNCTIONAL," chapter 13 of *Beautiful Architecture, edited by D. Spinellis and G. Cousios*, O'Reilly, 2009, pp. 315-348.

[2] S. Nanz and C. A. Furia, "A Comparative Study of Programming Languages in Rosetta Code," in *Proceedings of the 37th International Conference on Software Engineering - Volume 1*, Piscataway, NJ, USA, 2015, pp. 778–788.

[3] K. Normark, "Functional Programming in Scheme." [Online]. Available: http://people.cs.aau.dk/~normark/prog3-03/html/notes/theme-index.html. [Accessed: 02-Feb-2016].

[4] J. O. Coplien, "Multi-Paradigm Design," Ph.D. dissertation, Informatics Dept., Vrije Univ., Brussel, Belgium, 2000.

[5] D. Donko and S. Omanovic, Objektno orijentirana analiza i dizajn primjenom UML notacije, Elektrotehnicki fakultet, Sarajevo, 2009.

[6] Microsoft, ".NET Framework Versions and Dependencies." [Online]. Available: https://msdn.microsoft.com/en-us/library/bb822049(v=vs.110).aspx. [Accessed: 02-Feb-2016].

[7] Oracle, "Information about Java 8." [Online]. Available: https://java.com/en/download/faq/java8.xml. [Accessed: 02-Feb-2016].

[8] A. Scott, "Functional Programming vs. Object-Oriented Programming." [Online]. Available: http://functionspace.com/topic/1467/Functional-Programming-vs--Object-Oriented-Programming. [Accessed: 02-Feb-2016].

[9] D. Wampler and A. Payne, Programming Scala: Scalability = Functional Programming + Objects. O'Reilly Media, 2014.

[10] D. A. Turner, "Some History of Functional Programming Languages" - *an invited lecture given at TFP12, St Andrews University*, 12 June 2012.

[11] T. Kuhne, A Functional Pattern System for Object-Oriented Design, Verlag Dr. Kovac, Hamburg, Germany, 1999.

[12] Basili, V.R.; Selby, R.W., "Comparing the Effectiveness of Software Testing Strategies," in *Software Engineering, IEEE Transactions on* , vol.SE-13, no.12, pp.1278-1296, Dec. 1987

[13] Haskell, "Haskell Language." [Online]. Available: https://www.haskell.org/. [Accessed: 03-Feb-2016].

[14] Microsoft, "Visual C# Resources." [Online]. Available: https://msdn.microsoft.com/en-us/vstudio/hh341490.aspx. [Accessed: 03-Feb-2016].

[15] Oracle, "Java 8 Central." [Online]. Available: http://www.oracle.com/technetwork/java/javase/overview/java8-2100321.html. [Accessed: 03-Feb-2016].

[16] Microsoft, "Visual F#." [Online]. Available: https://msdn.microsoft.com/en-us/library/dd233154.aspx. [Accessed: 03-Feb-2016].

[17] G. Hutton, Programming in Haskell, Cambridge University Press, 2007.

[18] M. Leucker, T. Noll, P. Stevens, and M. Weber, "M.: Functional programming languages for verification tools: Experiences with ML and Haskell," in In: Proceedings of the Scottish Functional Programming Workshop, 2001.

[19] B. O'Sullivan, J. Goerzen, D.B. Stewart, Real World Haskell, O'Reilly Media, 2008.

[20] T. Liu, F# for C# Developers, Microsoft Press, 2013.

[21] Oracle, "Lambda Expressions (The Java™ Tutorials > Learning the Java Language > Classes and Objects)." [Online]. Available: https://docs.oracle.com/javase/tutorial/java/javaOO/lambdaexpressions.html. [Accessed: 03-Feb-2016].

[22] TIOBE, "TIOBE Software: Tiobe Index." [Online]. Available: http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html [Accessed: 03-Feb-2016].

[23] Microsoft, "Lambda Expressions (C# Programming Guide)." [Online]. Available: https://msdn.microsoft.com/en-us/library/bb397687.aspx. [Accessed: 03-Feb-2016].

[24] O. Angell and H. J. Godwin, "On Truncatable Primes," in *Mathematics of computation*, AMS, vol.31, No 177, pp. 265-267, Jan 1977

Improving the composition and assembly of APIs in service dominant ecosystem environments

Darije Ramljak
IBM Croatia Ltd/IBM Global Business Services
Zagreb, Croatia
darije.ramljak@hr.ibm.com

Abstract - Current research initiatives on API ecosystems mainly concentrate on the technical aspects of the API composition and assembly. API ecosystems usually have challenges such as publication and provisioning of APIs by providers and selection and consumption of relevant APIs by consumers. Difficulties and problems in API value oriented analysis and design are often overlooked. This paper introduces the new method and models for proper alignment of API value creation and value capture with business architecture, business processes and services.

I. INTRODUCTION

API economy is the concept that underpins the creation of new business ecosystems and new business models. An ecosystem is a complex web of interdependent enterprises and relationships aimed to create and allocate business value. Ecosystems are broad by nature, potentially spanning multiple geographies and industries, including public and private institutions and consumers.

Ecosystems are structurally complex environments comprising participants and their interactions. They have unique rules of engagement and share significant market characteristics, such as:

- Role within the environment
- Influence within the environment
- Capabilities and value propositions
- Governing rules
- Connections of elements
- Course of interaction

In the near and imminent future, ecosystems will enable organizations to create and offer tailored, customized experiences to consumers, who can choose from multiple, integrated options and thus reap the greater value from it.

The term value creation is essentially the main driver of ecosystem assembly and evolution. Ecosystem participants/partners can create value through innovating products, services, or experiences and they must collaborate to create and deliver something of mutually beneficial value. In effect, value creation and value capture reflect a networked, dynamic, everyone-to-everyone process of exchange and the main difference to standard value chains is that ecosystems produce more value as a whole, than the sum of the individual participants acting independently.

Creation and evolution of the ecosystems is founded on a number of enablers as follows:

- *Connectivity required for creation and evolution of the ecosystem*
- *Contextual, individualized experiences for consumers*
- *Simplicity to increase user accessibility while expanding sophistication*
- *Market adaptability by enabling experimentation with multiple, new partners*
- *Business scalability with limitless, cost-effective computing power*
- *Cost flexibility as costs shift from fixed to variable and users pay-as-they-go*

In a nutshell, ecosystem enablers are mandating efficient means of product creation, collaboration, data exchange, and interconnection of systems. Thus, APIs become the main technology and conceptual building blocks to create new ecosystem operational platforms and environments.

In this paper we are introducing a concept of *value network* [1] and define it as a value creating system in which all involved stakeholders in the ecosystem "co-create value". Since "value co-creation" is the ultimate objective of services, the method should firstly focus on high-level service value models, i.e. the identification of service participants (ecosystem participants), their value expectations and value exchange relations.

We position the starting point of the proposed method to the high-level service value model, which is described by service value network. Value model is subsequently transformed into a property graph representing the API model and further defined as an RDF model, capable of being queried as well as serving as a foundation for API execution models.

Majority of recently published papers dealing with the problem of API composition and assembly are mostly focusing on the technical issues such as interface matching, semantic service description and semantic service discovery. The proposed approach is aiming on the creation of the value oriented method for the definition of flexible, business aligned services to enhance the cumbersome process of API composition and assembly.

This paper is organized as follows. Section II gives a description of Service value definition and the role of Enterprise architecture in decisioning process. Section III introduces the API graph model and the role of RDF. Section IV presents the process of service selection based on the value capture and value creation criteria. The conclusion and future work are presented in section V.

II. BUSINESS MODEL AND VALUE NETWORK

In the context of this paper, an "enterprise" is defined as a collection of organizations having the same set of goals and/or the same common denominator. In this sense, an enterprise could be a government agency, government apparatus as a whole, a corporation, a corporate sector or just a department [5].

Business model is a widely used concept in business domain and value network analysis. A business model provides a description of the roles and relationships of a company, its customer, partners and suppliers, as well as the flows of goods, information and money between these parties and the main benefits for those involved. Business model usually covers the following 4 perspectives, i.e., what (i.e., value proposition), how (e.g., value composition, partnership, core capability), who (i.e., customer segment, distribution channel, and relationship), and how much (i.e., revenue and cost).

Business architecture is a tactical representation of a business model. It is also crucial for the definition of the enterprise wide services (APIs) since they have to be closely aligned with the actual business function required by the enterprise.

A very important concept introduced in the definition of Business architecture is the notion of Component Business Architecture (CBA). The decomposition of the entire enterprise into well defined and discrete components enables a clear separation of concerns, and the definition of business services offered within the enterprise or to the ecosystem. The concept is based on the Component Business Modeling (CBM) method [8]. CBM as a methodology is more comprehensive in terms of facilitating the transformation of the enterprise, while the new concept of CBA is more focusing on alignment of business services, business processes, and the service oriented information systems. One of the purposes of creating the business components is also the need to break complex business processes into manageable parts that have clearly defined boundaries (component boundaries) and interfaces (component services).

CBA differentiates the ecosystem wide, inter-component and intra-component processes. Inter-component processes are describing how the enterprise-wide processes are being choreographed using the business services provided and/or required by the business components. It also defines the value creation and value capture requirements as a set of constraints, i.e. circumstances in which the enterprise is willing to be a part of the ecosystem. Value capture and value creation flows are proposed to be modeled in a form of a value network.

Intra-component processes, on the other hand, are describing how the business service's logic, within the component that offers it, is defined in a form of processes and sub-processes (usually depicting the API that needs to be exposed from Systems of Record).

API definition is also managing the information aspect in a form of identification of business entities that are being exchanged between services and business activities in the business processes and exposed to the ecosystem.

A value network model is proposed to illustrate how the ecosystem participants work within an extended business environment. The traditional value network is depicted as a directed graph, in which nodes are service participants and directed edges are value flows between participants. It is used to characterize the creation and transference of various tangible and intangible values between participants [14]. Based on the value network analysis, ecosystem participants can make business decision whether the participation in the ecosystem is feasible and profitable or not, or the value network has to be optimized and re-constructed.

III. THE MODEL

The model introduced in this paper is focusing on the definition of relevant elements of API model that needs to translate the value network (ecosystem participants and value flows between them) into API graph suitable to be analyzed and executed. This is an actual translation of enterprise business strategy, tactical and execution model (component business architecture and associated business drivers and value network flows) into a technical API model, enabling the realization (capturing) of the value by executing API calls towards other ecosystem participants.

A. API Model structure

API model is structured as a property graph, meaning that both its nodes and edges (relationships between nodes) have properties. Properties provide characteristics of nodes and edges, and are defined as key-value pairs. The model is created in a formal, yet simple structure with extensibility in mind, by allowing new properties to be easily added to each node and edge. Also, Party-Role concept allows modeling of virtually any type of ecosystem participant, whether being provider, consumer or aggregator/hub for APIs.

The following elements are used as nodes in the API model property graph (*Figure 1 - API model structure*)**:**

- Party – representing a subject (actor) interacting in the ecosystem (humans, organizations)
- Role – definition of role that specific Party plays in the ecosystem
- API – a service being used or offered
- Application – a logical element used for API deployment
- Characteristic Value – key-value pairs representing characteristics/attributes of nodes

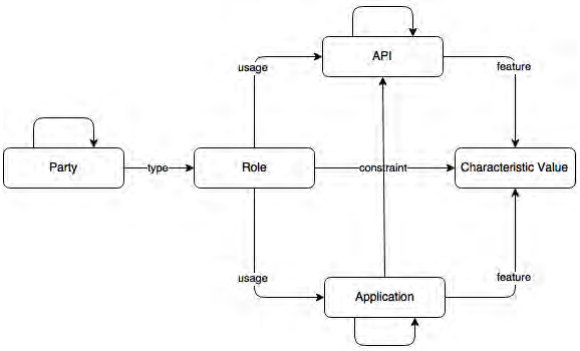


Figure 1 - API model structure

B. Resource Description Framework

The Resource Description Framework (RDF) is a language for representing meta-data. The *RDF data model* [9] defines the structure of the RDF language. The data model consists of three data types:

- Resources: All data objects described by a RDF statement are called resources. For example, resources are web sites or books.
- Properties: A specific aspect, characteristic or relation of a resource is described by a property. For example, properties are the creation date of a web site or the author of a book.
- Statements: A statement combines a resource with its describing property and the value of the property. RDF statements are the structural building blocks of the language.

An RDF statement is typically expressed as a "resource - property - value" - triple, commonly written as *P(R,V):* A resource *R* has a property *P* with value *V*. These triples can also be seen as object-attribute-value triple. Statements can also be expressed as graphs with nodes for resources and values where directed edges represent the properties.

RDF representation of API model property graph enables openness and integration with other datasets, using the principles of open linked data [13]. This extension mechanism would allow integration of additional semantic models and would provide better insight and service selection process, tailored to specific domain (e.g. specific industry usage).

RDF representation of the underlying API model graph will also enable usage of standardized queries [10] (e.g. SPARQL, RDQL) to analyze the graph and inspect nodes, relationships and characteristics. This feature will be used in determining dominant services from the value provision and value capture perspective.

C. API model support for execution

In order to the analyze the currently available APIs in the ecosystem, as well as transforming the API models described in RDF format into elements that support API execution (invocation of APIs) we introduce the concept of serialization/deserialization of API model RDF representation. Currently, de facto standard format for REST API data objects is JSON (JavaScript Object Notation), enabling a lightweight approach to exposure and invocation of services. Although RDF representation as JSON would seem like a good fit, the problems in the transformation of directed graphs into hierarchical tree structure of RDF have forced the creation and introduction of JSON-LD (JSON Linked Data) [12], and we are proposing the usage of this model for RDF serialization. JSON-LD is fully compatible with standard JSON and allows simplified serialization/deserialization of RDF and brings the proposed model to be consumable by APIs itself.

IV. SERVICE SELECTION

A. Service Dominance

A number of service selection methods have been developed, aiming to achieve a composite service with the best user desired characteristics. In some related research [3], APIs are selected depending on a predefined objective function. In this case, a weighting mechanism is leveraged where service users express preferences over different characteristics (parameters) as numeric weights. The composite service gaining the highest value is selected and returned to the user. As it proved to be very challenging to transform user preferences into numeric weights, some researchers have tried solving the problem using the dominance relationship between services [4] (also called skyline). With dominance relationship between services, for a task, only skyline services are selected for further composition, reducing the number of candidate services.

Given a set of *d*-dimensional points, a point *p* is said to dominate another point *q* if it is better than or equal to *q* in all dimensions and better than *q* in at least one, denoted as $p \prec q$. The skyline points are the subset of points that are not dominated by other points [2]. In a scenario of improving composing and assembly of services with proposed skyline technique, given a predefined value network schema, top-k skyline services are selected for each required value flow between ecosystem participants and composed into a composite service flow.

Definition 1. Service Dominance
A service ws_i is defined as a dominate another service ws_j on Q if and only if $\forall q_i \in Q$, $ws_i.q_i \geq ws_j.q_i$ and $\exists q_t \in Q$, $ws_i.q_t \geq ws_j.q_t$.

Definition 2. Service Skyline
A service ws_i is a skyline service if and only if there does not exist a service $ws_j \neq ws_i$ dominating ws_i. Service skyline could also be called full skyline or free skyline.

Definition 3. Value Preference Service Dominance
Given a value preference space P, a service ws_i is said to dominate another service ws_j on P if and only if $\forall p_i \in P$, $ws_i.p_i \geq ws_j.p_i$ and $\exists p_t \in P$, $ws_i.p_t > ws_j.p_t$.

Definition 4. Value Preference Service Skyline (VPS)
A service ws_i is a skyline service on P if and only if there does not exist a service $ws_j \neq ws_i$ dominating ws_i on P. We use *s-VPS(WS,P)* to denote the set of skyline services in *WS* on *P*, where *s* represents the number of attributes in *P*. *s-VPS(WS,P)* can be abbreviated as *s-VPS* or *VPS*.

The first two definitions describe skyline services over full service characteristics dimension, and Definition 3, 4

show what a value preference skyline service is. Given an *l*-dimensional value preference space v_i and *d*-dimensional service characteristic space *Q*, when $l = d$, the value preference service dominance is equivalent to the original dominance in full skyline.

B. *Finding top valued services*

Using VPS, top-k candidate services for each task are selected for the further composition. In order to rank services in a same candidate service class, a score is computed for each service under a user preference space P, denoted as $score_i(ws)$ for the service *ws*. The score is computed as follows: $score_i(ws) = \sum_{j=1}^{l} ws.q_j$, where $q_j (q_j \in P)$, $0 \leq ws.q_j \leq 1$ and $ws.q_j$ is the parameter value after normalization.

For selecting top-*k* services, a dominant number is defined for service ranking. For the service *ws*, its dominant number is equivalent to the number of services dominated by *ws* in *VPS*.

V. CONCLUSION

In this paper, a novel model for linking ecosystem value creation and value capture with representation of API parameters required for improvement of API composability and assembly, is given.

The paper presented "value co-creation" and "value capture" as the ultimate objective of services, and the proposed method starts with the identification of service participants (ecosystem participants), their value expectations and value exchange relations. Value network model is subsequently transformed into a property graph representing the API model and further defined as an RDF model, capable of being queried as well as serving as a foundation for API execution models (JSON-LD based). The task of finding the optimal API, given the CBA derived set of value parameters is provided using Top-k Service Skyline approach.

This approach will be further enhanced using the value oriented patterns that will enhance the matching of services in the ecosystem. Additional enhancements will include usage of Web Ontology Language (OWL) that bring additional features such as logical operators and expressiveness.

REFERENCES

[1] F. Li , C. Tian, R. Cao and S. Jiang, "Value network model for service ecosystem in business environment", Network Operations and Management Symposium, 2008. NOMS 2008. IEEE

[2] D. Skoutas, D. Sacharidis, A. Simitsis and T. Sellis, "Serving the Sky: Discovering and Selecting SemanticWeb Services through Dynamic Skyline Queries", 2008 IEEE International Conference on Semantic Computing, pp. 222-229, 2008.

[3] M. Alrifai, T. Risse, "Investigating Web Services on the World Wide Web", Proceedings of the 17th International Conference on World Wide Web, pp. 795-804, 2008.

[4] Q. Yu, A. Bouguettaya, "Efficient Service Skyline Computation for Composite Service Selection", IEEE Transactions on Knowledge and Data Engineering, 2011.

[5] The Open Group: The Open Group Architecture Framework (TOGAF), Version 8.1.1, Enterprise Edition, 2007.

[6] IEEE, IEEE Recommended Practice for Architectural Description of Software Intensive Systems, The Institute of Electrical and Electronics Engineers, Inc. IEEE Std 1471-2000, 2000.

[7] Armour, F.J., et al.: A Big Picture Look at Enterprise Architectures, IEEE IT Professional, January/February 1999.

[8] Flaxer, D., et al.: Using Component Business Modeling to Facilitate Business Enteprise Architecture and Business Services at the US Department of Defence, Proceedings of the 2005 International Conference on e-Business Engineering (ICEBE'05), 2005.

[9] "Resource Description Framework", https://www.w3.org/RDF/, 2014.

[10] "SPARQL 1.1 Overview", https://www.w3.org/TR/sparql11-overview/, 2013.

[11] "JavaScript Object notation – JSON", http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf, 2013.

[12] "JSON-LD 1.0", https://www.w3.org/TR/json-ld/, 2014

[13] F. Bauer, M. Kaltenbock, "Linked open data: The essentials", ISBN: 978-3-902796-05-9. 2012.

[14] J. Gordijn, P. van Eck, and R. Wieringa, "Requirements Engineering Techniques for e-Services. Service-Oriented Computing," The MIT Press, pp. 340-361, 2009.

Service Level Agreement - SLA, Raspoloživost servisa i kvalitet sistema u telekomunikacijama

A. mr Dušanka Glamočanin, dipl.inž.el.*
* Mtel ad, Banja Luka, Bosna i Hercegovina
e-mail: dusanka.glamocanin@mtel.ba

Sažetak - U ovom radu se govori o SLA *Service-level agreement* u vezi sa sistemom kvaliteta i pouzdanosti prilikom pružanja usluga i servisa u telekomunikacijama. Razmatraju se različiti nivoi primjene, različiti načini mjerenja SLA, karakteristične vrijednosti i specifični primjeri SLA u telekomunikacijama – davaoci internet usluga i informacionim tehnologijama. U vezi sa SLA govori se i o OLA *Operational-level agreement*.

I. UVOD

Service-level agreement (SLA) ili Sporazum o nivou usluga je dogovor između dvije ili više strana o nivou, kvalitetu i odgovornostima prilikom pružanja usluga. Pri tome je potrebno definisati uslugu koja se pruža, način mjerenja kvaliteta, način upravljanja uslugom, garancije i slično. SLA je ujedno i dio ugovora o uslugama gdje se formalno definiše nivo usluga. Između davaoca usluga i korisnika usluga definišu se aspekti usluge - obim, kvalitet, odgovornosti. Zajednička karakteristika jednog SLA je ugovoreni rok isporuke (usluga ili izvođenje). Davaoci internet usluga i telekomi bi trebali uključiti SLA kao dio njihovih ugovora sa korisnikom da definišu nivoe za uslugu koja se prodaje. Najčešće će SLA imati neku tehničku definiciju u smislu prosječno vrijeme između dva kvara (MTBF - *Mean Time Between Failures*), ili prosječno vrijeme za oporavak (MTTR - *Mean Time To Repear (Recovery);* precizirajući koja strana je odgovorna za izvještavanje o kvarovima ili plaćanje naknade; odgovornost za različite brzine prenosa; protok; *jitter* ili slične mjerljive detalje.

II. SPORAZUM O NIVOU USLUGA (SLA)

Sporazum o nivou usluga je sporazum između dvije ili više strana, gdje je jedan kupac, a ostali su davaoci usluga. Ovo može biti pravno obavezujući, formalni ili neformalni ugovor (na primjer, interni odnosi između različitih sektora). Sporazum može uključivati posebne organizacije ili različita odjeljenja unutar jedne organizacije. Ugovori između davaoca usluga i ostalih trećih strana se često (pogrešno) nazivaju SLA - jer je nivo usluga postavljen od strane kupca, te ne može biti dogovor između trećih lica; ovi sporazumi su jednostavno ugovori. Samo sporazumi definisani sa korisnikom usluga o nivou kvaliteta usluga su SLA. Sporazumi o operativnom nivou ili OLA - *Operational-level agreements*, međutim, mogu biti korišteni kao podrška za SLA.

Operational-level agreement (OLA) ili Sporazum o operativnom nivou definiše međusobno zavisne odnose u unutrašnjoj strukturi organizacije koja isporučuje uslugu, čiji je cilj isključivo podrška SLA. Ukoliko isporučilac usluge ima više organizacionih cjelina koje učestvuju u isporuci usluge, onda Sporazum o operativnom nivou opisuje međusobne odgovornosti svake organizacione cjeline prema drugoj cjelini, uključujući i proces i vremenski okvir za isporuku njihovih usluga. Cilj OLA je predstaviti jasan, koncizan i mjerljiv opis internih odnosa kao podrške za SLA.

OLA se definiše na sljedećim nivoima:

- Sporazum na organizacionom nivou,
- Sporazum na operativnom nivou,
- Sporazum na nivou operacije.

SLA treba da uključuje segmente: definiciju usluga, mjerenje performansi, upravljanje problemima, dužnosti kupca, garancije, oporavak od katastrofe, raskidanje ugovora [1] kako bi se osiguralo da SLA bude konstantno na željenom nivou. Ovi sporazumi treba da tačno definišu linije razgraničenja. Uključene strane su obavezne za otvorenu komunikaciju. Izvršenje ugovora (nagrade i kazne) treba da se striktno sprovode, mada najčešće SLA ostavljaju prostor za godišnju reviziju, tako da je moguće napraviti promjene na osnovu novih informacija [2].

SLA je uveden kasnih 1980-ih od strane telekom operatera u fiksnoj telefoniji kao dio ugovora sa svojim poslovnim klijentima. Danas je uobičajeno za kupca da angažuje davaoca usluga i pri tom definiše SLA u gotovo svim industrijama i tržištima za razne vrste ugovora o uslugama. Interni sektori (kao što je IT, HR, nekretnine) u većim organizacijama su usvojili ideju korištenja SLA sa svojim "internim" kupcima - korisnicima u drugim sektorima, unutar iste organizacije. Jedna od prednosti uvođenja SLA je poređenje kvaliteta usluga na više lokacija ili između različitih poslovnih jedinica. Može se koristiti i za testiranje na tržištu i da se omogući poređenje između internih sektora i spoljnih davaoca usluga.

SLA su bazirani na "izlaznim vrijednostima", jer je usluga koju je kupac primio, u stvari predmet sporazuma. Davaoci usluga mogu organizovati sve svoje kapacitete tako da pruže traženu uslugu. Kupci mogu odrediti način na koji se servis ili usluga isporučuje, kroz specifikaciju (specifikacija nivoa usluga) i pomoću zavisnih "ciljeva", osim onih koji se odnose na nivo usluga. Ova vrsta ugovora je poznata kao "ulaz" SLA. U posljednje vrijeme te specifikacije ne definišu kupci, kako organizacije

postaju zahtjevnije i kako se i metodologija rizika prebacuje na davaoca usluga.

SLA su najčešće definisani na različitim nivoima:

- SLA baziran na korisniku: Sporazum sa pojedinačnom grupom korisnika, koji pokrivaju sve usluge koje koriste. Na primjer, jedan SLA između davaoca usluge (IT davaoc usluga) i finansijskog sektora velike organizacije za usluge kao što su finansijski sistem, platni sistem, sistem naplate, sistem nabavke/kupovine, itd.
- SLA baziran na servisu: Sporazum za sve korisnike koji koriste usluge koje nudi davaocusluga. Na primjer: Auto servis nudi uobičajenu uslugu svim klijentima i održavanja kao dio ponude sa univerzalnim zaduženjima. Davaoc usluga mobilne telefonije nudi uobičajenu uslugu svim klijentima i određeno održavanje kao dio ponude s univerzalnim zaduženjima. Kod sistema e-pošte za cijelu organizaciju mogu se pojaviti problemi u ovoj vrsti SLA kao nivoa usluga, jer ponuđena usluga može varirati za različite kupce (na primjer, direktori mogu koristiti LAN sa velikim brzinama dok su ostalim zaposlenima dostupne niže brzine).
- Višenivoiski SLA: SLA je podijeljen na različite nivoe, svaki nivo je adresiran na različitu grupu korisnika za iste usluge, u istom SLA.
  - Korporativni nivo SLA: pokriva sve opšte tačke upravljanja nivoom usluga (često nazvane SLM - S*ervice Level Management*) i odnosi se na svakog korisnika u čitavoj organizaciji. Ove tačke će vjerovatno biti manje podložne promjenama i nisu potrebna česta ažuriranja.
  - Korisnički nivo SLA: pokriva kompletno upravljanje nivoom usluga za određenu grupu korisnika, bez obzira na usluge koje se koriste.
  - Servisni nivo SLA: pokriva kompletno upravljanje nivoom usluga za specifične usluge, u odnosu na ove specifične grupe korisnika.

Cilj SLA je da i klijent i davaoc usluga korektno urede svoje odnose, na obostranu korist. Pravilno definisan SLA prije početka razvoja samog projekta omogućuje kvalitetnu saradnju, prilikom razvoja samog projekta, za vrijeme upotrebe razvijenog proizvoda, nakon njegove implementacije i primopredaje.

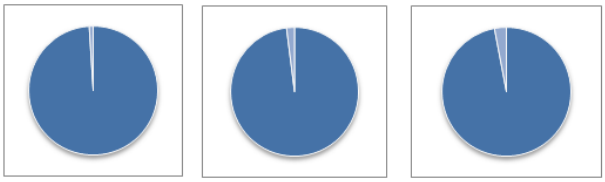
III. NAČINI MJERENJA - METRIKE

Performanse usluga, s odgovarajućim ciljevima za postizanje željenog nivoa usluga, mogu se mjeriti na različite načine. Najčešće se u upravljanju IT uslugama kao primjer navodi Pozivni centar. Uobičajeno je da se mjeri:

- Stopa odustajanja: Postotak poziva od kojih se odustalo dok se čeka na odgovor.
- Prosječna brzina na odgovor (ASA - *Average Speed to Answer):* Prosječno vrijeme potrebno za odgovor, koje se obično mjeri u sekundama.
- TSF - *Time Service Factor*: Postotak poziva na koje je odgovoreno u definisanom vremenskom okviru, npr, 80% u 20 sekundi.
- FCR - *First-Call Resolution*: Procenat dolaznih poziva na koje se može odgovoriti bez upotrebe povratnih poziva ili bez zvanja dodatne pomoći.
- TAT - *Turn Around-Time*: Vrijeme potrebno da se izvrši određeni zadatak.
- Srednje vrijeme za oporavak (MTTR - *Mean Time To Recover*): Srednje vrijeme potrebno za oporavak nakon ispada usluge.

Dostupnost ili neprekidni rad (*uptime*) je takođe uobičajena metrika, često se koristi za podatkovne usluge, kao što su *shared hosting*, virtualni privatni serveri i namjenski serveri. Uobičajeno je da SLA uključuju postotak vremena kada je mreža u radu, kada je napajanje u radu itd.

Primjer SLA za dostupnost dat je na sljedećoj slici:



99% dostupnost
7h 12m mjesečna nedostupnost
3d 15h 36m godišnja nedostupnost

98% dostupnost
14h 24m mjesečna nedostupnost
7d 7h 12m godišnja nedostupnost

97% dostupnost
21h 36m mjesečna nedostupnost
10d 22h 48m godišnja nedostupnost

Dostupnost	Mjesečna nedostupnost	Godišnja nedostupnost
100%	0m	0m
99.999%	0.4m	5m
99.99%	4m	52m
99.9%	43m	8h 46m
99.8%	1h 26m	17h 31m
99.7%	2h 10m	1d 2h 17m
99.6%	2h 53m	1d 11h 2m
99.5%	3h 36m	1d 19h 48m
99.4%	4h 19m	2d 4h 34m
99.3%	5h 2m	2d 13h 19m
99.2%	5h 46m	2d 22h 5m
99.1%	6h 29m	3d 6h 50m
99.0%	7h 12m	3d 15h 36m

Slika 1. Odnos SLA za dostupnost i stvarne (ne)dostupnosti usluge. Izvor: www.pingdom.com

Metrike koje su najčešće zastupljene kod definisanja SLA u IT su:

- Raspoloživost i neprekidan rad – procenat vremena kada su usluge raspoložive i dostupne korisniku,
- Broj korisnika koji mogu biti usluženi istovremeno,

- Tačno definisane vrijednosti performansi koje se postavljaju kao referentne vrijednosti i prema kojima se aktuelne performanse periodično uporeduju,
- Vrijeme odziva sistema ili aplikacije,
- Vrijeme odziva *Help Deska* za različite klase problema.

Osim definisanja metrike performansi, prilikom usaglašavanja SLA između kupca i davaoca usluga može se definisati i plan za rješavanje problema. Taj plan treba da objašnjava na koji način će davaoc usluga nadoknaditi ugovoreno kupcima u slučaju povrede ugovora. Vrlo je važno da SLA, jednom uspostavljen, treba periodično revidirati i ažurirati u skladu s promjenama u tehnologiji i raznih drugih promjena i uticaja.

IV. SPECIFIČNE PRIMJENE

SLA potiče od mrežnih usluga, ali se sada veoma često koristi na području kompletnih telekomunikacija - davaoci internet usluga, web usluge i usluga računarstva u oblaku (*cloud computing*).

A. Davaoci internet usluga na okosnici

Uobičajeno je da davaoci internet usluga na okosnici (*Backbone Network Servis Provider)* eksplicitno navode vlastiti SLA na svojoj web stranici [3][4][5]. Američkim Zakonom o telekomunikacijama iz 1996. godine ne propisuje se izričito da su kompanije dužne da imaju SLA, ali je to ipak preporuka firmama da ga definišu [6].

Za davaoce usluga, SLA je dio ugovora ili poseban ugovor u kom se definiše vrsta i nivo usluge između ponuđača usluge (prodavca, odnosno *service provider*) i klijenta (kupca, odnosno *customer*a). Potrebno je naglasiti da je u pitanju nivo usluge.

Potpisivanjem ugovora za SLA, kupac ne dobija samo željeni novo usluge, već se osigurava da će imati mogućnost održavanja sistema, kao i mogućnost dodatne nadogradnje već razvijenog sistema. Isto tako, kupac ima garancije da će cijena naplate usluge biti korektna, odnosno da prodavac neće naplaćivati uslugu više od njene realne cijene.

Sa druge strane, sa ovakvim ugovorom, prodavac usluge zna da ima zagarantovan priliv sredstava i da može računati na dalju saradnju sa klijentom.

Definisanje i potpisivanje SLA omogućava definisanje detalja, kao što je nivo tehničke podrške klijentu, kao i naknadna obuka klijenta za rad na sistemu.

Za davaoce nternet usluga, najbolje je SLA definisati po sljedećim principima:

- Definisati vremenski okvir za koji će da važi SLA (od-do). Najbolje po broju mjeseci ili broju godina.
- Definisati broj radnih sati, najbolje na mjesečnom ili sedmičnom nivou.
- Definisati cijenu po radnom satu.
- Definisati šta obuhvata nivo usluge i kako se tarifira. Cijena tehničke podrške, naknadne obuke, programiranja ili dizajniranja upotrebom različitih alata treba da se različito tarifira.

Svaki klijent kao i svaki prodavac imaju pravo na svoje viđenje načina definisanja stavki koje sadrži SLA. Prilagođavaju se svakom konkretnom slučaju. Najčešće se elementi ovakvih ugovora ne objavljuju javno.

Mnogi SLA se definišu tako da započinju nakon isticanja garantnog roka. Ovakav ugovor u suštini nije najbolji za klijenta, jer po tome znači da u garantnom periodu ne postoji mogućnost dodatne nadogradnje sistema, kao i tehničke podrške ili dodatne obuke korisnika sistema (praktično da sve zavisi od dobre volje prodavca i vremena koje ima na raspolaganju).

B. WSLA

WSLA - *Web service level agreement* ili Sporazum o nivou web usluga je neka vrsta standarda za praćenje usklađenosti SLA za web usluge. To omogućuje autorima da definišu način praćenja željenih performansi aplikacija za web usluge, kao i akcije koje treba obaviti kada performanse nisu ispunjene.

Primjer je WSLA Specifikacija, objavljena od strane IBM-a 28. januara., 2001.

C. Računarstvo u oblaku

Osnovna prednost računarstva u oblaku je dijeljenje zajedničkih resursa, koje je postignuto okruženjem u kome se dijeli infrastruktura. Usluga koja se nudi, pa i SLA, se može prilagoditi svakom korisniku usluge. Za svaku novu uslugu treba da bude definisan novi SLA. Ako se usluga mijenja, treba da se preradi SLA koji se odnosi na tu uslugu [7]. Iskustvo krajnjeg korisnika ili mogućnosti krajnjeg korisnika da koristi definisane resurse je glavna osnova za mjerenje, praćenje i izvještavanje o performansama računarstva u oblaku. Loša strana računarstva u oblaku u odnosu na SLA je problem u određivanju uzroka za prekid usluga zbog složene prirode okruženja.

Pošto su aplikacije preseljene sa tačno određenog hardvera u računarstvo u oblaku, sve aplikacije treba da dostignu isti ili čak zahtjevniji nivo usluge u odnosu na prije. SLA za usluge računarstva u oblaku najčešće su fokusirane na karakteristike podatkovnog centra i u posljednje vrijeme uključuju karakteristike mreže za podršku SLA s-kraja-na-kraj.

Svaka strategija upravljanja SLA razmatra dvije potpuno odvojene faze: pregovaranje dok se ne potpiše ugovor i praćenje njegove realizacije u realnom vremenu sa mogućnostima povremene revizije. Tako, upravljanje SLA obuhvata definisanje ugovora za SLA: osnovnu šemu parametara za kvalitet usluga - QoS; pregovaranje oko SLA; monitoring SLA; detekciju kršenja SLA; i izvršenje SLA - prema definisanoj politici [8]. Tipični primjeri kompanija koje su davaoci infrastrukture računarstva u oblaku a koji imaju definisane korisničke SLA su Microsoft, Amazon.com, Rackspace i Salesforce.com.

HP Helion Public Cloud nudi SLA za korisnika usluge sa iznosom 99,95% mjesečne raspoloživosti za sve proizvode unutar njihove infrastrukture. Ukoliko drugi davaoc usluga nudi uslugu sa mjesečnom raspoloživosti od 99,9%, razlika u procentima nije velika ali je veoma velika razlika u raspoloživosti sistema mjereno u satima.

D. *Outsourcing*

Outsourcing podrazumijeva prenos odgovornosti od jedne organizacije na njihove dobavljače. Upravljanje ovim novim saradnjama se vrši kroz jedan ili više SLA. U ugovoru se mogu definisati novčane kazne pa čak i pravo na raskid ugovora ukoliko su vrijednosti koje su definisane za SLA u velikom procenatu vremena izvan dogovorenih vrijednosti. Definisanje, praćenje i upravljanje SLA je važan dio upravljanja odnosa vezanih za prenošenje odgovornosti. Najčešće se SLA dogovaraju unaprijed i čine sastavni dio ugovora za prenošenje odgovornosti na druge organizacije, i upravo su oni jedan od primarnih alata za upravljanje takvim ugovorima.

E. *IT*

SLA su jednako važni u IT industriji. Ovakvim ugovorima definiše se vrsta, nivo i obim usluga koje dobavljač obezbjeđuje klijentu/korisniku.

SLA za IT oblast bi trebalo da pokrije sljedeće [14]:

- Opis usluga koje se isporučuju (npr. šta korisnik želi i na šta je dobavljač spreman da se obaveže isporučiti),
- Usaglašeni standardi kvaliteta (npr. svjetski, regionalni, lokalni standardi ili mjerljivi nivoi usluga koje dobavljač mora dostići), kao što su vrijeme odziva, vrijeme obrade zahteva, dostupnost sa specifikacijom: šta, kada, gdje, ko i kako,
- Načini isporuke usluga, servisa ili proizvoda,
- Posljedice neispunjavanja dogovorenih procedura ili standarda (npr. nadoknada u nekom obliku),
- Procedure kojima se korisniku omogućava da kontroliše da li dobavljač ispunjava definisane obaveze,
- Procedure za kontrolu i upravljanje promjenama ukoliko se ukaže potreba (npr. zamjena ili nadogradnja neke usluge koju je obezbjeđivao dobavljač po ugovoru),
- Uslovi koji se tiču sigurnosti, kao veoma bitnog aspekta IT (npr. pristup lokaciji korisnika ili podacima),
- Procedura za oporavak od prekida u radu ili pada sistema iz raznih razloga,
- Opciono: sastanci između korisnika i dobavljača radi pregleda ispunjavanja ugovornih obaveza, promjena i planiranja, njihova dinamika održavnja.

Idealno je da SLA bude ugovor sa kontinuiranim odnosom dvije strane. Sam proces pripreme i kreiranja SLA ugovora je isto onoliko važan koliko i sam ugovor, njegova realizacija, praćenje i korigovanje.

Ugovori o održavanju softvera i ugovori o održavanju računarske mreže i IT infrastrukture su primjeri SLA ugovora za IT. Alternativno, SLA može biti jedan dio šireg ugovora o uslugama.

V. PRIMJER SLA JEDNOG EVROPSKOG DAVAOCA USLUGA ZA NAJAM VLAKNA, KAPACITET, IP I USLUGE

Ugovorima između isporučioca i korisnika usluge u telekomunikacijama najčešće se definiše nivo kvaliteta usluge i korisnikova prava da aplicira za dodatni kredit u slučajevima kada nije dostignut definisani nivo usluge.

A. *Isporuka usluge i kredit za kašnjenje u isporuci usluge*

Usluga treba da bude instalirana i operativna u skladu sa definisanim specifikacijama, na vrijeme definisano kao RFS *(RFS - Ready for Service)* datum i u skladu sa svim ostalim specifičnim uslovima za definisanu uslugu. Ukoliko RFS datum definisan sa operatorom nije postignut, korisnik usluge ima pravo da ostvari kredit u skladu sa tabelom prikazanom u nastavku, za svaki dan kašnjenja usluge u odnosu na RFS datum.

TABELA 1. KREDIT PRIMJENLJIV NA IZNAJMLJENI KAPACITET ZAVISNO OD PERIODA KAŠNJENJA

Period kašnjenja	Kredit primjenljiv na iznajmljeni kapacitet
Broj dana neopravdanog kašnjenja između dogovorenog i stvarnog RFS datuma za definisanu uslugu	1% od dogovorenog mjesečnog zaduženja po danu kašnjenja do ukupno 30 dana

Korisnik nema pravo na bilo kakav kredit po osnovu RFS kašnjenja koji je nastao zbog korisnikovih postupaka ili propusta, grešaka trećih lica u isporuci ili omogućavanju usluge, ili u slučaju više sile.

B. *Izračunavanje raspoloživosti usluge*

Prilikom definisanja usluge u telekomunikacijama, najčešće se definiše raspoloživost, kao osnovna mjera kvaliteta usluge. Period mjerenja raspoloživosti usluge treba da bude u korelaciji sa periodom mjesečnog obračuna. Raspoloživost se mjeri mjesečno (ili kvartalno u slučaju *multicast* usluge), a počinje u mjesecu u kojem je isporučena kompletna usluga.

$$Raspoloživost = \frac{UkupnoVrijeme - \sum VrijemeSaGreškom}{UkupnoVrijeme} * 100\%$$

Gdje je:

UkupnoVrijeme	Ukupno vrijeme u periodu mjerenja
$\sum VrijemeSaGreškom$	Suma svih vremena sa greškom u periodu mjerenja

C. SLA Service Level Measurement – Dogovor o nivou usluge

Izgubljeni paket i kašnjenje u mreži su mjereni šaljući aproksimativno 1000 UDP (*User Datagram Protocol*) dužine 96 bajtova svakih 10 minuta do željenog servera lociranog u IP okosničkom čvoru koji je primarni čvor u mreži označen od strane davaoca usluge. Korisnikov interfejs u IP okosničkim ruterima/svičevima biće mjereni korištenjem SNMP protokola (*Simple Network Management Protocol).*

Garancije za IP okosnički nivo usluge, vrstu usluge koja se koristi i krediti su navedeni u nastavku.

D. Garantovana raspoloživost i krediti

Raspoloživost usluge unutar IP okosnice je 99.999%, mjereno kako je navedeno ranije u tekstu. Korisnik ima pravo da traži kredit ili olakšice na nivou zaduženja po ratama za uslugu koja je afektirana za 1/60 dijela dana za svaki minut sa greškama tokom kalendarskog mjeseca. Pri tom se smatra da IP okosnička usluga nije neraspoloživa u slučaju problema sa CPE opremom korisnika i u takvom slučaju nema prava na kredit.

Raspoloživost usluge se primjenjuje na usluge Ethernet, IP tranzit i IP konektivnost.

E. Garantovano kašnjenje u mreži i krediti

Termin „kašnjenje u mreži“ se odnosi na period potreban da se IP paket transportuje od izvora do destinacije unutar mreže i nazad. Najčešće se naziva RTT – *Round Trip Time*. Davaoc usluge garantuje da srednje vrijeme kašnjenja u mreži unutar IP okosnice u granicama posmatranog regiona neće prelaziti vremena navedena u nastavku, tokom kalendarskog mjeseca:

TABELA 2. GARANTOVANO SREDNJE VRIJEME KAŠNJENJA U MREŽI (TOKOM KALENDARSKOG MJESECA)

Region	Opis	Garantovano srednje vrijeme kašnjenja u mreži
Unutar Evrope	Srednji RTT kroz sve linkove između evropskih primarnih čvorova	35 ms
Evropa-SAD	Srednji RTT između Evrope i SAD (sa PoP u Njujorku)	99 ms
Unutar SAD	Srednji RTT kroz sve linkove između SAD primarnih čvorova	65 ms

Kašnjenje u mreži neće biti detektovano samo od korisnika već i od strane davaoca usluge na osnovu srednje vrijednosti mjernih uzoraka uzetih tokom kalendarskog mjeseca. Kada garantovana vrijednost nije postignuta, korisnik će pokrenuti zahtjev za kredit jednak zaduženju na nivou rate za afektovanu IP uslugu za dane tokom kalendarskog mjeseca u kom garantovano kašnjenje u mreži nije postignuto.

Kašnjenje u mreži se primjenjuje na usluge IP tranzita i IP konektivnost.

F. Garantovani izgubljeni paketi i krediti

Davaoc usluge garantuje da srednji broj izgubljenih paketa unutar IP okosnice kroz posmatrani region neće prelaziti pragove navedene u Tabeli 3, tokom kalendarskog mjeseca.

TABELA 3. GARANTOVANI SREDNJI BROJ IZGUBLJENIH PAKETA U MREŽI (TOKOM KALENDARSKOG MJESECA)

Region	Opis	Garantovani srednji broj izgubljenih paketa u mreži
Unutar Evrope	Srednji broj izgubljenih paketa kroz sve linkove između evropskih primarnih čvorova	0.1 %
Evropa-SAD	Srednji broj izgubljenih paketa između Evrope i SAD (sa PoP u Njujorku)	0.1 %
Unutar SAD	Srednji broj izgubljenih paketa između SAD primarnih čvorova	0.1 %

Izgubljeni paketi će biti detektovani i od strane davaoca usluge na osnovu srednje vrijednosti mjernih uzoraka uzetih tokom kalendarskog mjeseca. Kada garantovana vrijednost nije postignuta, korisnik će pokrenuti zahtjev za kredit jednak zaduženju na nivou rate za afektovanu IP uslugu za dane tokom kalendarskog mjeseca u kom garantovani srednji broj izgubljenih paketa u mreži nije postignut.

Srednji broj izgubljenih paketa u mreži se primjenjuje na usluge IP tranzita i IP konektivnost.

Bez obzira na višestruke neželjene događaje ili greške, postizanje garantovanog nivoa usluge navedenih u ovom SLA, maksimalni kazneni kredit koji može koristiti korisnik tokom mjesečnog obračunskog perioda je ograničen na 30% od prethodnog mjesečnog zaduženja za afektovanu uslugu.

VI. PRIPREME OPERATORA MOBILNE TELEFONIJE IZ BIH ZA USPOSTAVLJANJE SLA

Sva nova tehnička rješenja za usluge koje nude operatori mobilne telefonije treba da ukažu na sposobnost operatera mobilne telefonije da obezbjedi veću skalabilnost, pouzdanost, sigurnost i kvalitet usluge QoS, kao preduslove za uspostavu SLA za postojeće i nove usluge, posebno saobraćaja u realnom vremenu. Potrebno je obezbijediti i sljedeće funkcionalnosti:

- Otpornost na određene greške mrežnih sekcija - mogućnost dinamičke rekonfiguracije mreže, bez ograničenja protoka za krajnje korisnike na radnim putevima, rezervne puteve i mogućnost da podržava višestruke uzastopne ili paralelne kvarove linija u mrežnoj topologiji.
- Fleksibilnost - mogućnost dinamičkog ili programiranog konfigurisanja mreže ili mijenjanja topologije u zavisnosti od trenutnih potreba.

- Prioriteti - mogućnost prenosa kanala većeg prioriteta, umjesto kanala sa nižim prioritetom, u slučaju greške na mreži.

Svi projekti kojima se planiraju i realizuju nove usluge za koje će se definisati SLA treba da su urađeni u skladu sa važećim tehničkim propisima pridržavajući se trenutnih planova izgradnje kablovske i bežične pristupne mreže, transportne mreže, okosničke mreže i servisnih platformi.

VII. ZAKLJUČAK

SLA - *Service-level agreement* kao sporazum o nivou usluga između davaoca i korisnika usluga postaje nezaobilazan dio ugovora između te dvije strane u području telekomunikacija i informacionih tehnologija. Pomoću SLA definiše se usluga koja se pruža, nivo i kvalitet usluge ili servisa, odgovornosti, način mjerenja kvaliteta, način upravljanja uslugom, garancije, krediti. Postoje različiti nivoi primjene, različiti načini mjerenja SLA i karakteristične vrijednosti. U današnje vrijeme, svaki korisnik usluge u telekomunikacijama i informacionim tehnologijama želi visok nivo usluge u odnosu na naknadu koju plaća, a davaoci usluge žele tačno da znaju šta su njihove obaveze. Zato je potrebno definisati SLA u svakom ugovoru između davaoca i korisnika usluge ili servisa, posebno u oblasti davaoca internet usluga i informacionih tehnologija.

LITERATURA

[1] Orla O'Donnell (October 2014), „The Service Level Agreement", IPA Institute of Public Administration, preuzeto sa www.ipa.ie/pdf/ServiceAgreementsReport_2014.pdf

[2] Five Key Points for Every SLA, preuzeto 03.12.2015. sa www. Dell.com.

[3] SLA - Global IP Network - NTT America, preuzeto sa www.us.ntt.net

[4] Global Latency and Packet Delivery SLA - Verizon Business, preuzeto sa www.verizonenterprise.com/terms/global_latency_sla.xml

[5] „AT&T High Speed Internet Business Edition Service Level Agreement", preuzeto sa www.att.com

[6] Telecommunications Act of 1996#SEC. 101. ESTABLISHMENT OF PART II OF TITLE II. preuzeto sa http://thomas.loc.gov/cgi-bin/query/F?c104:1:./temp/~c104ZqPyd9:e19323:

[7] Cloud Service Level Agreement Standardisation Guidelines, European Commission, Brussels, 2014, preuzeto sa https://ec.europa.eu/digital-single-market/en/news/cloud-service-level-agreement-standardisation-guidelines

[8] Georgina GALLIZO, Roland KUEBERT, Karsten OBERLE, Andreas MENYCHTAS, Kleopatra KONSTANTELI, „Service Level Agreements in Virtualised Service Platforms", eChallenges, 2009, preuzeto sa www.researchgate.net/publication/228954833_Service_level_agreements_in_virtualised_service_platforms

[9] SLA@SOI, preuzeto sa: http://sla-at-soi.eu/

[10] Alana Scheikowski „Service Level Agreement Best Practices", , preuzeto sa: http://www.amazon.com/Service-Level-Agreement-Practices-content-theartofservice-com/dp/1742443060

[11] Jimmy Desai, „Service Level Agreement, Service Agreements", preuzeto sa: http://www.amazon.co.uk/Service-Level-Agreements-Legal-Practical-x/dp/184928069X

[12] Robert Benyon, Rob Johnston, „Service Agreement", preuzeto sa: http://www.abebooks.com/book-search/isbn/9077212914/

[13] Gerard Blokdijk, Ivanka Menken, „The Service Level Agreement SLA Guide", preuzeto sa: http://www.tower.com/service-level-agreement-sla-guide-book-templates-gerard-blokdijk-paperback/wapi/112516547

[14] Miodrag Ranisavljević, „Elektronsko poslovanje", preuzeto sa: http://www.poslovnisoftver.net/ugovor-o-nivou-usluga/

[15] Dušanka Glamočanin, „Nivoi primjene i načini mjerenja kvaliteta pružanja usluga", STED 2015

Challenges of a Service Transition in Multi Domain Environment

Ivana Golub, Branko Radojević *
* Croatian Academic and Research Network (CARNet), Zagreb, Croatia
Ivana.Golub@CARNet.hr, Branko.Radojevic@CARNet.hr

Abstract- Service transition from a development to the full production operations is well defined in best practice documents with a set of steps that must be followed towards the success. Includes a lot of preparations and in advance planning, covering resources, policies, change management, definition of critical success factors and key performance indicators, challenges, risks and many other things.

In multi domain environment, with multiple domains and multi nation teams participating in the creation of the service, in parts of service transition such as service validation and testing as well as service operations, the best practice provides two-edge sword. It can at the same time help in understanding and agreeing in individual steps and present a stumbling-stone if followed blindly and literally.

The paper presents challenges and guidelines for service transition in a multi domain environment.

I. INTRODUCTION

During its lifecycle, a service has several potentially impacting milestones. Challenges follow the service from its development, strategical positioning on the market, including everyday service operations. One important point related to the service launch is when a service transitions from the development into the operations. Service transition includes a number of tests of the validity and sustainability of the service, checking all necessary functionalities and applicability, testing if it is fit for use. At this point, number of issues can arise that can lead to service retreat back into development, to re-assessment of the service readiness for production. Or the service can successfully transition into production and everyday operations.

Multi-domain environment implies existence and relevance of more then one area; it could be area of responsibility, origin, expertize, interest, geographical area, cultural, etc. In the service development and operations, multi-domain aspects can be seen as developers from multiple countries participating in the same team, the team including experts in different areas, coming from more then one organization, the infrastructure used for the service development spread in several countries and/or organisations. Additionally, the service can be aimed for multiple user categories, coming also from different countries and organizations, bringing with them different expectations and potentially various cultural and process requirements.

Each and every point of difference can present a challenge for the service development and take up in each of the service lifecycles as defined by ITIL→ Best Practice [1]. If not handled timely and appropriately, the challenge can become a risk for the service survival.

In this paper, second chapter describes multi-domain environment, third and fourth chapter present service transition from the development into operations in a multi-domain environment and its challenges. Chapter 5 proposes the resolution for those challenges, which are applied on use cases in Chapter 6. Related work is presented in the Chapter 7 and conclusions in Chapter 8.

II. SERVICES IN MULTI-DOMAIN ENVIRONMENT

A "domain" is defined in dictionaries as “an area of knowledge or activity” [2], "a field of action, thought, influence", "the territory" as well as a group of computers and devices on a network "governed by a single ruler or government" [3]. Multi-domain environment brings more then one area of responsibility, interest, expertize, multiple organizations with different visions, strategies, values, governance, processes, target market, and many more differences. Domains can be based on geographically different and distant locations.

Still, multiple domains can come together when they have a common goal and can work together towards the goal, dealing on the way with all the differences in the service of the achievement of the wanted goal. The goal can be to provide one or more services to customers. successful merge of one or more companies, to gain material or non-material benefits of any kind.

In such situations, the differences can become advantage, bridge the gaps that a single domain might have, but can as well bring a lot of challenges and risks.

For a service offering, multi-domain can mean that service developers and operators are coming from different countries bringing with them human capital, the expertize as well as some of the hardware and/or software resources. The joint can be based on a project participation, joint venture, acquisitions, merge, or any other formal and legislative way.

Examples of multi-domain environment can be found in international projects such as those funded by European Commission [4].

III. SERVICE TRANSITION

Service transition is a process a service follows on its way from development to operations. According to ITIL→ Best Practice of Service Transition, the "purpose of the service transition stage of the service lifecycle is to ensure that new, modified or retired services meet the expectations of the business as documented in the service strategy and service design stages of the lifecycle" [5].

The authors participated in and analysed transition of services in the following scenarios:

- service transition from development to production in multi-domain environment of a pan-European project GÉANT [6] and
- transition of ICT infrastructure and services of Ministry of Social Politics and Youth [7], Croatian pension insurance institute [8], Ministry of Environmental and Nature Protection [9] and five Croatian institutes under the Ministry of Health to CARNet infrastructure [10].

Each of those transitions started from the stabile starting point, solid and justified reason for change and gaps between the existing and the final wanted state.

Service transition has its organizational, technical, administrative and strategic part. Organizational part includes planning and managing of all changes related to the service and risk management. Technical part includes preparation of the needed infrastructure; testing infrastructure is needed for validation and testing phase where all relevant parts of the service will be examined to check if the service is fit for use, if the service provides all functionalities as expected, if the performance is of acceptable quality. Production infrastructure is needed in addition to the testing one, as it includes all the necessary hardware and software that presents the last resort for this service release that the users will use in the everyday operations. Technical part includes execution of the validation and testing process, successful deployment of the service instance on the production infrastructure, as well as comprehensive performance and sustainability testing. Establishment of the service desk and service support processes, establishment of the processes for knowledge database maintenance and development falls into the administrative part of the transition. Strategic part includes the check that during the transition, and once when the transition is completed, the service still remains fit for purpose.

The transition usually introduces changes, and changes can affect some of the following categories:

1. the infrastructure - when the infrastructure used during the development is not the same as the one that will be used for a production service; infrastructure includes all the hardware and software needed for the service candidate release;
2. the service itself - if during the validation and testing phase it turns out that current version of the service should be changed in some way;
3. people - when persons from the development team will not be involved with the service when the service goes into production and/or when additional people should be introduced in the service-related tasks;
4. procedures - if procedures used during the development phase are no longer applicable and/or used in the operations phase and when operations phase brings new or changed procedures that needs to be followed.
5. multi-domain aspects - resources provided by one domain could be replaced by or amended with resources from another domain (resource can be for example a person, hardware or software), tasks performed by one domain is being transferred to another domain.

Any of those changes should be analysed, evaluated and managed.

Starting situation, as well as any of the aforementioned changes can impose challenges and/or risks to the service offering.

IV. CHALLENGES AND RISKS

Each and every service lifecycle phase brings some challenges and with it associated risks to its stakeholders.

Challenges and risks recognized from the analysed cases, extended with other that changes in a service transition can bring and observed from the same five categories are:

1. the infrastructure - existing infrastructure not available for use in the production, new infrastructure resources not (yet) available, available resource not sufficient, possible delay for the infrastructure preparation that can impact the service launch, the service requires specific software that is not available (including potential issues with licences and licencing), ...;
2. the service itself - the service is not fit for purpose or not fit for use, the services require bigger changes which means that it is not ready for production, the service does not satisfy needed requirements or does not pass validity tests, ...;
3. people - lack of needed expertize or experience, insufficient number of team members, changes in the team (e.g. caused by individuals leaving to other teams/companies or handover from one team to another), not enough motivation, ...
4. procedures - existing procedures not applicable, procedures too complicated, additional procedures not coordinated with existing procedures, too many procedures, slow support caused by inappropriate procedures, old teams face difficulties accepting new procedures (relates also to the #3 - people), ...
5. multi-domain aspects - resources provided by one domain should/should not/should but cannot be replaced by or amended with resources from another domain (resource can be for example a person, hardware or software), tasks performed by one domain are in some part overlapping with another domain, a domain not aware of their area of responsibility, ...

For each risk it should be defined what is the possibility that this risk will happen, what could be the impact and what are the possible options to mitigate the risk. For each of the option, additional analysis should be performed to define the optimal solution at the time of the analysis.

V. CHALLENGES RESOLUTION

The potential risks cannot completely be eliminated. Necessary prevention can help to decrease the possibility of the appearance of a risk and/or to lower down the potential impact, but some risk will always be present.

However, the challenges can be minimized and mitigated with the implementation of the following methods and techniques, gathered in the five letter word - DARCS:

D - Description of work should be clear for each and every participant of the team, disregarding the role the person takes. From the service perspective - this should provide basic definitions for the service.

Work description influences each human-related service aspect. Therefore, if a challenge is related to the participation of one or more persons, D should be checked and acknowledged.

A - Assignment or job assignment should be clear for each and every person at each and every point in time related to the service.

Each person gets its own assignments within the team that together brings the whole puzzle together. A challenge may arise if the assignment is not well understood and by each person matched to exact set of steps to its resolution. Thus, together with the broader scope of D, A should be checked for all team members in each service aspect.

R - RACI (Responsible, Accountable, Consulted, Informed) matrix [11] of roles important for the service should be clear to all participants for all important situations.

R is implemented each time when there are more actors involved, in order to define relationships between them. If the associations are not defined or not well defined, risks and challenges may arise, e.g. in the form of overlapping tasks, unrecognized gaps, delays.

C - Communication is the most important glue that can keep everything together; exchange of data and information horizontally and vertically is even more important in the multi-domain environment and in the phase of service transition which is expected to be the most dynamic, the fastest and the shortest comparing to other service lifecycle phases.

C should be checked each time when a challenge is recognized, hopefully before it becomes a risk, in human interactions, as well as in protocols and procedures.

S - SMART goals - last but not least, all goals related to the service transition and offering should follow the SMART principle [12] and be defined to be very **s**pecific, **m**easurable, **a**chievable (agreed upon), **r**ealistic (with reasonable resources) and **t**ime-based.

SMART approach should be checked every time when the infrastructure is involved. Right measure, resources, timing, specific requirements are extremely important, and if omitted - raises risk possibility. SMART should also be checked against procedures, processes and multi-domain aspects. With regard to people, SMART can help to determine exact goals of the participation and expectations and therefore should not be ignored.

Proper combination of the techniques enables managers and team members to address every challenge and minimize the more or less expected risk, as can be seen in the following Chapter.

VI. DARCS IN ACTION

Following Tables match challenges from Chapter 4 with the proposed DARCS principle described in Chapter 5, and lists one or more techniques that can help mitigate individual challenges.

Challenge - Infrastructure	DARCS
existing infrastructure not available for production	RCS
new infrastructure resources not (yet) available	RCS
available resource not sufficient,	RCS
delay for the infrastructure preparation	RCS
software dependencies not resolved	RCS
…	?

Table 1: Challenges - Infrastructure

As it can be seen, all listed infrastructure challenges (except for the last) are related to the fact that the resources are not made available, which is included in the letter R of the SMART goal definition principle. SMART goal requires that the needed resources are known and made available, so if the SMART approach is taken, with the analysis of the resources it will be known what should be ensured before the transition can take place. If the resources are recognized in advance, then the transition can start only when the resources are ready, otherwise it can be delayed or cancelled. There is a possibility that the needs of the service were not communicated well and/or in advance to the resource owner/stakeholder, and that it caused the challenge to appear. If so, then the communication channels and messages should be checked to ensure that all relevant parties are appropriately involved. If there are more parties involved,

the need for RACI should be checked and introduced if needed.

The last option enables the possibility that not all challenges are listed. There is no ultimate solution for an un-known problem, therefore the question-mark in the field. The same is applicable for all the following tables and will therefore be omitted.

Table 2 lists challenges related to the service itself:

Challenge - Service	DARCS
not fit for purpose	DACS
not fit for use	DACS
not ready for production	DACS
does not pass validity tests	DACS

Table 2: Challenges - Service

All listed challenges can appear as a result of the execution of service validation and testing process during the service transition. One of the first things that should be checked next is - is the overall work defined well enough, and then secondarily - was is clear to all involved (including their assignments). In addition, if the testing process provided bad results, it should be clearly communicated to all involved. In order to prevent this, D and A should be checked, as well as goals re-assessed across the SMART principle.

Table 3 lists challenges related to the people involved in the service:

Challenges - People	DARCS
lack of expertize	DACS
lack of experience	DACS
insufficient number of team members	CS
changes in the team	CS
lack of motivation	CSA

Table 3: Challenges - People

Lack of the right combination and number of people that should be engaged with the service signals that it should be re-checked if all involved know their exact work and roles and to match it with existing expertize and engagement. If the challenge still exists, then it means that resources (from S) should be re-checked. Communication is here of vital importance, to help recognize that there is a gap, and to help prevent a situation that because of continuous lack of resources existing people burn out (faster spending of existing resources).

Table 4 contains a list of challenges related to procedures. Since the service operational procedures deal with different roles and responsibilities, it is not strange that R has important role in this area. Procedures non-existing, not applicable, too complicated or overlapping, point to the need to define and/or redefine the roles (R) and existing resources (people) (S) and to communicate (C) necessary changes and procedures to all involved. In addition, it is important to check if all know their assignments (A), to avoid misunderstandings.

Challenges - Procedures	DARCS
existing procedures not applicable in production	CRS
procedures too complicated	CS
additional procedures not coordinated with existing procedures	CR
too many procedures	SR
slow support caused by inappropriate procedures	ARC
old teams face difficulties accepting new procedures	ARC

Table 4: Challenges - Procedures

Challenges that might appear in the area of multi-domain service support are listed in the Table 5.

Challenges - Multi-domain	DARCS
resource provisioning is changing from one domain to another	ARCS
tasks performed by one domain are overlapping with another domain	DARC
a domain not aware of area of responsibility	DARC

Table 5: Challenges - Multi-domain

Three selected challenges require strong communication between domains. It has to include discussions about the exact area of work (D), exact assignments (A), the matrix of responsibility (R) that should be synchronous on both sides (or all sides if more of two domains are involved). The first example is directly related to the available resources and how they are determined. The last two can have any of the S aspects as a secondary after thought, after the DARC is analysed and resolved.

It can be seen that some of the techniques are better for some area of challenge sources, and some for the others. Some techniques such as Communication and SMART are applicable very often, especially in the phase of service transition. RACI should be engaged when there are more parties involved, D and A when one or more persons are included.

Additionally, even if some of the letters are not mentioned in the table, the resolution of one challenge might introduce the need to use other techniques, that were not used at the beginning.

For more complex and important services and systems DARCS can be applied in circles and each of the techniques re-assessed each time when the challenges and/or risks are recognized.

VII. Related work

In [13], Cusick explains production readiness reviews that are added to the architecture review in the transition phase.

Wolski et al. [14] presented a Service and Validation Process that includes a set of tests that are performed on a service candidate before the service goes into production. Only successful execution of all defined tests ensures that the service will transition directly. Partial success means that the service will stay with the developers until the minor or major are corrected.

Bovend'Eerdt et al. define in [15] a novel approach to set SMART goals by "using up to four parts: the target activity, the support needed, quantification of performance and the time period to achieve the desired state". In this paper, SMART goal setting is seen as a useful tool that enables all participants from multiple domains to share the same view on the goals of every process and every task related to the service operations.

VIII. Conclusion

This paper examines service offering from the perspective of the time when the service goes from the development in the production, and when there are multiple area of knowledge or influence. Service transition and especially service transition in a multi-domain environment can be very challenging. Five categories of challenge sources are recognized for a service transition in a multi-domain environment, and for each of them a few potential challenges and risks have been listed.

Based on the multiple sources of best practices, a set of five existing techniques is proposed to help avoid most of the challenges and mitigate their potential impact. However, none of it can help to eliminate the existence or appearance of any risk in the system.

Proposed approach has been applied on a listed example of challenges (and risks) with a proposal of the techniques that can help in a particular situation.

LITERATURE

[1] ITIL→ Best Practice, https://www.axelos.com/best-practice-solutions/itil/what-is-itil (available 02/2016)

[2] http://www.merriam-webster.com/dictionary/domain (available 02/2016)

[3] http://dictionary.reference.com/browse/domain (available 02/2016)

[4] http://ec.europa.eu/programmes/horizon2020/en/h2020-sections-projects

[5] ITIL→ Service Transition Book, ISBN 9780113313068

[6] www.geant.org

[7] https://vlada.gov.hr/UserDocsImages//Sjednice/Arhiva//74148.%20-%207.pdf

[8] https://vlada.gov.hr/UserDocsImages//Sjednice/2015/229%20sjednica%20Vlade//229%20-%2010%20a.pdf

[9] https://vlada.gov.hr/UserDocsImages//Sjednice/Arhiva//80.%20-%2016.pdf

[10] https://vlada.gov.hr/UserDocsImages//Sjednice/2015/235%20sjednica%20Vlade//235%20-%2016.pdf

[11] https://www.projectsmart.co.uk/raci-matrix.php

[12] https://www.projectsmart.co.uk/smart-goals.php

[13] J. Cusick, "Architecture and Production Readiness Reviews in Practice" http://arxiv.org/pdf/1305.2402.pdf (available 02/2016)

[14] M. Wolski, I. Golub, et al., "Service Validation and Testing Process", GÉANT Project Deliverable 8.1, 2016

[15] T. J. H. Bovend'Eerdt, R. E. Botell, D. T. Wade, "Writing SMART rehabilitation goals and achieving goal attainment scaling: a practical guide", *Clinical Rehabilitation*, Vol. 23, No. 4. (1 April 2009), pp. 352-361



Teaching "Ten Commandments" Of Software Engineering

Z. Putnik *, M. Ivanović *, Z. Budimac * and K. Bothe **
* Department of Mathematics and Informatics, Faculty of Science, University of Novi Sad, Serbia
{zoran.putnik; mirjana.ivanovic; zoran.budimac}@dmi.uns.ac.rs
** Institute of Informatics, Humboldt University Berlin, Germany
bothe@informatik.hu-berlin.de

Abstract - Besides learning smaller, individual issues of computer science like certain programming languages, algorithms, techniques, or methods, students are obliged to study and accept some more general subjects, meet introductory versions of problems they will encounter in professional life, and be introduced to matters of larger importance. In this paper, we will present how some of these "commandments" are delivered to students at our institutions, within two courses – one on the final year of undergraduate studies, and one on master studies (also electable for undergraduate students). Out of the larger set of universal, general issues, we selected ten that seemed more important to us, teaching our students proper approaches via collaborative, teamwork-solved assignments. Finally, we will describe the way in which eLearning helped us teaching such general matters properly.

I. INTRODUCTION

DAAD foundation (German Academic Exchange Service) under auspices of Stability Pact for South Eastern Europe since 2001 supports educational project "Software Engineering: Computer Science Education and Research Cooperation" [1]. The project comprises at the moment 17 universities from 10 countries. As the title shows, the original intention of the project was "...to cooperatively develop a joint web-based course in software engineering, the materials of which are now in use by several partner universities..." [2] The second main focus of the project was development of scientific cooperation between project members on all issues they have in common.

Over time, project surpassed original intentions by far. Besides "Software Engineering", several additional joint courses were developed and are in use at various institutions, project participants. Exchange of professors and students has been taking place each year very successfully. Joint meetings are held in autumn each year, where current situation within project is presented, experiences are exchanged, and interesting issues for future work and cooperation are developed.

As one of the bigger achievements, from 2004 to 2007 a TEMPUS project comprising core participants of the original DAAD project was created, with participation of several more European universities and individual experts [3]. Within that project, a whole set of around 20 courses was created for master studies in the field of "Software Engineering". Together with the teaching and assessment materials, additionally rules for the exam and studies, and other necessary resources have been developed. While the project lasted, there were even guest lectures held by professors from Germany, Spain, United Kingdom, and Romania. They visited and gave lectures at universities in Serbia and FYR of Macedonia, beneficiaries of the project. Exchange of students between these two countries also took place at the time, while exchange of professors still lasts, covered by local faculty funding, even after the discontinuation of the TEMPUS project.

Among other courses, two key courses that will be presented in this paper were developed within the "Software Engineering: Computer Science Education and Research Cooperation" project. The first is the course on "Software Engineering" for students of the final year of undergraduate studies, (originally fourth year, and later third year). This course is obligatory for the students of the "information science" direction, and elective for the students of "computer science" direction. The second one is course on "Software Project Management" for the students of the first year of master studies. It is obligatory course for the "Software Engineering" direction, while it is elective for all other directions of master studies of computer science, and also for students of the final year of undergraduate studies of the same direction.

Paper is further organized as follows: In the second section, we present opinions collected from articles dealing with the issues of teaching software engineering principles at the undergraduate level. The third section presents some of the comments and opinions we received within surveys that were conducted with the students involved in these two courses. The fourth section presents our approach in details, describe "ten commandments" or rules-of-thumb we are trying to teach, including comments on lessons learned while conducting our two courses. The next fifth section briefly enumerates some of the elements of eLearning that helped us conduct the whole process more efficiently. Finally, the sixth section concludes the paper and gives some final comments.

II. RELATED WORK

Since the beginning of teaching of software engineering, it has been recognized that this discipline is

This work was supported by Ministry of Education and Science of the Republic of Serbia within the project entitled 'Intelligent techniques and their integration into wide-spectrum decision support' (no. OI174023).

very important, final stage of studying computer science, concluding milestone that will finish ones education. Or, as nicely stated in [4] "…it is usually at least partly a project-based course, with the intention that student groups can deploy their already acquired skills on programming, verification, databases, and human-computer interaction, while applying the new material about requirements, architecture, and project management on a project." In the line of that statement, author also depicted his course on software engineering that uses rather similar methodology as we did namely teamwork, assignments, and attempts to gather all of the previously acquired knowledge.

A good example of teaching software engineering course can also be found in [5]. Authors of the course insisted on teaching software engineering using active learning, and in particular "…team-work, balance between theory and practice, imparting both technical and managerial skills encountered in real-world and problem-based learning." A survey conducted after the course showed that students were satisfied stating that "…the instruction model was new, interesting and had a positive impact on the motivation in addition to meeting the learning outcome of the course."

The fact that the teamwork, project-based learning, and practical employment of acquired skills are the basic element of teaching in the field of software engineering is well established, known from many books and articles. Still, let us mention paper [6] that together with detailed explanation of pedagogical aspects of teaching within a software engineering course, managed to summarize basic principles in four short words: "LEARNING = DOING + REFLECTING + FEEDBACK".

While we tried to teach basic principles of software engineering by giving our students several (ten, actually) different assignments, there are also attempts to illustrate all of them within one large project, conducted by a large teams of students, in a much longer time-frame. One example of such an approach is presented in [7]. As authors claim "…approach combines carefully coordinated lectures and homeworks, a hierarchical project management structure, modern communication technologies, and a web-based project tracking and individual assessment system." As a result, in their opinion, this approach "provides a more realistic project experience for the students, without incurring significant additional overhead for the instructor."

An interesting teaching approach using role-playing is presented in [8]. As authors say "…this paper has described teaching approach developed at Lancaster University that aims to address the teaching challenges posed by RDSE (Reuse-Driven Software Engineering) by combining role-playing with a realistic problem scenario and effective project support environment". While this approach and the whole organization are very tempting and potentially successful, they require a lot more time and personnel. Lucky for the authors, they have the whole undergraduate software engineering degree program, and three years to conduct it!

Finally, let us mention a paper [9] that probably doesn't present any particularly novel approach to teaching, but nicely summarizes the whole situation with teaching of software engineering. Claiming that software engineering *is* a key concept in computing education, authors remind the readership that "…many schools offer a project-oriented course, or multi-course sequence, to teach students both the theoretical concepts of software development as well as the practical aspects of developing software systems in a team environment." But, the point interesting for this paper is that authors recognize rather similar points as the most important ones, such as: principles of requirements analysis, project management, configuration management, use of collaboration tools, and tools for team communication, needed for a development methodology, and effective teamwork.

III. Student's Opinion About The Courses

The main intention of two courses we are using here as examples of teaching basics of software engineering is to summarize all of the knowledge students achieved over the years of their studies. These ambitions seem to be well accomplished, since in all of the surveys conducted each year in order to collect student's opinion about courses, the same facts are repeated. For the purpose of reporting at the annual joint project meetings the answers in these surveys are well analyzed on several occasions, and some conclusions of our students are repeated each year for the course "Software Engineering":

- there is appropriate quantity of knowledge within the teaching resources;
- lecture content is appropriate and in accordance with the previously acquired knowledge;
- knowledge and engagement of lecturers is excellent;
- Enough new useful things are also adopted.

On a more "personal" level, students answered within the commenting parts of the survey:

- "Very interesting course. Different from other courses, and it should be obligatory for all CS students";
- "Good student-teacher relation. Students can be active and participate in discussions";
- But also "I learned the light and dark side of working in teams".

Finally, at some point there was an anonymous comment that recapitulates the general opinion. What is the most important – achieving of exactly that reaction was the intention of both courses from the beginning: "...I'm generally very satisfied with the course. It taught me a lot of things that were previously unknown to me, but now everything makes sense (on a global level). I understand its significance and importance. Also I've become very attached to it, and finally started to see my path and place in the future."

For the second course, "Software Project Management", we have an additional interesting situation. Students who did not elect this course originally, together with the students whose study direction does not offer

selection of this course, came back to attend the course, sometimes even several years later. Even though they do not gain any ECTS points, grades, or anything, they are back, just for the knowledge! Our inquiry showed that since the large percentage of master students are employed, they claimed that "...they were told by their colleagues who attended the course that it is very useful in practical work, since it gives a general overview and feeling about large (software) projects..."

Some numerical and basic statistical analysis of student achievements with the results accomplished with the assignments/commandments was performed, but these go out of the scope of this paper. For interested readers, parallel analysis of results from several institutions, project participants can be found in [2].

IV. Selected "Ten Commandments" of Software Engineering

In order to sum-up things for our students, two courses were organized in such a way to imitate real working environments as much as it is possible in educational circumstances. Students are given several assignments, complex and comprehensive, and appropriate, relatively long time to solve those. Each of the assignments was agreed upon by the project participants, and more-or-less the same structure of the course exists in all of the countries where it is conducted. Theory that stands behind each assignment is presented at theoretical classes and if necessary further discussed within classical classroom exercises. More interesting for this paper is the fact that each of these assignments students have to solve within a fixed time frame as a team of usually 4 to 5 persons – just as they will work in the future professional life.

A. *The first commandment: "Make sure you know what you should do!"*

At the master level, there is a whole course dealing with the important matters and understanding of "requirements specification and engineering". Still, the intention was to introduce students to this point at the undergraduate level also, and show them possible problems with ambiguity, misunderstandings, or errors they can encounter while negotiating with the clients.

Consequently, for the first assignments, students are given requirements specification of an imaginary software system that should be developed. It consists of all of the usual elements of specification, including: goals of the system, ways of product usage, overview of the system in a form of use-case diagram, list of needed product functions in the system described through use-cases, description of product data, questions of needed efficiency and other quality requirements, elements of user-interface, non-functional and technical environment, and so on.

Being a large document (at least to inexperienced students) of 13 pages, requirements specification contains also a lot of uncertainties, impreciseness, and of course errors. The task of the students is to read the document carefully, clear-up in their minds what they should do, what functions are required, are there any contradictions that might lead to problems later during the software development and similar. They are supposed to fix the document, suggest corrections, recognize faults and similar. To help them, but more importantly to teach them to work collaboratively, we divided our students into teams of 4-5 members, and gave them two weeks to solve the assignment. In the beginning, teams were created by self-selection, yet after few years, we introduced some psychological tests, and automate team creation process, as presented in [10].

Lessons learned: This is a very difficult assignment for our students! Even after two weeks they had for solving the task, a lot of the things were left unnoticed. To help them, we organize one class where "the most intriguing" solution is presented by the authors, while the rest of the class, together with the lecturer, challenges their opinions and attitudes. Here, "intriguing" doesn't necessarily means the best, or the worst solution. It is the solution that we think will provoke the most discussion, fairly correct of course, the one that recognized the most of the things other solutions didn't, even if their comments on those things are not proper sometimes.

B. *The second commandment: "Learn the algorithm, and follow it!"*

As for any other profession, craft part of software production is something that should be followed to the letter! When some routine task is to be performed, there should be no time waste on inventions and creativity. When you have to for example sort a sequence of numbers, simply apply the most appropriate sorting algorithm, and do the job, do not (over)think!

Still – the task of sorting is not challenging enough. We had to think of something more difficult, and by far more global. So, in line with the first assignment, we first taught our students "Function point method" for the cost estimation of a future software system, even showed them example on how to apply it with the preliminary requirements specification, and then asked them to perform the same method on the final requirements specification they read and corrected in the first assignment. Again, students had two weeks to work in their team and to solve the assignment.

Lessons learned: This was rather inconvenient assignment for our students. It looked very easy and straightforward, especially after so ambiguous and "confusing" first assignment. We were rather tolerant and accepted very wide range of final results as correct, as long as they had a proper explanation. (Results varied from 170, up to 330 function points, and with the proper justification, each one was accepted.) Still, improper application of the algorithm was often obvious, input functions were declared as output, queries were calculated in cost estimation even when they should not be, and all other sorts of errors were observed. This assignment is with the lowest grade average among all of the assignments, as our analysis shows [11].

C. *The third commandment: "Do not trust (authorities) implicitly, check for errors!"*

It happens sometimes that there is a need to use some piece of software, or model/project/scheme/pattern that

some other member of the software production team created for current, or some previous software systems. It also happens that afterwards, there are some errors found in the final version of the software. Was it us who made the error, or was that error already built-in?

The third assignment is mostly the visual one. From the book that is used as the bases for the whole course [14], a set of data-flow diagrams was selected, depicting communication between program subsystems. Even though the quality of the book is high, some errors (and not only a few) can be found within those data-flow diagrams. Students were supposed to check the diagrams, find the errors, find inconsistencies between diagrams of higher and lower order, and similar.

Lessons learned: Different results for this assignment were achieved with different students. Some of them were pretty sure that it is impossible that there are more than just a few errors, so after finding those, they stopped looking. On the other hand there were those who declared almost everything the error! Yet, the most of the students at this moment were able to recognize actual mistakes without involvement of much trust or mistrust, just by their honest, pure and straight logic, and of course theoretical pre-knowledge we gave them at classes.

D. *The fourth commandment: "Think about the problem, use your previous knowledge, and find the best solution!"*

Opposing the second assignment, there are situations that ask for and require creative thinking. There are situations when software engineer has to apply not just the knowledge gained within one subject, course, or topic, but the sum of knowledge collected over the courses and years.

Naturally, this time we were not able to use the same task, there was a need for something new and challenging. So, we first explained in details one real-life situation from students' life, and then ask our students to become creative and try to apply gained knowledge. They had to create use-cases, use-case diagram, class diagram (using UML notation learned within some other course), and define all of needed operations and attributes for the future possible software application.

Lessons learned: This we hoped and expected to be nice, inspired, and imaginative assignment. It wasn't – either our students do not know enough, or they simply do not care for a higher grade, and are satisfied with any positive result, as long as they didn't have to invest too much effort into their work. So, results are fair, but far from creative. Compared to other assignments, point-wise, this is one of those with the lower grade on the average.

E. *The fifth commandment: "Honour theory and learn to follow through (mathematical) proofs!"*

As one of the more difficult and un-liked part of software engineering education, at least with our students, there is a need to study mathematics and formal methods. We decided to introduce them to: algebraic specification, model based Z specification, and Hoare logic.

Within the lecture, students learn about mentioned formal techniques, and see examples of their use. In order not to make it too difficult to students, we gave the assignment of creating formal specification of some other similar system using algebraic and Z specification and assignment of extension/modification of an existing Z specification, by adding a specific set of needed operations.

Lessons learned: Contrary to our expectations – which of course might mean that we have wrong expectations – this assignment proved to be difficult, yet straightforward in the opinion of our students. Having enough time and all of the necessary theory at their disposal, students on the average manage to solve the larger part of this assignment correctly and achieve quite large percentage of points on the average.

F. *The sixth commandment: "Recognize knowledge of your colleagues and either teach them, or learn from them!"*

Since it is certain that in real life our students will have to cooperate and collaborate with other individuals, teams, even companies, it was decided to start their education in this area as early as possible. Grades for previously mentioned "creative" assignment number four were not distributed to students immediately. Instead, each of the teams was given the solution for this task of some other team, and the assignment to analyze it, criticize it, suggest improvements, recognize errors and similar.

To make the assignment more challenging, the most opposite solutions were always exchanged. This way, teams were faced with very different view on the same problem and with quite different solution. We hoped that this will provoke more discussion and better explanations.

Lessons learned: Rather interesting results were received for this assignment. Teams defended their solution with a lot of eagerness, criticized the opposite opinion strongly and with detailed explanations. This actually presented a problem with grading for us. IF a team has a firm and strong opinion on something and is ready to defend it with reason and logic, and yet that opinion is incorrect, what should be done? As much as correct solution is wanted, we would also like to teach students to stand behind their opinions.

In addition, let us mention that with this assignment, there was yet another interesting situation noticeable. While students (as students commonly do) complained about their grades and points for previous assignments, as a rule, grades that they gave to other team within this assignment were almost always lower than what we gave. At additional dedicated class, we presented all of the teams with both grades asking them which one they want. No need to mention, they wanted the higher grade.

G. *The seventh commandment: "Be prepared to manage and use unknown tools!"*

Another common situation in a life of software engineer is the one in which (s)he encounters some unknown tool(s), needed for quick solving of some emergency situation. To get students ready for such thing,

after receiving lecture on "software metrics", students are faced with a set of Java methods and a request to measure their quality. As tools, they are given "Eclipse" programming environment - which they already know - and metrics plug-in, which they don't know. First, they have to install it, learn how to use it, and finally to apply it on a given set of methods.

Lessons learned: With this assignment, we perhaps underestimated students. Installation, learning and usage went smoothly, without any problems. Unfortunately, that enabled us to notice another problem. Part of the assignment was strict (do this and that), while the other part was more relaxed (... and report any other point you find interesting). The most of teams does just obligatory part, not caring for additional points for research.

H. *The eight commandment: "Trust your colleagues, but TEST their work."*

Since in general opinion testing is the most expensive and time consuming phase of software development, thus making it one of the most important phases, two of the assignments are dedicated to this phase.

Eighth task for our students was preceded by lecture on testing in general, and after that lectures about two most common types of testing: structured and functional testing. As an excellent method to actually perform functional testing in practice, students are introduced to "classification tree method" first theoretically, and after that practically, through presentation of "Classification Tree Editor". After that, students are given a program that needs testing, and are required to implement mentioned methodology and perform functional testing. This assumed creation of classes and classifications, selection of actual test-cases, and performing of the testing.

In addition, this task again required students to download, install, register for, and use an unknown tool, namely CTE XL tool.

Lessons learned: Quite nicely accepted task, with very good results and discussions of these results. No surprise, since classification tree method was even at classes accepted by students as an "interesting game".

I. *The ninth commandment: "Save time and money, test completely, yet minimally."*

Having at our disposal some of the locally (Humboldt University of Berlin) created testing tools we were able to involve students with some more practical work and some more unknown tools, applied again for one of the most important phases of software development, testing phase. Students are first introduced to theory of testing from another angle, teaching them about various kinds of testing-coverage measures (branch coverage, statements coverage...). After that, students hear about the testing tool called SOTA [13]. Finally, students are given a program that needs testing, and their assignment was to find minimal set of test-cases that will guarantee various types of coverage for a given program.

Lessons learned: At this moment, within a course, students are already experienced and, with the interesting topic such as testing, usually highly motivated to achieve good results. Consequently, this task is largely appreciated and welcomed, results are usually excellent, and detailed explanations following the submitted results are in the most cases highly professional! This was one of the best accepted assignments.

J. *The tenth commandment: "Get a complete picture of a project!*

The final commandment asked for the larger assignment, so this one is performed within another course, "Software Project Management". After teaching about planning techniques and methods, we gave a two hour lecture on practice of "organizing a scientific conference". Besides, two documents are available to students: written informal explanation of one common method for conference organization, and formal, use-case based requirements specification, for a project of planning the conference organization.

One of many project planning tools had to be selected, and in order not to burden students with too much studying of the tool itself, we selected Microsoft Project as a tool very similar to other Microsoft Office tools, already known to our students. Being a large task, we gave students unlimited time for solving it, with the only condition that it must be solved before approaching the final exam. On the hard side, we asked students to solve this assignment individually!

Lessons learned: Being a pre-requisite for the exam, this assignment was solved by all students, more or less successfully. Large majority of students managed to achieve more than 50% of the points, which was required, while sufficient number of "better" students managed to achieve more than 80% of the points, which was considered excellent. One more time, it seemed that the greater problem was students wish to invest their time and knowledge, their satisfaction with the (minimal) passing grade, than actual knowledge needed.

V. USE OF ELEARNING

Let us just briefly mention here that the whole effort in teaching mentioned "ten commandments" was largely relaxed due to our use of eLearning facilities.

At our Department, LMS Moodle, the most popular and the most often used open-source eLearning system [14] is successfully employed in everyday teaching for about twelve years now. While in the beginning it was just a simple repository of teaching resources, over the years it grew into an integral part of our teaching.

Considering the issues discussed in this paper, two particular elements of LMS Moodle were of large importance. Namely, in order to be able to differentiate between individual students' efforts, we required from students to use forums and wiki service within Moodle to solve their assignments. While more on our experiences can be found in [15] and [16], let us just mention here some of the most important issues:

- Within forums, students were allowed to ask questions, and exchange and share experiences

and opinions. Exchange of solutions was of course *not* allowed;

- Wikis were defined in such a way that only a members of a team were able to approach their solution, so plagiarism and copying of other teams' solution was not possible;
- Using "history" feature of wiki service, lecturers were able to recognize individual contributions of students to the final result and thus to commit fair and correct grading.

VI. Conclusion

In this paper, we presented our methodology and experiences with teaching courses on "Software Engineering" and "Software Project Management". Both courses are created within a large and long lasting educational project, and are as such conducted at various countries, by various lecturers.

Created and enhanced with experiences of a large number of professors and assistants, these courses proved their value in several ways. They were successful at undergraduate and master level, at different universities, with the same set of teaching resources. On the other hand, courses were also successful as a two-semester, one-semester, and even one-week crash-course [2], [17], [18].

In particular, we presented one aspect of our teaching, or as we named it "ten commandments of software engineering". These are taught through assignments, usually solved by collaborative teamwork. The set of assignments selected is not a unique set definitely. The project participants are allowed to select those assignments that best fit their teaching methods, which we tried to use to our advantage. Used set of assignments can be shortened, or expanded by few tasks, depending on available time, quality of students, or style of a lecturer. Here we presented the set used in Novi Sad, but the point is that teaching of some common software engineering rules-of-thumb is a must, and what better ways to do it, than involve students in actual, practical work with these things. By our experience that emerged while selecting, changing, and enhancing assignments for students, after finishing about half of mentioned tasks, students started to enjoy their work, learned what they have to do, and become eager for their next challenge.

The whole organization requires some additional work performed by lecturers, and asks for extra effort carried out by students, that is certain. Still, benefits that can be recognized with this kind of organization of case-study and project based approach, with learning conducted through assignments solved by team endeavors, and special classes organized for reflection on the issues and problems encountered is worth it. Results achieved on both assignments and theoretical tests conducted over the years, proved that students learned a lot. Results of a survey conducted with students, proved that they were satisfied with the methodology and became more interested in the field of software engineering.

References

[1] "Joint Course on Software Engineering" (JCSE) Project home-page, http://www2.informatik.hu-berlin.de/swt/intkoop/daad/, Accessed on 20. February 2016.

[2] Budimac, Z., Putnik, Z., Ivanović, M., Bothe, K., Zdravkova, K. and Jakimovski, B. "Multi-Country Experience in Delivering a Joint Course on Software Engineering - Numerical Results", The International Review of Research in Open and Distance Learning, Vol. 15. No. 1, (pp. 84-110), 2014.

[3] Joint Master Studies in Software Engineering, TEMPUS CD-JEP 18035-2003 Project home-page, http://perun.dmi.rs/msc-se/, Accessed on 20. February 2016.

[4] Flener, P. "Realism in Project-Based Software Engineering Courses: Rewards, Risks, and Recommendation", Chapter in Computer and Information Sciences – ISCIS 2006, Volume 4263 of the series Lecture Notes in Computer Science, pp 1031-1039, DOI 10.1007/11902140_107, 2006.

[5] Sureka, A., Gupta, M., Sarkar, D. and Chaudhary, V.. "A Case-Study on Teaching Undergraduate-Level Software Engineering Course Using Inverted-Classroom, Large-Group, Real-Client and Studio-Based Instruction Model", CoRR abs/1309.0714, 2013.

[6] Razmov, V. "Effective pedagogical principles and practices in teaching software engineering through projects", Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, FIE '07. 37th Annual, pp. S4E-21-S4E-26, DOI: 10.1109/FIE.2007.4418158, 2007.

[7] Coppit, D. "Implementing large projects in software engineering courses", Computer Science Education, 16:1, 53-73, DOI: 10.1080/08993400600600443, 2006.

[8] Kotonya, G. and Lee, J. "Teaching reuse-driven software engineering through innovative role playing". In Companion Proceedings of the 36th International Conference on Software Engineering (ICSE Companion 2014). ACM, NY, USA, pp. 276-282. DOI: 10.1145/2591062.2591166, 2014.

[9] Teel, S., Schweitzer, D., Fulton, S. "Teaching Undergraduate Software Engineering Using Open Source Development Tools", Issues in Informing Science and Information Technology (IISIT), Vol. 9, pp. 63-73, 2012.

[10] Komlenov, Z., Budimac, Z., Ivanovic, M., Bothe, K. "Team formation process automatization", In Proc. of EDUCON, IEEE Global Engineering Education Conference, DOI: 10.1109/EduCon.2013.6530190, 2013.

[11] Budimac, Z., Putnik, Z., Ivanović, M., Bothe, K., Schuetzler, K. "On the Assessment and Self-assessment in a Students Teamwork Based Course on Software Engineering" Computer Applications in Engineering Education, ISSN: 1099-0542, Volume 19, Issue 1, pp 1–9, DOI: 10.1002/cae.20249, 2011.

[12] Balzert, H., Lehrbuch der Software-Technik, Band 1 und 2, Heidelberg: Spektrum Akademischer Verlag. 1998.

[13] https://www2.informatik.hu-berlin.de/swt/lehre/TestTools/sota/, Accessed on 20. February 2016.

[14] https://moodle.org/, Accessed on 1.April 2016.

[15] Putnik, Z., Budimac, Z., Ivanović, M., Bothe, K. "Analysis of Students' Behaviour Based on Participation and Results Achieved in a Wiki-based Team Assignments", In ACM Intl. Conf. Proc. Series, Proc. of the 6th Balkan Conference in Informatics, Greece, pp. 179-186, DOI: 10.1145/2490257.2490277, 2013.

[16] Putnik, Z., Ivanović, M., Budimac, Z., Samuelis L. "Wiki - A Useful Tool to Fight Classroom Cheating?", Advances in Web-Based Learning - ICWL 2012, Lecture Notes in Computer Science, Volume 7558/2012, pp. 31-40, DOI: 10.1007/978-3-642-33642-3_4, 2012.

[17] Budimac, Z., Putnik, Z., Ivanović, M., Bothe, K. "A View on a Successful International Educational Project in Software Engineering", e-Informatica Software Engineering Journal, Vol. 6, Issue 1, pp. 47-59, DOI: 10.5277/e-Inf120104, 2012.

[18] Budimac, Z., Putnik, Z., Ivanović, M., Bothe, K. "Transnational Cooperation in Higher Education in Balkan Countries", Novi Sad Journal of Mathematics, Vol. 43, No. 1, pp. 167-177, 2013.

Methodologies for Development of Mobile Applications

Z. Stapić, M. Mijač and V. Strahonja
University of Zagreb, Faculty of Organization and Informatics, Varaždin, Croatia
zstapic@foi.hr, mmijac@foi.hr, vstrahon@foi.hr

Abstract – The paper presents the results of literature review performed in order to identify newly created methodologies targeting the development of mobile applications. As mobile development differs from development of other software systems we also discuss the challenges of such development and recon that methodological approach in development is necessary. However, there were no systematic studies identifying methodologies for mobile applications development. Systematic literature review we performed included more than 6700 initial studies, identified 49 (0.73%) potentially relevant studies, and after extracting data and applying quality check procedures 14 methodologies are identified. Only one of these methodologies is reported to be used in practice.

I. INTRODUCTION

In accordance with the current state-of-the-art stream, the development of mobile applications and systems should satisfy special requirements and constraints and thus differs from traditional software development in many aspects. Back in 2008, Rahimian and Ramsin [1], identified portability issues, various standards, protocols and network technologies, specific capabilities of devices and strict time-to-market as specific requirements and constraints which are present, in domain of mobile development, today as well. Even before, in 2005 Hosbond [2] recognized a high level of uncertainty in mobile applications development as a result of two set of challenges, namely *business* related challenges (e.g. tough competition, conflicting customer interests, establishment of revenue-share models etc.) and *development* specific challenges (e.g. rapidly changing technology, lack of standardization, integration with existing systems etc.). Apart from these challenges, there are other specifics in mobile development that are worth noting, but are out of scope of this paper. Some of them are the use of platform specific and dependent but not interoperable development environments [3], fragmentation problem [4], [3], [5], [6], testing [6], number of target platforms [5], innovation over standardization [2] et cetera.

This made scholars and practitioners try to find a systematized way in mobile applications development. We agree with authors saying that first issue that should be addressed in development of mobile applications is the usage of methodology [1], [7], [8]. However, soon, it is concluded that classic or agile software development methodologies should be adapted for the development of mobile applications as the existing ones do not cover the specific mobile targeted requirements [8]. Even more, in our undergoing research we went one step further with the hypothesis that methodological approach is a necessary but not sufficient and that *methodological interoperability* [9] supported by *ontologically described semantics* [4] could help when development teams are targeting two or more target platforms.

Thus, as the use of methodologies is recognized in theory as starting point in dealing with stated challenges, issues and specifics, and as existing methodologies do not provide enough merits for use in this field, in this paper we present our research and results on newly created development methodologies targeting mobile applications development. Second chapter of the paper specifies basic concepts that are used in our research and then the research method is explained. In fourth chapter we bring main results and in final chapter we conclude the topic and bring forward possible directions of future research.

II. CONCEPTS PUZZLE

Systems development methodologies (SDM) are of an academic interest since the early 1980s when the IFIP WG8.1 (Working group of the International Federation for Information Processing on Design and Evaluation of Information Systems) organized three conferences named Comparative Review of Information Systems Design Methodologies (CRIS). The first conference [10] aimed to present and compare spectrum of methodologies, while second [11] and third [12] conferences had goal to analyze and to evaluate methodologies respectively. These conferences also resulted in the definition and distinction of basic concepts and terms like methodology, method, tool, approach, and development cycle. However, the used concept of "methodology" was limited only to the design stage of the system development life-cycle [13]. Since these origins, different definitions for the term "software development methodology" which cover full development life-cycle are created. Some of them are available in [14], [15], [16], [17] and in [18].

Consequently, today SDM could be observed as a noun and as a verb. As a noun, "software development methodology is a **framework** that is used to structure, plan, and control the process of developing an information system" – this includes the pre-definition of specific deliverables and artifacts that are created and completed

This research was partially sponsored by University of Zagreb, Program of Support in Scientific and Artistic Research in 2015.

by a project team to develop or maintain an application [19]. As a verb, the software development methodology could be considered as an **approach** used by organizations and project teams to apply the software development methodology framework. Every software development methodology approach acts as a basis for applying specific frameworks to develop and maintain software. As software development approaches do not define all aspects of development process but rather give general direction, they are not in the focus of this paper. Thus, term *methodology* in this work will annotate full development framework, while term *approach* will be used to denote approach as a verb. In general, adopting the definition from Avison and Fitzgerald [20] in this paper, methodology will be considered as "*a collection of procedures, techniques, tools and documentation aids which will help the systems developers in their efforts to implement a new information system.*"

III. METHOD

In order to identify existing body of knowledge on our research topic we tried to find existing systematic literature reviews on software development methodologies for development of mobile applications. Thus we performed exactly the same method, **Systematic Literature Review**, as it is defined by Kitchenham and Charters in [21] or by Stapić et al. in [22].

The IEEExplore, ACM Digital library, INSPEC, CiteSeerX and GoogleScholar databases were searched by the following search query: ("literature review" OR SLR) AND (mobile development). This query implicitly includes "systematic literature review" phrase. Additionally, more rigorous search queries, like ("literature review" OR SLR) AND (mobile development methodologies) or similar have been discarded as returning only a few or no results.

The search returned following number of papers: IEEExplore (61), ACM Digital library (624), INSPEC (62), CiteSeerX (22) and GoogleScholar (128). Additionally, the original query on GoogleScholar returned more than 22.300 results, so there was used a narrower concept searching for „mobile development" as a phrase instead of searching for both words independently as in other databases.

Almost all obtained papers were excluded as not being literature reviews or not being literature reviews in mobile applications development. Only one paper [23] passed the inclusion criteria, but the focus of the SLR performed in this paper was to review the literature in the domain of four mobile systems development perspectives (requirements, technology, application, business) but unfortunately did not include methodologies or approaches to be used when developing mobile applications.

Thus, according to information available in the mentioned databases, there are no existing systematic literature reviews covering the subject of software development methodologies for mobile applications development, which brings us to the conclusion that such systematic literature review is highly necessary. As an additional proof of this claim, the results of SLRs on *Systematic Literature Reviews in Software Engineering* presented in [24] and in [25] show that no literature reviews were conducted in the domain of software development methodologies or software development methodologies for mobile devices.

By taking into consideration all aspects of new systematic literature review that will identify newly created methodologies for mobile applications development, the newly defined search query was much more complex: "*(mobile AND ("software development" OR "system development" OR "application development" OR "program development") AND (methodology OR method OR approach OR framework OR process OR procedure OR model))*. However, it was not possible to automatically exclude the use of existing methods or the use of adapted methods in mobile applications development. While the execution of this search string returned 6761 initial studies. After applying inclusion and exclusion criteria through several phases, the final list of relevant studies included 49 relevant studies (0.73%).

IV. RESULTS

The data extraction procedure are applied on identified 49 studies and are followed by the quality check procedures. The final result of identified newly created methodologies for mobile applications development included 14 methodologies as presented in Table 1.

TABLE I. - IDENTIFIED METHODOLOGIES

Name	Study	QA score
Agile Methodology for Mobile Software Development	[1]	3.0
Agile Solo	[26]	2.0
Agile usability process	[27]	2.0
DEAL	[28][29]	2.5
Integrated Product Development Process for Mobile Software	[30]	2.0
Inter-combined Model	[31]	3.0
MASAM methodology	[32]	2.5
Methodology for Building Enterprise-Wide Mobile Applications	[33]	4.0
MicroApp visual approach	[34]	2.5
Mobile Application Development Methodology	[35]	1.5
Mobile-D	[36]	2.5
	[37]	1.0
New media application prototyping	[38]	3.0
Systems Development Methodology	[39]	4.0
ViP (Virtual Platform)	[40]	4.0

However, we filtered these results through final filter to identify those methodologies that are reported to be used in practice and unfortunately identified only one method that is reported to be used in practice: **Mobile-D from Abrahamsson et al.** [36], [37].

A. Mobile-D

Mobile-D methodology was firstly introduced by Abrahamsson in 2004 [36], and it was composed of five iterations i.e. phases: *set-up, core, core2, stabilize* and *wrap-up*. According to technical documents available on

TABLE II. - MOBILE-D PHASES, ACTIVITIES AND TASKS

Mobile-D Phases	Development days / Activities	Tasks
Explore	Stakeholder establishment	Customer establishment
		Stakeholder group establishment
	Scope definition	Initial requirements collection
		Initial project planning
	Project establishment	Environment selection
		Personnel allocation
		Architecture line definition
		Process establishment
Initialize	Project set-up	Environment setup
		Training
		Customer communication establishment
	Planning day in 0 iteration	Architecture line planning
		Initial requirements analysis
	Working day in 0 iteration	
	Release day in 0 iteration	
Productionize	Planning day	Post-iteration workshop
		Requirements analysis
		Iteration planning
		Acceptance test generation
		Acceptance test review
	Working day	Wrap-up
		Test-driven development
		Pair programming
		Continuous integration
		Refactoring
		Inform customer
	Release day	System integration
		Pre-release testing
		Acceptance testing
		Release ceremonies
Stabilize	Planning day	Post-iteration workshop
		Requirements analysis
		Iteration planning
		Acceptance test generation
		Acceptance test review
	Working day	Wrap-up
		Test-driven development
		Pair programming
		Continuous integration
		Refactoring
		Inform Customer
	Documentation wrap-up	
	Release day	System integration
		Pre-release testing
		Acceptance testing
		Release ceremonies
System test & fix	System test	System test
	Planning day	Post-iteration workshop
		Requirements analysis
		Iteration planning
		Acceptance test generation
		Acceptance test review
	Working day	Wrap-up
		Test-driven development
		Pair programming
		Continuous integration
		Refactoring
		Inform customer
	Release day	System integration
		Pre-release testing
		Acceptance testing
		Release ceremonies

the authors' web site, for example [41], the methodology included 34 principal inputs and outputs (like *action point list, architecture line plan, base process description, daily status report* etc.) and 9 different roles (like *customer group, exploration team, project team, steering group*, etc.). The method evolved and according to presently available documents such as web application presenting the methodology [42] and set of documents and templates describing the methodology in detail [43] the main phases, activities and tasks are presented in Table 2.

The practices included in execution of tasks during different phases and activities comprise nine principal elements which are mainly well-known agile practices specialized for mobile software development [36], [44]:

- Phasing and pacing – The projects are performed in iterations of which each begins with a Planning Day
- Architecture Line – Architecture line approach is utilized together with architectural patterns and Agile Modeling
- Mobile Test Driven Development – Test-first approach is utilized together with automated test cases
- Continuous Integration – Effective Software Change Management (SCM) practices are applied through multiple means
- Pair Programming – Coding, testing and refactoring are carried out in pairs
- Metrics – Few essential metrics are collected rigorously and utilized for feedback and process improvement purposes
- Agile Software Process Improvement – Post-Iteration workshops are used to continuously improve the development process
- Off-Site Customer – Customer participates in Planning and Release Days
- User-Centered Focus – Emphasis is placed on identifying and fulfilling end-user needs

Mobile-D is based on combination of Extreme programming in terms of practices, Crystal family of methodologies in terms of scalability and Rational Unified Process in terms of life-cycle coverage [45]. Additionally, the analysis of the methodology showed that Mobile-D is leaning on, and is strongly connected with, Test Driven Development approach.

V. CONCLUSION

Despite the mentioned problems in mobile application development that could be interesting for the scientific community, a relatively few researches aimed to enhance the methodologies for mobile application development, and most of the work performed in this field has been focused on the implementation-oriented aspects of the mobile software development, while methodology-oriented issues still remain to be properly addressed [1].

Reviewing the existing solutions for mobile application development, we identified Mobile-D methodology from Abrahamsson et al. [36] as an agile approach to mobile application development as the only one methodology reported to be used in practice. The obtained papers and other documents that include detailed

guidelines are sufficient to make this methodology fully applicable and usable.

Finally, it is important to mention that this research had some limitations, primary in accessing all identified sources which were not available for scientific community in Croatia nor in Spain. We also find it important to perform more detailed analysis on extracted data in order to identify development approaches that are used as well as classic or agile methodologies that are adapted or adopted for mobile applications development. The specific research topic might be the analysis focused on methodologies that are suitable for multiplatform mobile applications development.

ACKNOWLEDGMENT

The research presented in this paper was conducted as a part of a PhD research and is partially founded by the scholarship granted to the first author from Croatian Science Foundation and from Support program granted by University of Zagreb.

REFERENCES

[1] V. Rahimian and R. Ramsin, "Designing an agile methodology for mobile software development: A hybrid method engineering approach" in *Proceedings of Second International Conference on Research Challenges in Information Science, RCIS (2008)*, Marrakech, 2008, pp. 337–342.

[2] J. H. Hosbond, "Mobile Systems Development: Challenges, Implications and Issues" in *Mobile Information Systems II*, vol. 191, J. Krogstie, K. Kautz, and D. Allen, Eds. Springer Boston, 2005, pp. 279–286.

[3] V. Agarwal, S. Goyal, S. Mittal, and S. Mukherjea, "MobiVine: a middleware layer to handle fragmentation of platform interfaces for mobile applications" in *Proceedings of the 10th ACM/IFIP/USENIX International Conference on Middleware*, New York, NY, USA, 2009, pp. 24:1–24:10.

[4] Z. Stapić, "Dealing with Mobile Platforms Fragmentation Problem: Ontology Oriented Approach" *Res. Pap. Fac. Mater. Sci. Technol. Slovak Univ. Technol.*, vol. 21, no. Special Issue, pp. 98–103, 2013.

[5] A. Manjunatha, A. Ranabahu, A. Sheth, and K. Thirunarayan, "Power of Clouds in Your Pocket: An Efficient Approach for Cloud Mobile Hybrid Application Development" in *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, 2010, pp. 496–503.

[6] Y. Ridene, N. Belloir, F. Barbier, and N. Couture, "A DSML For Mobile Phone Applications Testing" in *Proceedings of 10th Workshop on Domain-Specific Modeling in SPLASH*, France, 2010.

[7] A. C. Spataru, "Agile Development Methods for Mobile Applications" PhD Thesis, University of Edinburgh, The University of Edinburgh, Edinburgh, 2010.

[8] H. J. La and S. D. Kim, "A service-based approach to developing Android Mobile Internet Device (MID) applications" *2009 IEEE Int. Conf. Serv.-Oriented Comput. Appl. SOCA*, vol. 00, no. MID, pp. 1–7, 2009.

[9] Z. Stapić, L. de Marcos Ortega, and J. M. Gutiérrez Martínez, "Approaches in Development of Multi-platform Mobile Applications: State of the Art" in *Actas del IV Congreso Internacional sobre Aplicación de Tecnologías de la Información y Comunicaciones Avanzadas (ATICA 2012)*, Loja, Ecuador, 2012, pp. 429–436.

[10] T. W. Olle, H. G. Sol, and A. A. Verrijn Stuart, Eds., *Information systems design methodologies: a comparative review: proceedings of the IFIP WG 8.1 Working Conference on Comparative Review of Information Systems Design Methodologies, Noordwijkerhout, The Netherlands, 10-14 May 1982*. Amsterdam˜ ; New York˜ : New York, N.Y: North-Holland Pub. Co.˜ ; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co, 1982.

[11] T. W. Olle, H. G. Sol, and C. J. Tully, Eds., *Information systems design methodologies: a feature analysis: Proceedings of the IFIP WG 8.1 Working Conference on Feature Analysis of Information Systems Design Methodologies, York, U.K., 5-7 July, 1983*. Amsterdam˜ ; New York˜: New York, N.Y: North-Holland˜ ; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co, 1983.

[12] T. W. Olle, H. G. Sol, and A. A. Verrijn Stuart, Eds., *Information systems design methodologies: improving the practice: proceedings of the IFIP WG 8.1 Working Conference on Comparative Review of Information Systems Design Methodologies, Improving the Practice, Noordwijkerhout, The Netherlands, 5-7 May, 1986*, Post-conference ed. Amsterdam˜ ; New York˜ : New York, N.Y: North-Holland Pub. Co.˜; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co, 1986.

[13] S. Gasson, "The role of methodologies in IT-related organisational change" in *Proceedings of BCS Specialist Group on IS Methodologies, 3rd Annual Conference, The Application of Methodologies in Industrial and Business Change*, Wrexham, 1995.

[14] S. Dyck and T. A. Majchrzak, "Identifying Common Characteristics in Fundamental, Integrated, and Agile Software Development Methodologies" 2012, pp. 5299–5308.

[15] R. Ramsin and R. F. Paige, "Process-centered review of object oriented software development methodologies" *ACM Comput. Surv.*, vol. 40, no. 1, pp. 1–89, 2008.

[16] D. E. Avison and G. Fitzgerald, "Where now for development methodologies?" *Commun. ACM*, vol. 46, no. 1, pp. 78–82, 2003.

[17] IEEE Computer Society, "IEEE Standard Computer Dictionary. A Compilation of IEEE Standard Computer Glossaries (610-1991)" IEEE Std 610-1991, 1991.

[18] "Guide to the software engineering body of knowledge (SWEBOK V3) - Software engineering models and methods (Chapter 10 - Unpublished - In Review)" Technical report, 2012.

[19] Centers for Medicare and Medicaid Services (CMS), Office of information Services, "Selecting a development approach" 2008.

[20] D. E. Avison and G. Fitzgerald, *Information systems development: methodologies, techniques, and tools*. Oxford [England]˜ ; Boston: Blackwell Scientific Publications, 1988.

[21] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3" Keele University and University of Durham, Technical report EBSE-2007-01, 2007.

[22] Z. Stapić, E. García López, A. García Cabot, L. de Marcos Ortega, and V. Strahonja, "Performing Systematic Literature Review in Software Engineering" in *Proceedings of 23rd Central European Conference on Information and Intelligent Systems*, Varaždin, Croatia, 2012, pp. 441–447.

[23] J. H. Hosbond and P. A. Nielsen, "Mobile Systems Development - A literature review" in *Proceedings of IFIP 8.2 Annual Conference*, 2005.

[24] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review" *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, 2009.

[25] B. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner, M. Niazi, and S. Linkman, "Systematic literature reviews in software engineering - A tertiary study" *Inf. Softw. Technol.*, vol. 52, no. 8, pp. 792–805, 2010.

[26] A. Nyström, "Agile Solo - Defining and Evaluating an Agile Software Development Process for a Single Software Developer" 2011.

[27] P. Wolkerstorfer, M. Tscheligi, R. Sefelin, H. Milchrahm, Z. Hussain, M. Lechner, and S. Shahzad, "Probing an agile usability process" in *CHI '08 extended abstracts on Human factors in computing systems*, New York, NY, USA, 2008, pp. 2151–2158.

[28] N. Alyani and S. Shirzad, "Learning to innovate in distributed mobile application development: Learning episodes from Tehran and London" in *2011 Federated Conference on Computer Science and Information Systems (FedCSIS).*, Piscataway, NJ, USA, 2011, pp. 497–504.

[29] H. Charaf, "Developing Mobile Applications for Multiple Platforms" in *Engineering of Computer Based Systems (ECBS-*

EERC), 2011 2nd Eastern European Regional Conference on the, 2011, p. 2.

[30] C. Zeidler, C. Kittl, and O. Petrovic, "An integrated product development process for mobile software" *Int. J. Mob. Commun.*, vol. 6, no. 3, pp. 345–356, 2008.

[31] Y. Xiong and A. Wang, "A new combined method for UCD and software development and case study" in *Information Science and Engineering (ICISE), 2010 2nd International Conference on*, 2010, pp. 1–4.

[32] Y.-J. Jeong, J.-H. Lee, and G.-S. Shin, "Development Process of Mobile Application SW Based on Agile Methodology" in *Proceedings of 10th International Conference on Advanced Communication Technology, (ICACT 2008)*, Gangwon-Do, 2008, pp. 362–366.

[33] M. Chen, "A methodology for building mobile computing applications" *Int. J. Electron. Bus.*, vol. 2, no. 3, pp. 229–243, 2004.

[34] S. Cuccurullo, R. Francese, M. Risi, and G. Tortora, "A Visual Approach supporting the Development of MicroApps on Mobile Phones" in *Proc. of 3rd International Symposium on End-User Development*, Brindisi, Italy, 2011, pp. 289–294.

[35] R. Rupnik, "Mobile Applications Development Methodology" in *Handbook of research in mobile business: technical, methodological, and social perspectives*, Second Edition., B. Unhelkar, Ed. IGI Global Snippet, 2009.

[36] P. Abrahamsson, A. Hanhineva, H. Hulkko, T. Ihme, J. Jäälinoja, M. Korkala, J. Koskela, P. Kyllönen, and O. Salo, "Mobile-D: an agile approach for mobile application development" in *Companion to the 19th annual ACM SIGPLAN conference on Object-oriented programming systems, languages, and applications*, New York, NY, USA, 2004, pp. 174–175.

[37] P. Abrahamsson, T. Ihme, K. Kolehmainen, P. Kyllönen, and O. Salo, "Mobile-D for Mobile Software: How to Use Agile Approaches for the Efficient Development of Mobile Applications" 2009.

[38] A. Biswas, T. Donaldson, J. Singh, S. Diamond, D. Gauthier, and M. Longford, "Assessment of mobile experience engine, the development toolkit for context aware mobile applications" in *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology*, New York, NY, USA, 2006.

[39] M. Binsaleh and S. Hassan, "Systems Development Methodology for Mobile Commerce Applications: Agile vs. Traditional" *Int. J. Online Mark. IJOM*, vol. 1, no. 4, pp. 33–47, 2011.

[40] J. Um, S. Hong, Y. T. Kim, E. Chung, K. M. Choi, J. T. Kong, and S. K. Eo, "ViP: A Practical Approach to Platform-based System Modeling Methodology" *J. Semicond. Technol. Sci.*, vol. 5, no. 2, p. 89, 2005.

[41] O. Salo and J. Koskela, "Mobile-D Glossary, VTT Technical Research Centre of Finland, Available at: http://agile.vtt.fi/mobile-d.zip" VTT Technical Research Centre of Finland, 2004.

[42] VTT Technical Research Centre of Finland, "Mobile-D Online Presentation (Web Application)" *AGILE Software Technologies Research Programme*, 2006. [Online]. Available: http://agile.vtt.fi/mobiled.html. [Accessed: 16-May-2012].

[43] VTT Technical Research Centre of Finland, "Mobile-D Description and Templates (ZIP Archive Document), Available at: http://agile.vtt.fi/mobile-d.zip" VTT Technical Research Centre of Finland, 2006.

[44] VTT Technical Research Centre of Finland, "Mobile-D Product Description" *AGILE Software Technologies Research Programme*, 2004. [Online]. Available: http://agile.vtt.fi/prodserv.html. [Accessed: 16-May-2012].

[45] D. Supan, K. Teković, J. Škalec, and Z. Stapić, "Using Mobile-D methodology in development of mobile applications: challenges and issues" in *Razvoj poslovnih i informatičkih sustava CASE 25*, Rijeka, 2013, pp. 91–98.

Upravljanje promjenama na primjeru telekom operatera u Jugoistočnoj Evropi

Adnan Gabela*
* BH Telecom d.d., Sarajevo, Bosna i Hercegovina,

I. UVOD

Dominantni telekom operateri se suočavaju sa sve značajnijom konkurencijom od strane alternativnih pružaoca telekomunikacijskih usluga. Osim toga, regulatorne agencije, ali i brojni međunarodni sporazumi direktno utiču na snižavanje cijena telekomunikacijskih usluga, što rezultuje u padu prihoda. Razvojem tehnologije i popularizacijom alternativnih načina komunikacije i prenosa podataka, također se ugrožava nivo prihoda ostvaren od strane osnovnih telekomunikacijskih aktivnosti.

Konkretno, u 2013. godini je prvi put broj OTT poruka premašio broj SMSova. Over the top (OTT) servisi za razmjenu poruka predstavljaju novu generaciju takvih servisa i uključuju mobilni e-mail, Instant messaging te društvene mreže koje nude razmjenu poruka.

Navedeni servisi su unijeli potpunu revoluciju u način komunikacije, čime je razmjena poruka, poziva i fileova podignuta na novi nivo. Pojava i rast navednih servisa, za koji se predviđa da će se nastaviti i u narednim godinama je direktno uticala na smanjenje prihoda telekom operatera u cijelom svijetu.

Razvojem terminalnih uređaja, vidljivo je kako smartphonei sve više i više predstavljaju „male kompjutere“ na način da je većinu poslova povezanih sa komunikacijom moguće završiti na taj način.

Otvara se potpuno novi prostor za usluge koje su dostupne direktno ili indirektno putem smartphonea. Budući da su tehnološki uslovi već obezbjeđeni, potrebno je osmisliti i prilagoditi ideju za realizaciju.

Tržište telekomunikacija se nalazi u zreloj fazi i zbog toga je jako teško nalaziti potpuno nove korisnike usluga. Umjesto toga, preostaju mogućnosti primamljivanja korisnika od strane konkurenata, te zadržavanje postojećih korisnika i povaćenje prihoda po korisniku (ARPU – Average revenue per user).

Početkom 21. stoljeća većina telekom operatora je pružala osnovne telekomunkacione usluge, uz mogućnost kupovine telefona. U posljednje vrijeme intenziviran je kontakt i poslovna saradnja sa bankama, te danas većina t.o. nudi alternativne načine plaćanja (SMSom, putem POSa, Online..), dostavu računa e-mailom i sl..

Usljed kombinacije navedenih faktora, telekom operateri se okreću prema novim djelatnostima, koje ne predstavljaju osnovne telekomunikacione usluge. Za ulazak u nova područja poslovanja, neophodna je saradnja sa drugim poslovnim subjektima, prije svega iz finansijskog i tehnološkog domena poslovanja kao što su banke, osiguravajuće kuće, softverske kompanije, proizvođači uređaja i sl.

Usljed uslova na tržištu, novih konkurenata, telekom operateri se suočavaju sa neminovnim padom prihoda od osnovne djelatnosti, te se okreću novim alternativnim kanalima ostvarivanja prihoda, koji su najčešće indirektno povezani sa osnovnim tk poslovima. Osim toga, pozicija telekom operatera je u projektima poslovne saradnje sa partnerima je tradicionalno bila dominantna, odnosno svi potencijalni partneri su se morali prilagođavati zahtjevima telekom operatera. [1] U narednim periodima, za veliki broj potencijalnih usluga i tržišta telekom operateri predstavljaju samo još jednog u nizu potencijalnih partnera, odnosno morati će se prilagođavati i pojedinim zahtjevima partnera. Konkretno rad se fokusira na sljedeća područja poslovanja: m2m Machine to Machine, NFC Near Field Communications, učešće na tržištu energenata, mobilno oglašavanje, mobilno bankarstvo, Internet of Things i Smart City.

U radu su detaljno obrazložene karakteristike navedenih područja poslovanja, sa opisom trenuntog stanja, kao i projekcijom trendova u bližoj budućnosti. Dio teorijskog rada objašnjava pojam poslovne prilike i pojam rizika, kao ključne aspekte pri ulasku u nova područja poslovanja.

Prikupljeni su podaci koji karakterišu trenutno stanje tržišta telekomunikacija, sa konkretnim činjenicama vezano za veličinu tržišta, broj korisnika, nivo prihoda i sl. Osim toga, analizirani su trendovi koji se mogu očekivati u narednom periodu, koji su izvedeni na osnovu egzaktnih pokazatelja, te potencijala koje pojedina područja poslovanja pružaju.

U ovom radu je fokus na telekom operaterima u regionu i njihovom odnosu prema novim tehnologijama i područjima poslovanja. Svaka zemlja ima svoje specifičnosti i posebnu situaciju na tržištu, ali prosječno se u svim zemljama nalaze po 3-4 dominantna operatera, od kojih su neki u državnom, a neki u privatnom vlasništvu. U svim zemljama je prisutan i značajan broj alternativnih operatora, no u ovom radu fokus je na dominantim operatorima i njihovom odnosu prema tržištu koje se konstatno mijenja i razvija.

1 Crawford, Sudan, Captive Audience: The Telecom Industry and Monopoly Power in the New Gilded Age (Hardcover), 2013.

II. GLAVNI TEKST

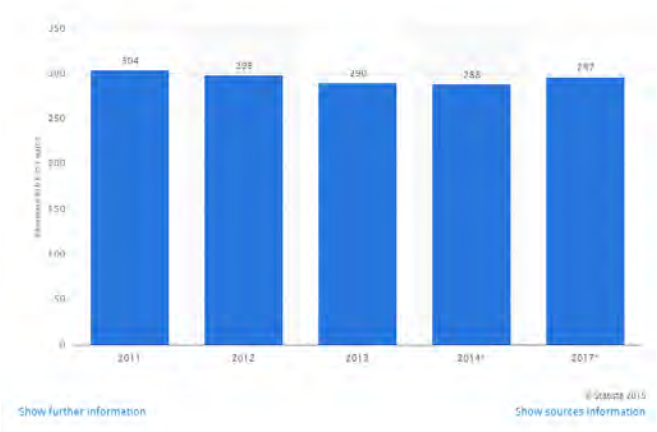
A. Opis postojećeg stanja na tržištu telekomunikacija

Poduzimanjem mjera za veću regulaciju tržišta, te podsticanjem veće konkurentnosti dominantni telekom operateri se suočavaju sa novim prijetnjama. Potpisivanjem sporazuma o ukidanju roaminga na području Evropske unije (Evropski parlament 2014.), te drugim sličnim sporazumima (Budva 2014.), kojima se ukidaju ili snižavaju cijene roaminga telekom operateri su suočeni sa novim prijetnjama. Osim toga masovna upotreba alternativnih načina komunikacije i prenosa podataka (Facebook, Twitter, Viber, WhatsApp, Skype) predstavlja značajne nove konkuretne, protiv kojih se gotovo nemoguće boriti. Navedeni servisi su unijeli potpunu revoluciju u način komunikacije te je pojava i rast navednih servisa, za koji se predviđa da će se nastaviti i u narednim godinama direktno uticala na smanjenje prihoda telekom operatera u cijelom svijetu.

Over the top (OTT) servisi za razmjenu poruka predstavljaju novu generaciju takvih servisa i uključuju mobilni e-mail, IM te društvene mreže koje nude razmjenu poruka. U 2013. godini je prvi put broj ovakvih poruka premašio broj SMSova.

Instant messaging (IM) predstavlja servis sa aplikacijom baziranom na TCP/IP protokolu koji omogućava komunikaciju između korisnika u realnom vremenu, korištenjem raznih terminalnih uređaja. Pomoću ovakvih aplikacija, osobe koje su prijavljene na mrežu (online) mogu dobiti odgovor u jako kratkom vremenu od slanja poruke.

Ukupni prihodi telekom operatera se u Evropi trenutno nalaze u blagom padu, ali se očekuje porast i to ponajviše zahvaljujući novim tehnologijama, što je prikazano na grafikonu ispod.



Slika 1: Prihodi telekom operatera u Evropi za period 2011.-2017. (2014.-2017. procjena); Izvor: Statista – jedna od najvećih online statističkih kompanija, istraživanje provedno 2015.

Slična situacija je i u regionu Jugoistočne Evrope. Konkretno, u analizi konsultantske kuće iCapital za period I-IX 2014. u odnosu na I-IX 2013. od svih Telekom operatora u regionu jedino je Telekom Srpske ostvario rast prihoda.

B. Potencijalna nova područja poslovanja

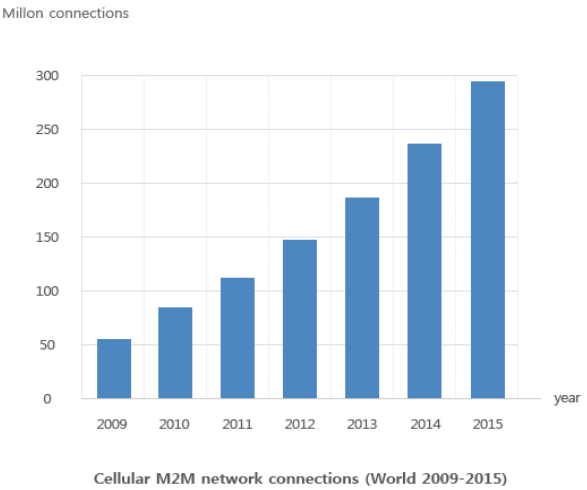
Machine to Machine (m2m) je termin korišten da opiše tehnologije koje omogućavaju kompjuterima, procesorima, smartphoneovima i sl. da komuniciraju jedan sa drugim, vrše mjerenja i donose odluke često bez ikakve ljudske intervencije[2]. Procjenjuje se da će do 2020. godine 50 milijardi uređaja imati pristup internetu, odnosno imati mogućnost komunkacije sa drugim uređajima. Ograničenja za ovo tržište zapravo i ne postoje, budući da se neko rješenje može primijeniti na gotovo sve oblasti poslovanja.

Glavna tržišta gdje m2m ima ili vrlo brzo može imati značajnu ulogu su: zdravstvo, prijevoz, energija, sigurnost i nadzor, upravljanje javnim službama, trgovina na malo, industrijska/kućna automatizacija i kontrola, te poljoprivreda.[3]

Kompleksnost i raznovrsnost potencijalnih rješenja uključuje veliki broj igrača u ovom tržištu. Za mnoga rješenja telekom operateri predstavljaju poželjnog, ali ne i nužno neophodnog partnera.[4] S obzirom na veliki broj mogućnosti koje tržište nudi i trenutnu situaciju da je tržište još uvijek „otvoreno“, odnosno ne postoje ustaljeni dominantni igrači za svaku pojedinačnu oblast, neophodno je djelovati brzo, u cilju zauzimanja djela tržišta.

Prema podacima iz siječnja 2014. 428 mobilnih operatora nude m2m usluge u preko 187 zemalja diljem svijeta. Najveći udio operatora koji nude m2m povezivanje nalaze se u Europi. [Izvor: GSMA Intelligence]

Na slici ispod je prikazan broj konektovanih m2m uređaja zaključno sa 2013. godinom, kao i projekcija do 2015. godine iz čega je uočljiv značajan rast konektovanih uređaja.



Slika 2: Broj konektovanih m2m uređaja Izvor: Gartner INC – Vodeća svjetska kompanija za istraživanje informacionih tehnologija i konsalting iz IT oblasti, osnovana 1979. SAD. Istraživanje o m2m tehnologiji, provedeno 2014.

2 Harbor Research, Market Overview, m2m and smart systems, 2010.

3 Boswarthick, D. Omar Elloumi, O. Olivier Hersent, O. M2M Communications: A Systems Approach, str. 4, 2012., John Wiley & Sons

4 Popović, Ž. Vanesa Čačković, V. Darijo Burjan, D. Arhitektura mreže za M2M komunikacije, str. 9, 2012. Ericsson Nikola Tesla Revija

Opskrba energentima je još jedna mogućnost koja se, ovisno u reguliativama i odnosima na tržištu konkretne države, otvara za telekom operatere. Konkretan primjer je Republika Hrvatska i Hrvatski telekom, koji je ušao u djelatnost opskrbe električnom energijom za krajnje korisnike.
Donošenjem Zakona o energiji, Zakona o izmjenama i dopunama zakona o energiji, Zakona o tržištu električne energije i Zakona o regulaciji energetskih djelatnosti, u Republici Hrvatskoj stvorene su pretpostavke za postupno otvaranje tržišta električne energije. S tim u vezi Hrvatski Telekom je uvidio svoju priliku za učešće na ovom tržištu, te je prema Hrvatskoj energetskoj regulatornoj agenciji HERA podnio zahtjev za obavljanje energetske djelatnosti opskrbe električnom energijom. HERA je izdala navedenu dozvolu, te od kraja 2013. godine Hrvatski Telekom pruža navedene usluge.
Trgovina energentima je određena regulativom konkretne države, te pravila igre mogu biti značajno različita, pogotovo u državama koje nisu članice Evropske Unije.
Za zemlje EU najvažniji dijelovi energetskog zakonodavstva su Direktiva Europske komisije ounutrašnjem tržištu električne energije i Direktiva o unutrašnjem tržištu prirodnog plina, koje su definirale okvire restrukturiranja i demonopoliziranja tržišta mrežnih energenata, u cilju povećanja efikasnosti sektora i snižavanja cijena energenata. Primarni razlog stvaranja tržišta energenata je snižavanje cijena po ugledu na američko tržište, posebno za velike potrošače koji su od iznimne važnosti za konkurentnost ekonomije.
Konkretne prednosti koje telekom operateri nude u ovom domenu poslovanja su razgranata mreža, baza velikog broja korisnika i dobar i uhodan sistem naplate. Predviđanja pokazuju da će se ukupna svjetska potrošnja električne energije u periodu od 2008. godine do 2035. godine povećati za 53 posto. (prema International Energy Outlook 2011, U.S. Energy Information Administration) S obzirom na to, postoji jako značajno područje u kojem se može ostvarivati prihod.

Near Field Communication (NFC) je bežična komunikaciona tehnologija koja se koristi za razmenu podataka na kratkoj udaljenosti do 10 cantimetara.
NFC je standardizirana bežična, bez-kontaktna tehnologija, kratkog dometa koja je napravljena za intuitivnu, jednostavnu i sigurnu komunikaciju između dva elektronska uređaja.[5]
Za komunikaciju putem NFC tehnologije, jedan uređaj mora da radi kao aktivan čitač/pisač, dok je druga strana pasivna i funkcioniše kao tag.
NFC tehnologije primjenjuju se kod bezkontaktnih transakcija kao što su plaćanja računa i karte za prijevoz, a također se koristi za jednostavan i brz prijenos podataka kod elektronskih posjetnica i pristupu online digitalnim sadržajima.
Prednosti i mogućnosti NFC tehnolgoije vrlo brzo su prepoznali pre svega proizvođači hardwera i mobilni telekomunikacioni operateri koji su i pokrenuli i NFC Forum kao nezavisno udruženje kompanija koje žele da razviju i standardizuju NFC tehnologiju u cilju njene aplikativnosti u mnogim oblastima poslovanja. Upravo oni su i pokrenuli, na različitim tačkama u svetu, niz pilot projekata koji su bili uspešni i koji su već prerasli u ozbiljne projekte elektronskog poslovanja u mnogim oblastima ljudskog delovanja kao što su: trgovina, bankarstvo, transport, elektronska uprava, pošta, telekomunikacije, zdravstvo, socijana briga, zabava, hotelijerstvo, marketing i istraživanje tržišta. Kasnije su ovo područje prepoznale i finansijske institucije, prije svega banke i izdavaoci kartica.
Trenutno, MasterCard i Visa preduzimaju značajna ulaganja u Evropi kako bi se stvorila infrastruktura za bezkontaktne transakcije. Još 2009. godine Visa je uložila 10 miliona eura kako bi pomogla malim trgovcima da uposlovanje uvedu novu tehnologiju. Najpoznatiji Visa International bezkontaktni projekt vezan je za plaćanje javnog prevoza bezkontaktnim karticama u Londonu.
Telekom Slovenije je prvi u regionu ušao na ovo tržište. Konkretno, omogućeno je plaćanje namirnica u pojedinim maloprodajnim trgovinama putem telefona. Zbog nedovoljne promocije i tehničkih problema pri rješenju, usluga nije polučila previše dobre rezultate. Ostali telekom operateri još uvijek nemaju razvijena rješenja, te za razliku od m2m projekata, nisu svi još uvijek uvidjeli potencijal ovog tržišta.

Mobilno bankarstvo je sa pojavom smartphoneova i tableta dobilo značajan prostor, te ima mogućnosti za omasovljenje i dalji rast. Kao dio elektronskog bankarstva nudi dodatne pogodnosti korisnicima, prije svega komoditet, dostupnost 24/7, visok kvalitet i nove usluge uz visok stepen sigurnosti podataka i novca. U najrazvijenim zemljama m-banking predstavlja ne samo izvor dodatne zarade, nego i uslov zadržavanja pozicije banke na tržištu. Apsolutno je neizostavno uključiti se u igru i na ovom, relativno novom tržištu. Prilikom razvijanja ovih usluga, banke trebaju pažljivo da ispitaju tržište, izvrše segmentaciju i zajedno sa mobilnim operatorima razviju adekvatne platforme i aplikacije.
Cilj realizacije servisa plaćanja računa jeste da se poveća naplativost i redovnost naplate računa i želja Telekoma da se proširi spisak servisa koji nudi svojim korisnicima i Zato se smatra da će mobilni telefon, odnosno mobilne tehnologije, biti značajan kanal za razvoj elektronskog poslovanja u budućnosti, a po nekim procjenama će prevazići Internet i preuzeti od njega poziciju dominantnog poslovnog kanala elektronskog poslovanja.
Telenor Srbija je krenuo u projekat TelenorBanka, kojim se omogućava korisnicima upravljanje finansijama putem mobilnog uređaja. TelenorBanka omogućava transfer novca u i van zemlje, plaćanje računa na klik, određivanje limita, te čak i podizanje novca sa vlastitih bankomata.
Telenor banka je prva mobilna online banka u regiji. Usluge banke su u prvoj fazi namijenjene samo građanima, koji će preko mobilne aplikacije moći plaćati usluge, slati i primati novac te mijenjati valute.

5 Vasković, J.: Primena NFC tehnologije u sistemima plaćanja, 2012.

Mobilno oglašavanje je način komuniciranja proizvoda ili usluge koristeći mobilne uređaje. Spektar mobilnog oglašavanja je veliki i sadrži razne forme od jednostavnih SMSova do poruka sa interaktivnim sadržajem.
Ova vrsta oglašavanja može ciljati korisnike prema njihovim specifičnim socio demografskim karakteristikama. Mobilni operatori identifikuju mobilne profile, njihove preferencije i prikazuju relevantna oglašavanja kada korisnici download-uju i koriste data servise kao što su online igrice, smartphone aplikacije ili melodije. Ovo su samo primjeri mogućnosti mobilnog oglašavanja, ali ovaj vid oglašavanja ima mnogo veći potencijal i penetraciju, kao što su društvene mreže, navigacijske aplikacije ili mobilne web stranice.[6]
Mobilno oglašavanje u zapadnoj Evropi trenutno čini otprilike trećinu svih digitalnih investicija, dok su u SAD-u ti brojevi još i veći, s daleko najvećim trendom rasta od svih načina oglašavanja. Samo u Evropi u 2013. mobilno oglašavanje zabilježilo je rast od čak 123 posto. Prema definiciji, mobilno oglašavanje podrazumijeva oglašavanje roba i usluga korisnicima mobilnih uređaja i pametnih telefona.
Brojne su mogućnosti za komunikaciju s kupcima; pasivni potrošač postao je prošlost i očekuje se interakcija. Za to služe nagradna pitanja, kvizovi, glasovanja, ankete mobilni kuponi, skupljanje bodova – time brend dobiva kontinuiranu individualnu komunikaciju sa strankama i korisnicima, što mu je i najveća vrijednost. Kompanije tako dobivaju trajnu dostupnost klijentima, proširenje ciljane skupine potrošača, pokretanje dijaloga s njima i nadogradnju vlastita portala ili internetske stranice multimedijalnim sadržajima, od kojih ostvaruje prihode, kao i od mobilnog marketinga, čime korist postaje višestruka.

Internet of Things (IoT) predstavlja koncept u kojem su uređaji, vozila, građevine i drugi objekti opremljeni elektronikom, softwareom, senzorima uvezani na način da mogu vršiti razmjenu podataka. Osnovne karakteristike sistema baziranih na Internetu stvari su:

1. Sve komuncira: Pametne stvari imaju mogućnost bežične komunikacije između sebe i između međusobno povezanih objekata
2. Sve se identifikuje: Pametne stvari se identifikuju preko digitalnog imena
3. Sve interreaguje: Pametne stvari mogu interreagovati sa lokalnim okruženjem kroz očitovanja i aktiviranja postojećih mogućnosti.[7]

Internet stvari obećava mnogo ali prvenstveno je neophodno odgovoriti na mnoge tehničke i ekonomske izazove. Jasno je da razvoj Interneta stvari zavisi od dinamike inovacija u brojnim oblastima tehnike i da će za potpuno usvajanje Interneta stvari trebati vremena.
Od fleksibilnosti i sposobnosti samih kompanija da apsorbuju nove pametne informacione tehnologije, zavisiće i njihova spremnost da se u konkurentskom smislu nadmeću na tržištu, bez obzira na njihovu veličinu i finansijsku moć.
Preduzeća koja uspiju da se povežu na Internet stvari i koriste podatke i analitiku kako bi razvili predviđajuće algoritme mogu da ubrzaju efikasnost, dramatično povećaju produktivnost, smanje upotrebu energije i drugih resursa, te dramatično smanje granične troškove proizvodnje i distribucije proizvoda ili usluga.
Kako se umreženost i dalje sve više razvija, a tehnička rješenja postaju sve dostupnija, paralelno s njima rastu i očekivanja korisnika.
U tom kontekstu izazovi za istraživanja i razvoj u svrhu stvaranja „pametnog svijeta“ koji povezuje realno, digitalno i virtualno su ogromni.
BH Telecom i QSS su 2015. godine u Sarajevu predstavili u regionu jedinstvenu Internet of Things (IoT) platformu sa projektima na kojima zajednički rade: Connected Home sistem – Sistem za upravljanje i nadzor privatnim objektima, Smart Metering – Sistemi za upravljanje obračunom i potrošnjom komunalnih usluga, Connected Office – Sistem nadzora za poslovne objekte i data centre, kao i Solar-Meteo sistem za različite namjene u industriji poput upravljanja poljoprivrednim dobrima, nasadima voća i povrća i sl.

Smart City (SC) predstavlja grad u kojem je uvezana fizička, poslovna, IT i društvena infrastruktura sa ciljem uvećavanja ukupnog kvaliteta života, kroz bolju povezanost, obrazovanje, zdravstvo, sigurnost, transport uz očuvanje prirodne sredine.
Sa tehničkog aspekta, za implementaciju koncpeta Smart Cityja neophodno je prisustvo ICT tehnologija u svim aspektima. Realizacija koncepta u mnogome ovisi o ICT komponenti, gdje je neophodna integracija sa ključnim elementima gradske infrastrukture u cilju prikupljanja i razmjene podataka o svim ključnim komponentama grada u real timeu.[8]
Za realizaciju navedenih projekata neophodan je dalji razvoj tehnoloških rješenja, uvođenje 4G, a vremenom i 5G tehnologije, ali i integracija sa drugim potencijalnim učesnicima IOT i SC koncepta (gradske uprave, softverske kompanije, hardverske kompanije, komunalna preduzeća, obskrbljivači energentima, gradski prevoz i sl.)
Na temu Smart Cities su u Zagrebu održane dvije konferencije (15.5.2015. i 17.3.2016.) godine sa ciljem realizacije projekata, ali za sada se nijedan grad u regionu ne može okarakterisati kao SC, a kao prvi takav je u saradnji sa Hrvatskim telekomom najavljen grad Dubrovnik.

6 Prašljivić, A., Njemčević, L. Služba za marketing istraživanje, BH Telecom, Izvještaj o Rezultatima Istraživanja: Mobilno Oglašavanje, 2014.

7 Vujanović V, Maksimović M: Internet stvari – tehnički i ekonomski aspekti primjene, str. 2. 2015., Infoteh Jahorina

8 Chourabi H, Nam T, Walker S: Understanding Smart Cities: An Integrative Framework, str 2. 2015. Hawaii International Conference on System Sciences

C. Poslovna prilika i rizik

Poslovna ideja se može definisati kao kratak opis osnovnih funkcija budućeg poslovnog poduhvata. Ona predstavlja kombinaciju naših resursa i vrijednosti za krajnjeg kupca ili korisnika.

Poslovna ideja je prvi korak prema realizaciji zacrtanih ciljeva i mora dati odgovore na sljedeća pitanja:

1. Koji proizvod/uslugu ćete proizvoditi, odnosno pružati?
2. Ko će biti vaši kupci / korisnici?
3. Na koji način ćete prodavati/pružati svoj proizvod/uslugu?
4. Koje će potrebe vaš proizvod/usluga pružiti kupcima/korisnicima?

Često se pogrešno misli kako poslovna ideja nužno mora biti inovativna. Naprotiv, iznenađujuće je koliko je malo ideja zasnovano na apsolutno novim idejama. Značajno više uspješnih poduhvata zasnovano na adaptaciji starih, iskušanih i provjerenih ideja.

Poslovna ideja i poslovna prilika često se koriste kako bi opisali isto - ono na čemu se zasniva svaki poslovni poduhvat. Međutim, iako povezani, ova dva pojma ne predstavljaju sinonime. U osnovi svake poslovne prilike je ideja, dok obrnuto ne vrijedi: od 100 ideja prezentiranih potencijalnim investitorima u obliku poslovnog plana, obično 1, a vrlo rijetko 2-3 ideje dobiju potvrdan odgovor za finansiranje. Samo su one predstavljale poslovnu priliku.

Poslovna prilika predstavlja tržišno provjerenu poslovnu ideju. Postoje 4 osnovna elementa koja razlikuju ideju od poslovne prilike:

1. Poslovna prilika stvara ili dodaje vrijednost za kupca, odnosno krajnjeg korisnika proizvoda /usluge
2. Dodavanje vrijednosti ostvaruje se kroz rješavanje problema, odnosno zadovoljenje, neke želje ili potrebe,
3. Pretpostavlja se postojanje određenog tržišta na kojem se može ostvariti određena marža i postoji potencijal za rast
4. Postoji dobar odnos između nosioca ideje , njegovih želja i sposobnosti tima, s kojim radi na realiziranju poslovne prilike; [9]

Rizik se definiše kao djelovanje nesigurnosti na ciljeve organizacije. Nesigurnost može imati za posljedicu odstupanje od očekivanih rezultata. Ciljevi organizacije mogu biti različiti, npr: finansijski ciljevi, ciljevi zaštite zdravlja i sigurnosti i zaštite okoliša, a mogu se primijeniti na cijelu organizaciju ili različita područja njezina djelovanja. Rizik se često opisuje odnosom mogućih događaja (pojava ili promjena posebnoga skupa okolnosti) i posljedica (rezultatima događaja koji utječu na ciljeve), ili njihovom kombinacijom. On se često izražava kao kombinacija posljedica nekoga događaja (uključujući promjene u okolnostima) i pridružene vjerojatnosti njegove pojave.

U literaturi se mogu pronaći različite definicije i objašnjena rizika. Robert L. Simons definira rizik kao: "Neočekivani događaj ili niz okolnosti koji u značajnoj mjeri reducira sposobnost menadžera da implementiraju svoju namjeravanu poslovnu strategiju". Tačnije rečeno rizik je vjerojatnost prijetnje da iskoristi neku ranjivost imovine te time ugrozi imovinu.

U konkretnom slučaju rizik predstavlja ulazak u nova, neistražena i neregulisana tržišta, koja ne predstavljaju primarnu djelatnost firme, ali također rizik predstavlja i nezaiteresovanost za nove prilike.

Svjesno ili nesvjesno propuštanje poslovnih prilika, odnosno pristup „ne činiti ništa“ u ovom slučaju samo po sebi predstavlja rizično poslovanje. Kompanije imaju obavezu i odgovornost pratiti sve potencijalne zone poslovanja, koje direktno ili indirektno mogu da ostvare uticaj na njih. To se ne odnosi samo na konkurenciju nego sve zone poslovanja koje su naslonjene na industriju kojom se kompanija primarno bavi. Vlasnik ključnih informacija ima mnogo veću šansu za uspjeh.

U oblasti u kojima djeluju telekom operateri, navedena činjenica je još više izražena [10] . Zbog konstantnih inovacija, te napretka tehnologije, ulazak u nova područja poslovanja predstavlja neophodnost, ne samo mogućnost.

Rizik nije moguće izbjeći jer nema poduzetničke aktivnosti bez rizika. U rizike treba ulaziti, ali planski. Rizik treba prihvatiti kao realnost i učiniti sve kako bi se došlo u fazu upravljanja rizicima.[11]

III. ZAKLJUČAK

Telekom operateri se nakon dugog procesa dominacije na tržištu i konstanog rasta prihoda suočavaju sa značajnim poteškoćama i problemima.

Stagnacija ili pad prihoda telekom operatera od osnovne djeltanosti, ne znači da nevedene kompanije trebaju sjediti skrštenih ruku, nego se okrenuti ka novim mogućnostima.

Naravno da osnovne telekomunikacione aktivnosti trebaju i dalje ostati u fokusu operatera, budući da i dalje generišu najveći dio prihoda. Razvoj i uspješna implementacija osnovnih telekomunikacionih usluga trebaju ostati primarna djelatnost prema kojoj se treba odnositi sa najvećim prioritetom.

Međutim, kako je i detaljno prezentirano u radu, mogućnosti se otvaraju u novim područjima: m2m, NFC, mobilna plaćanja, trgovina energentima, mobilno oglašavanje, IOT, SC...

Inovativne funkcije u modernim sistemima, na osnovu iznijetih primjera u radu doslovno ne poznaju granice i mogu biti različite zavisno od područja primjene. Mobilni uređaji danas predstavljaju gotovo neodvojivi dio ličnosti

9 Ekonomski fakultet u Osijeku, Vodič za sve, koji razmišljaju o ulasku u poduzetničke vode, str. 4,

10 Marušić, N. Pregled metoda ublažavanja utjecaja endogenih faktora na neuspješnost IT projekata, 2013., Mipro

11 Drljača, M. Održivi uspjeh i upravljanje rizicima poslovanja, str. 15, 2010., Centar za kvalitet Crne Gore

svakog korisnika i konstatano nude sve više i više mogućnosti, a sudeći prema trendovima, korištenje i broj usluga koje moblni uređaji nude će se samo povećavati.

Dva segmenta koja su ključna za uspjeh su istraživanje i razvoj (R&D), odnsono tehnički dio prilagodbe inovacijama, te marketing, odnosno pravilan pristup i fokus na korisnike. Oba segmenta su neophodna za uspjeh, te njihova kvalitetna balansiranost i praćenje jedan drugog. Marketing predstavlja kariku koja podržava odnos cijene, usluge i kanala distribucije, ali uprkos svom značaju, ne može nadoknaditi evenutalne propuste u kvalitetu jednog od pobrojana tri elemetna.

Pored R&D i marketinga, za poslovni uspjeh je neophodno pažljivo izabrati strateške partnere za nova područja poslovanja, budući da je za sve pobrojane potencijalne aktivnosti neophodan razvoj i implementacija usluga sa partnerima. Zajednički ulazak na navedena tržišta i razvoj proizvoda predstavlja kompleksan strateški i operativni zadatak, ali je neophodan za poslovni uspjeh.

Za kvalitetno upravljanje promjenama i praćenje trendova u industriji koja je konstatno podložna promjenama neophodno je kvalitetno donošenje i implementacija odluka. Današnja kretanja ukazuju na brze i opsežne promjene, kao i dominaciju znanja kao temeljnog resursa na kojem se zasniva društveni i privredni razvoj. Najvažniji korak u procesu upravljanja promjenama koji treba da učini menadžment preduzeća je da uoči situaciju u kojoj su promjene potrebne i neophodne.

Veća i povoljnija dostupnost telekomunikacijskih usluga ima za posljedicu veću dostupnost informacija za sve građane a time posredno utiče i na obrazovanje. Dobro razvijena telekomunikacijska mreža predstavlja nužan preduslov ekonomskog napretka zemlje.

Prostor za djelovanje u posmatranim zonama poslovanja definitivno postoji, ali je neophodno djelovati na vrijeme, ići u korak sa trendovima i iskoristiti postojeće potencijale koje kompanija posjeduje.

Vremenski okvir je od ključnog značaja za iskorištenje poslovne prilike. Politika „ne raditi ništa“ predstavlja značajan rizik, koji može dovesti do gubitka prihoda.

Na osnovu informacija iznesenih u radu, zaključuje se da postoji mogućnost ulaska u posmatrane zone poslovanja , ali je za realizaciju tog cilja neophodan angažman zaposlenih, kako na strateškom tako i na operativnom nivou.

LITERATURA

[1] Adizes, I. Corporate Lifecycles: How and Why Corporations Grow and Die and What To Do About It., 1988., Prentice Hall

[2] Boswarthick, D. Omar Elloumi, O. Olivier Hersent, O. M2M Communications: A Systems Approach, str. 4, 2012., John Wiley & Sons

[3] Buntak, K. Drožđek, I. Košćak, M. Metodologija implementacije upravljanja rizicima fmea metodom, str. 27, 2014., Tehnički glasnik, Sveučilište Sjever

[4] Čagalj, M. Sigurnost u M2M (Machine-to- Machine) komunikacijama, str. 6, 2013., Mipro

[5] Chourabi H, Nam T, Walker S: Understanding Smart Cities: An Integrative Framework, str 2. 2015. Hawaii International Conference on System Sciences

[6] Crawford, Sudan, Captive Audience: The Telecom Industry and Monopoly Power in the New Gilded Age (Hardcover) , 2013.

[7] Drljača, M. Održivi uspjeh i upravljanje rizicima poslovanja, str. 15, 2010., Centar za kvalitet Crne Gore

[8] Drucker, P. Innovation and Entrepreneurship, str.45, 1985., Heinemann, London

[9] Ericsson White Paper, More than 50 billion connected devices, February 2011

[10] ETSI, Universal Mobile Telecommunications System (UMTS);LTE;IP

[11] Evans, D. The Internet of Things How the Next Evolution of the Internet Is Changing Everything, CISCO white paper

[12] Gabela J, Uslovi za razvoj poduzetništva u BiH, 2015.

[13] Gertner, Jon The Idea Factory: Bell Labs and the Great Age of American Innovation (Hardcover) , 2012.

[14] Harbor Research, Market Overview, M2m and smart systems, 2010.

[15] Kirzner, I. Competition and Entrepreneurship, str. 56, 1973., Chicago: University of Chicago Press

[16] Krpan, LJ. Maršanić, R. Značenje i uloga hrvatskoga prometnog sistema u europskim prometnim tokovima, Suvremeni promet, Opatija, 2004.

[17] Lukičić, M. M2M komunikacije u primjeni naprednih elektroenergetskih mreža, str. 56, 2012., Ericsson Nikola Tesla Revija

[18] Maračić, H. Energentska učinkovitost u komunikaciji stroja sa strojem, str. 6, 2014., Mipro

[19] Marin, D.Telekomunikacijska legislativa i standardizacija, 2006. Zagreb: Kigen

[20] Marušić, N. Pregled metoda ublažavanja utjecaja endogenih faktora na neuspješnost IT projekata, 2013., Mipro

[21] Multimedia Subsystem (IMS) emergency sessions, TS 123 167 V11.9.0,December, 2013.

[22] Poletan-Jugović, T. Integracija R Hrvatske u svjetske tokove energenata, str. 584, 2009., Sveučilište u Rijeci, Pomorski fakultet u Rijeci

[23] Popović, Ž. Vanesa Čačković, V. Darijo Burjan, D. Arhitektura mreže za M2M komunikacije, str. 9, 2012. Ericsson Nikola Tesla Revija

[24] Porter, M. Konkurentska prednost, str. 27, 1985., Masmedia

[25] Prašljivić, A., Njemčević, L. Služba za marketing istraživanje, BH Telecom, Izvještaj o Rezultatima Istraživanja: Mobilno Oglašavanje, 2014.

[26] Skočir, P. Suradnja umreženih uređaja u komunikacijskim sistemima stroja sa strojem, str. 2, 2014., Mipro

[27] Vasković, J.: Primena NFC tehnologije u sistemima plaćanja, 2012.

[28] Vujanović V, Maksimović M: Internet stvari – tehnički i ekonomski aspekti primjene, str. 2. 2015., Infoteh Jahorina

Tehnologije integracije informacijskih sustava

Aleksandar Stojanović *, Nikolaj Lazić ** i Željko Kovačević ***

* Tehničko veleučilište u Zagrebu/informatičko-računarski odjel, Zagreb, Hrvatska
** Sveučilište u Zagrebu, Filozofski fakultet/odsjek za fonetiku, Zagreb, Hrvatska
*** Tehničko veleučilište u Zagrebu/informatičko-računarski odjel, Zagreb, Hrvatska
{ astojanov@tvz.hr, nlazic@ffzg.hr, zkovacev1@tvz.hr }

Sažetak - Danas je uobičajeno za veće organizacije da imaju više informacijskih sustava od kojih svaki obuhvaća jedan segment podataka s kojima radi dotična organizacija. Također je uobičajeno da su ti sustavi razvijani tokom dužeg vremenskog perioda, pa je stoga i vrlo vjerojatno da su se poslovni uvjeti unutar kojih posluje dotična organizacija promijenili. To znači da se pojavila potreba za analizom novih vrsta podataka, a time i za novim informacijskim sustavom čijom je funkcionalnošću potrebno obuhvatiti te podatke. Zbog ovoga organizacije često moraju koristiti više zasebnih informacijskih sustava, što usporava, a time i poskupljuje funkcioniranje same organizacije. Kao rezultat, javlja se potreba za povezivanjem, odnosno integracijom takvih sustava. U ovom radu razmatraju se neke popularne tehnologije za integraciju informacijskih sustava. Nadalje, uspoređuju se razlike između dvije različite arhitekture ovakvih sustava – SOA, kao dobro utvrđene arhitekture, i mikroservisi, kao relativno nove arhitekture koja je posljednjih godina u usponu.

Ključne riječi: integracija, mikroservisi, SOA, ESB

I. UVOD

U informacijskoj tehnologiji integracija sustava odnosi se na povezivanje različitih programskih sustava tako da se sa njima može raditi kao s jednom cjelinom, odnosno kao s jednim sustavom. Iako se informacijski sustavi najčešće razvijaju na osnovu zasebnih zahtjeva i grupa korisnika, s tehničkog aspekta često je lakše napraviti dva ili više manja informacijska sustava koji se kasnije mogu povezati, nego jedan kompleksni sustav koji bi uključivao puno veći skup zahtjeva i korisnika. Integracija informacijskih sustava je stoga korisna ne samo za povezivanje s "legacy" sustavima, nego i kao tehnička strategija izrade informacijskih sustava s ciljem veće jednostavnosti i fleksibilnosti.
Integracija informacijskih sustava obično uključuje tri vrste integracije:

- Integracija podataka - omogućuje korištenje podataka od strane dva ili više različitih sustava i osigurava da su podaci konzistentni.
- Integracija procesa - povezivanje operacija iz različitih odjela organizacije što omogućuje veću efikasnost i spremnost na promjene.
- Integracija servisa – povezivanje funkcionalnosti različitih modula; najčešći način ovakve integracije su SOA (engl. *Service-Oriented Architecture*) i ESB (engl. *Enterprise Service Bus*).

SOA je paradigma za arhitekturu fleksibilnih programskih sustava koji se sastoje od više modula koji se nazivaju servisima. Iako ih se može promatrati odvojeno, jedna od ključnih komponenti SOAe je ESB, a njegova osnovna uloga je povezivanje servisa koji su dio funkcionalnosti nekog šireg sustava. ESB omogućava integraciju aplikacija iz heterogenih okruženja i ima ulogu u sljedećim aktivnostima:

- Upravljanje i koordinacija poruka koje razmijenjuju servisi
- Usmjeravanje poruka
- Transformacija podataka
- Sigurnost
- Osiguravanje pouzdanosti komunikacija
- Kontrola instaliranja servisa

Ako govorimo o informacijskim sustavima, većina takvih sustava se sastoji od većeg broja aplikacija koje mogu (ali i ne moraju) imati pristup zajedničkoj bazi podataka. Te su aplikacije uglavnom *monolitne*, što znači da se sastoje od „jednog komada“, to jest, svaka promjena u kodu zahtijeva ponovno instaliranje cijele aplikacije. S obzirom da mnogi informacijski sustavi uključuju takozvane „legacy“ aplikacije koje su gotovo uvijek monolitne, takve se aplikacije obično prilagode tako da mogu biti dio jednog šireg sustava koji uključuje ESB komponentu.
Jedna relativno nova paradigma za arhitekturu aplikacija zove se *mikroservisi*. Najvažnije obilježje mikroservisa je to da su servisi sastavni dio jedne aplikacije (koja je napravljena na taj način). To znači

da je izrada aplikacije u više servisa i komunikacija među tim servisima sastavni dio ove arhitekture. Drugim riječima, za razliku od monolitnih aplikacija koje se poslije mogu prilagoditi i uključiti u neki veći informacijski sustav upotrebom ESBa, aplikacije napravljene mikroservisima imaju strukturu koja je sama po sebi takva da joj naknadne prilagodbe i tehnologije (kao ESB) nisu potrebne. To znači da je kod mikroservisa integracija komponenti karakteristika arhitekture samih aplikacija, a ne zasebna paradigma ili tehnologija kojom možemo naknadno povezati skup monolitnih aplikacija.
Druga važna karakteristika mikroservisa je *decentralizacija*: svaki servis može raditi sa svojim podacima, a drugi servisi mu pristupaju samo kroz njegovo sučelje, nikad direktno njegovim podacima. Nadalje, na ovaj način svaki servis ne mora biti vezan za jednu te istu tehnologiju podataka; na primjer, jedan servis može koristiti klasičnu SQL bazu podataka, dok drugi može koristiti neku od NoSQL tehnologija, kao što su dokument ili graf baze podataka. Općenito, servisi su nezavisni moduli aplikacije koji imaju precizno definirano sučelje preko kojih se koristi njegova funkcionalnost.

II. OSNOVNI PRISTUPI INTEGRACIJI APLIKACIJA

Najčešći pristup integraciji aplikacija je integracija poslovnih procesa (engl. *Business Process Integration, BPI*), integracija servisa i integracija podataka (engl. *Enterprise Information Integration, EII*).
Arhitektura modernih aplikacija općenito se sastoji od tri razine [3]:

- Prezentacijska razina - to je razina na kojoj se podaci prikazuju korisniku.
- Aplikacijska razina - ova se razina sastoji od poslovne logike i drugih stvari koje određuju funkcionalnost same aplikacije.
- Podatkovna razina - ovo je obično razina koja se sastoji od baze podataka. Ova razina ima aplikacijsko sučelje, API (engl. *Application Programming Interface*), koji aplikaciji omogućava pristup podacima.

Ove razine djelomično se preslikavaju i na integraciju:

- Integracija procesa
- Integracija servisa
- Integracija podataka

Svrha integracije procesa je da automatizira njihovo izvršavanje čime bi se na efikasniji način postiglo isto što i njihovim izvršavanjem putem pojedinačnih informacijskih sustava. Dva osnovna pristupa integracije procesa su BPM (engl. *Business Process Management*) i BPEL (engl. *Business Process Execution Language*).

Integracija servisa osigurava cjelovitu i potpunu povezanost integriranih sustava. Servisi se sastoje od neovisnih komponenti koji se povezuju sa za to predviđenim protokolima. Ti su protokoli detaljna specifikacija komunikacije među servisima koja određuje kako se izmjenjuju poruke i podaci. Jedan od općeprihvaćenih pristupa integraciji servisa su SOA i ESB. SOA je vrsta arhitekture programskih sustava koje je orijentirana ka servisima. U tom kontekstu, servis je logički prikaz neke poslovne aktivnosti sa specificiranim ishodom. ESB je komponenta SOA arhitekture koja se upotrebljava za implementaciju komunikacije između aplikacija. To je sloj iznad sustava za slanje poruka koji omogućava povezivanje sustava na višoj razini apstrakcije. SOA koristi razne tehnologije, kao što su XML web-servisi, SOAP (engl. *Simple Object Access Protocol*), RPC (engl. *Remote Procedure Call*), CORBA (engl. *Common Object Request Broker Architecture*) i drugo. Još jedna tehnologija koja je u sve široj upotrebi u novije vrijeme je i REST (engl. *REpresentational State Transfer*), koja se često uspoređuje sa SOAPom, iako je SOAP više protokol, a REST arhitektura.

Integracija podataka ima za cilj da korisniku pruži cijeloviti pregled podataka koji mogu dolaziti s različitih izvora ili se nalaziti na različitim lokacijama. Na primjer, korisnik može tražiti popis kupaca koji su kupili neki proizvod; informacije o kupcima se mogu nalaziti na jednom mjestu, a one o proizvodima na drugom, pa se integracijom ovih podataka korisniku može pružiti prikaz na kojem su ti podaci povezani.

III. INTEGRACIJE INFORMACIJSKIH SUSTAVA

Neki od često korištenih informacijskih sustava u velikim organizacijama su CRM (engl. *Customer Relationship Management*), GIS (engl. *Geographical Information Systems*), ERP (engl. *Enterprise Resource Planning*), SCM (engl. *Supply Chain Management*) [3], zdravstveni informacijski sustavi i financijski informacijski sustavi. CRM sustavi se koriste za upravljanje suradnjom i odnosom s klijentima, te uključuju sustave za korisničke usluge, prodaju, marketinštvo i tehničku podršku. GIS sustavi upotrebljavaju se za spremanje, obradu i analizu prostornih i geografskih podataka. ERP sustavi se sastoje od skupa integriranih aplikacija za rad s podacima iz raznih područja aktivnosti neke organizacije. SCM sustavi koriste se za upravljanje protokom robe i servisa. Zdravstveni informacijski sustavi koriste se u zdravstvenim ustanovama za vođenje evidencije o povijesti bolesti pacijenata. Financijski

informacijski sustavi općenito obuhvaćaju sve aspekte obrade financijskih podataka.
U ovom dijelu ukratko su opisane neke mogućnosti integracije gore opisanih vrsta informacijskih sustava.

A. Integracija GIS i ERP informacijskih sustava

Kod integracije ovakvih sustava postoji nekoliko mogućnosti:

- Izrada ili nabava programskih konektora koji direktno spajaju neke GIS i ERP module;
- Korištenje posredničkih (engl. middleware) programa koji predstavljaju vezu između sustava ili pojedinačnih modula;
- Instaliranje okruženja za cijelovitu integraciju s GIS sustavom.

Neka od najuspješnijih riješenja za integraciju ovakvih sustava napravile su firme ESRI i SAP [4]. Postoji nekoliko osnovnih tehničkih sučelja za integraciju SAPovog ERP sustava sa ESRIjevim GIS sustavom:

- SAP RFC konektori – to je sučelje koje omogućava pozive funkcija između SAPovog ERP sustava i ESRIjevog GIS sustava;
- Konektori drugih proizvođača – ovi uključuju Way Control Broker od firme Information Builders koji omogućava direktnu vezu sa starijim sustavima;
- SAP generički GIS konektor – ovo je "middleware" koji koordinira između SAP i ESRI sustava;
- EAI drugih proizvođača su kombinacija platformi, poslovnih procesa, standarda i aplikacija koji omogućavaju integraciju dva ili više informacijskih sustava;
- SAP EAI – SAP Exchange Infrastructure je SAPov EAI sustav za razmjenu informacija između SAP ERP i ESRI GIS;
- SICAD-APX je ESRIjevo riješenje za EAI integraciju njihovog GIS sustava sa SAPovim ERP modulima

Integracija GIS i ERP postoji i za Oracle kroz njihov Oracle Spatial 10g sustav. To je baza podataka koja može spremati i obrađivati geospacijalne podatke koristeći SDO_GEOMETRY tip podataka, te prostorne funkcije i operatore.

B. Integracija CRM informacijskih *sustava*

CRM informacijski sustavi su implementirani u mnogim organizacijama koje su fokusirane na dugoročne odnose s klijentima. Korištenje CRM sustava podrazumijeva organiziranje, automatiziranje i usklađivanje poslovnih procesa – što uključuje aktivnosti oko poslovnih procesa prodaje, marketinga, tehničke podrške i korisničkih usluga.

Integracija CRM sustava s drugim sustavima nije jednostavan proces jer on može uključivati veći broj informacijskih sustava koji pokrivaju različite domene podataka. Iako integracija CRM sustava s drugim sustavima može rezultirati vrlo kompleksnim integriranim sustavom, prednosti ovakve integracije i dalje nadmašuju nedostatke. Te se prednosti uglavnom odnose na kvalitetu i efikasnost poslovanja, smanjene troškove, bolju podršku u odlučivanju, te bolju agilnost organizacije.

C. Integracija SCM informacijskih sustava

Osnovni fokus SCM (Suppy Chain Managemet) sustava je proizvodnja robe i servisa kada i gdje je to potrebno, s ciljem bolje efikasnosti i zadovoljavanja potreba tržišta. Neke od prednosti integracije SCM sustava s drugim sustavima jesu

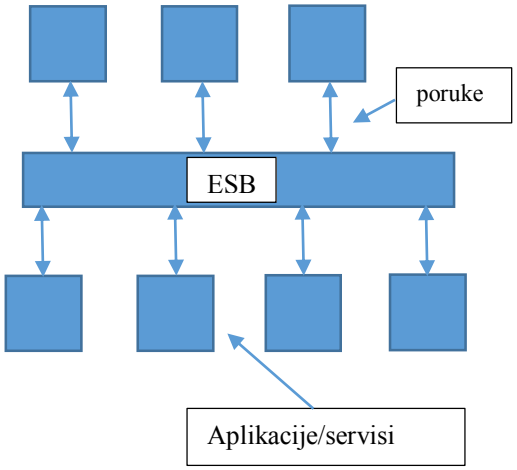
- Bolja kontrola nad dobavljačima i njihove kvalitete;
- Efikasniji procesi i veća spremnost i iskoristivost opreme;
- Poboljšana komunikacija i suradnja između članova u lancu opskrbe;
- Efikasnije isporučivanje robe i odgovaranje na potrebe naručitelja;
- Veća produktivnost i osjetljivost na potrebe tržišta
- Reduciranje troškova i zaliha

Općenito, glavna prednost u korištenju SCM sustava je bolje operativno i poslovno planiranje, što organizacijama omogućava da se bolje usklade s potrebama tržišta.

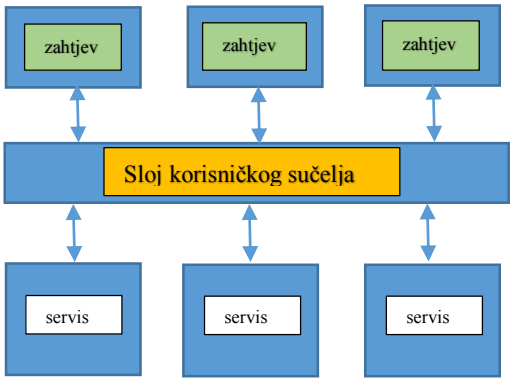
D. Integracija zdravstvenih informacijskih sustava

Zdravstveni informacijski sustavi omogućavaju pristup informacijama o povijesti bolesti pacijenata. Takvi se sustavi često sastoje od većeg broja aplikacija koje su u upotrebi u raznim odjelima bolnica, kao što su radiologija, pedijatrija, ortopedija, dermatologija i drugi. Svrha povezivanja ovih aplikacija je efikasnije i kvalitetnije osiguravanje zdravstvene njege korisnika zdravstvenog sustava. Ovi informacijski sustavi često se moraju povezati sa sustavima drugih organizacija, kao što su ljekarne i privatne klinike [9]. Za povezivanje tih sustava često se koristi SOA arhitektura, zajedno sa posredničkom komponentom za razmjenu poruka, a česta je i upotreba mrežnih servisa i dotičnih tehnologija, kao što su SOAP, REST, XML i JSON.

Slika 1: Model SOA arhitekture programskog sustava [2].



Slika 2: Struktura sustava s mikroservisima [7, 8].



E. Integracija financijskih informacijskih sustava

Financijski informacijski sustavi koji se upotrebljavaju u većim organizacijama, kao što su banke, u pravilu se sastoje od više podsustava. Na primjer, takav se jedan sustav može sastojati od sustava za vođenje računa, sustava za odobravanje kredita i sustava za odobravanje zajmova [10]. Takvi sustavi zahtijevaju integraciju na više razina, specifično i integraciju servisa i integraciju podataka. SOA je kod ovakvih sustava jako zastupljena arhitektura koja je odgovorna za povezivanje velikog broja aplikacija, u nekim slučajevima na tisuće njih [11].

IV. DISKUSIJA

Iako postoje razne vrste ESBa, osnovni model takve arhitekture sustava može se prikazati kao na slici 1. Naizgled konceptualno slični SOAi, mikroservisi se od SOAe razlikuju u nekoliko stvari. Kao prvo, cilj SOAe je integracija više aplikacija, dok više mikroservisa u pravilu pripada samo jednoj aplikaciji. Kao drugo, prethodno navedene uloge ESBa u principu nisu dio ove arhitekture. Iako neke varijante mikroservisa uključuju komponentu „posrednika“ za poruke (engl. *message broker*), ta je komponenta daleko manje kompleksna od tipičnog ESBa – ona uglavnom služi za prijenos poruka, omogućava asinkrono slanje poruka i adekvatno balansiranje količine poruka (engl. *load balancing*). Međutim, to je samo jedna kompleksnija varijanta ove arhitekture; mnoge druge varijante uopće ne uključuju posredničku komponentu.

Kao treću krakteristiku ove paradigme treba spomenuti i pogodnost za automatsko stavljanje servisa u produkciju. S obzirom da su servisi nezavisni moduli jedne aplikacije, postoje alati kojima se servisi mogu lako i brzo zamijeniti ili dodati u sustav, bez da se mora zamijeniti cijela aplikacija. Ovo je puno teže napraviti s klasičnom SOA arhitekturom, jer ona ne podrazumijeva da je cijeli sustav skup nezavisnih servisa, već je to često skup monolitnih aplikacija koje se moraju instalirati posebnim procedurama. Sve ovo doprinosi efikasnosti izrade aplikacija korištenjem mikroservisa.

Slika 2 pokazuje osnovnu strukturu sustava baziranog na mikroservisima. Za razliku od SOA arhitekture, ovdje nema ESBa, već servisi komuniciraju direktno kroz svoja sučelja.

V. PREPORUKE

Mikroservisi su relativno nova paradigma arhitekture aplikacija i sve je više projekata u kojima se ona upotrebljava. Ako govorimo o integraciji servisa, onda su mikroservisi bolje prilagođeni za tu svrhu od SOA tehnologija. Osnovni razlog je taj da je kod mikroservisa integracija sastavni dio arhitekture aplikacija, dok se kod SOAe za to podrazumijeva posebna tehnologija. Iako je integracija jedan od osnovnih ciljeva SOA arhitekture, ona ne diktira neki specifičan način izrade aplikacija; one su najčešće napisane puno prije nego što se pojavi potreba za specifičnom vrstom integracije, pa se moraju koristiti posebni posrednički (engl. *middleware*) sustavi, kao što je ESB, kojima se te aplikacije (nakon prilagodbe) mogu povezati. Kod mikroservisa stvari su drugačije: aplikacije su same po sebi napravljene tako da se njeni dijelovi, to jest servisi, povezuju (integriraju), i to tako da su detalji povezivanja najvećim dijelom sastavni dio servisa, pa nema potrebe za korištenjem posebnih komponenti ili tehnologija samo za to.

Bez obzira na to što mikroservisi pripadaju jednoj aplikaciji, ne postoje nikakve tehničke prepreke da jedan mikroservis komunicira s drugim koji pripada drugoj aplikaciji. Na taj način mikroservisima možemo integrirati dvije ili više aplikacija direktno, bez upotrebe dodatnih integracijskih komponenti i tehnologija.

S obzirom da se radi o novoj arhitekturi potrebno je još kolektivnog iskustva da se pokažu sve prednosti (i nedostaci) tog pristupa. U tu svrhu potrebno je istražiti nekoliko stvari:

1. Kada su mikroservisi prikladni za primjenu, a kada nisu?
2. Da li su mikroservisi jednako primijenjivi i za male i za velike sustave?
3. Kakve je principe potrebno istražiti i razviti za nadgledanje rada servisa i osiguravanje od slučaja kada oni prestanu raditi zbog unutrašnje greške?
4. Koji sigurnosni mehanizmi su prikladni za ovakvu arhitekturu sustava (s obzirom da mikroservisi podrazumijevaju distribuirani programski sustav)?

Pored gore opisanih aspekata, potrebno je istražiti i podršku mikroservisima u postojećim platformama, kao što su Java i .NET, i programskim jezicima, s ciljem utvrđivanja moguće direktne podrške izradi aplikacija s ovom arhitekturom, slično kao što kod takvih platformi postoji direktna podrška objektno-orijentiranom programiranju [6]. Međutim, ovo bi bilo moguće tek nakon što se stekne bolji uvid i veće iskustvo u radu s mikroservisima, tako da se mogu utvrditi njihovi osnovni elementi koji bi trebali biti na raspolaganju i koji bi pojednostavnili implementaciju aplikacija s mikroservisima.

VI. ZAKLJUČAK

Povećanjem količine i vrsta podataka integracija informacijskih sustava postaje sve važnija, dok istovremeno takvi sustavi postaju veći i složeniji. Mogućnost efikasnog proširenja jednog takvog sustava novim modulima ili modifikacije postojećih modula može biti od strateškog značaja za organizaciju koja upotrebljava taj sustav. Iako se i SOA i mikroservisi temelje na komponentizaciji sustava, integracija komponenti je inherentna mikroservisima [5], dok je SOA više primjenjiva kao tehnologija za naknadnu integraciju aplikacija [1]. Stoga koncept mikroservisa može poslužiti kao polazište za istraživanje i razvoj jednog općenitog modela izrade i integracije sustava koji bi obuhvaćao šire područje aplikacija, a cilj toga bio bi razvoj metoda i tehnika izrade modernih informacijskih ili općenito programskih sustava.

LITERATURA

[1] Gregor Hohpe, Bobby Woolf, *Enterprise integration patterns – designing, building, and deploying messaging solutions*, Addison Wesley, 2003.

[2] Nicolai M. Josuttis, *SOA in practice – the art of distributed system design*, O'Reilly, 2007.

[3] D. Litan, M. Velicanu, L. Copcea (Teohari), M. Teohari, A. M. Mocanu (Virgolici), I. Surugiu, and O. Raduta, *Business' new requirement: information systems integration – methods and technologies*, International Journal of Computers and Communications, Issue 3, Volume 5, 2011.

[4] Esri Partners | SAP | SAP ERP Integration Options site, [Online]. Izvor: http://www.ESRI .com /partners/ alliances/sap/sap-erp.html

[5] Martin Fowler, *Microservices*, preuzeto 7/2015. iz http://martinfowler.com/articles/ microservices.html

[6] Galen Gruman and Alan Morrison, *Microservices: the resurgence of SOA principles and an alternative to the monolith*, preuzeto 7/22015. iz http://www.pwc.com/en_US/us/technology-forecast/2014/cloud-computing/features/assets/pwc-microservices.pdf

[7] Sam Newman, *Building microservices*, O'Reilly, 2015.

[8] ThoughtWorks, *Principles of microservices*, Sam Newman, Microxchg, Berlin 2015, preuzeto 7/2015. iz http://samnewman.io/talks/principles-of-microservices

[9] U. Batra, S. Mukharjee, *National Healthcare Information System Integration: A Service Oriented Approach,* International Journal of Computers, Issue 3, Volume 5, 2011.

[10] *Service Oriented Architecture Based Integration*, http://www.omg.org/news/meetings/workshops/MDA-SOA-WS_Manual/01-A1_Rosen.pdf

[11] *Fifteen Years of Service-Oriented Architecture at Credit Suisse*, http://www.infoq.com/articles/service-oriented-architecture-at-credit-suisse

TELECOM PRODUCTS, SERVICES AND MARKET

Restructuring of telco products

Ivan Vrbovčan, Tomislav Pavić, Marija Šoša Anić
Hrvatski Telekom d.d., Zagreb, Croatia
ivan.vrbovcan@t.ht.hr, tomislav.pavic@t.ht.hr, marija.sosa.anic@t.ht.hr

Abstract - In the daily struggle to increase revenues in telecoms, improving the sales process is one of the key areas in which changes are taking place. Through it telecoms want to achieve an increase in sales and longer customer retention. Products and their implementation in IT systems are one of the key elements in the sales process and sales process improvements will require redesign of current product definition.

In this paper will be given an overview of the current telecom product structures, the difficulties that telecoms are facing due to the product structure and suggestions for future improvements.

I. INTRODUCTION

In the situation of economic crisis in the market, telecoms are struggling to gain an economic advantage. One of the key area in which changes are taking place are improving customer retention and revenue growth, and they require improvement in the following areas:

- Simplifying the pricing, packaging and purchase process,
- Understanding the usage and customer behavior patterns
- Understand what drives customer requests
- Streamlining the on-boarding process in order to avoid customer frustration and high support cost [1].

Products and their implementation in IT systems are one of the most important elements in the sales process. They are communicated to customers to enable selling, using of goods and services, and at the end – earning of money [2]. As sales process is so fundamentally connected to products, it makes sense to analyze current product structure and their improvements possibilities in order to improve sales processes.

II. CHALLENGES WITH EXISTING PRODUCT STRUCTURES

When we are analyzing product structures in current telco IT systems, the first thing that we could notice is that each IT system has its own product structure. This is on one hand a consequence of the fact that each system is using products to achieve different function - for example products in CRM (Customer Relationship Management) system are used for order entry and customer assets representation, products (or services) in OSS are used for provisioning of services, and in the billing system products are used for calculating how much customer has to pay. So each system only contains a part of product catalogue relevant for its own operations. On the other hand, product catalogue distribution across systems is also a consequence of telco strategy to use COTS (completely of the shelf) and best-of-the-breed IT application. In this strategy, telecoms are buying on the market COTS IT applications for specific functions (CRM, billing, …), usually from different vendors. COTS applications have to function independently, because they are sold separately, they usually bring their own product catalogue (see picture 1).



Picture 1. Example of product catalogue distribution across IT systems

Product catalogue distribution across IT systems is one of the reasons for prolonged time-to-market in telco. When telco introduces a new product, it has to be configured in each IT systems separately, and after that it has to be checked that integration between systems for that product is working correctly. As the product configuration has to be entered in several systems, this also requires experts for different areas, usually from IT side. Some companies are using centralized product catalogue to manage products in one place and then propagate product definition to specific IT system.

The other challenge on the product structure is caused by increased use of customer self-care portals and mobile devices. Legacy IT systems are like CRM are usually primarily built for internal usage by internal employees. System design is focused more on functionalities and less on user experience the product presentment on the screen. This is not sufficient for customer self-care portals and mobile applications, where user experience has to be

maximized. As a result, customer self-care portals and applications on mobile devices introduces additional, enhanced versions of product catalogues, with simplified product structures suitable for presentment on small screens, product pictures and videos, reviews, evaluations, improved searching capabilities, etc. Additionally, customer assets presented on self-care portals and mobile devices sometimes differs from CRM view and billing view, which may bring communication problems between customer and telco employees. Introduction of new product on customer self-care channel and mobile applications requires additional work to configure related product catalogues. This on the other hand cause that only a fragment of the whole product portfolio is accessible through those channels, and that even more prolongs time-to-market.

To achieve flexibility in product definition, customer services in telecoms are built from several elemental products (installation, subscription, tariff option…) which are used as "Lego bricks". This enables telco employees to choose from existing components and compose some new elements from those building blocks. However, using such products in order entry to compose a final service introduce complexity and require additional knowledge. Even with a real Lego bricks, instructions are needed to build some more complex things like a house or a car. The IT system provides instructions for order entry and has to check rules which defines how components are entered in the system in the right way. All this are resulting in a complex order entry system. To even more complicate the situation, telco systems has to support a whole customer assets lifecycle: activation, deactivation, resumption, modification of attributes, upgrade, movement to other location, contract takeover, providing services in rented apartment, etc. The "Lego bricks" product model is not suitable for self-care, as it is too complicated for end customers.

Product presentation capabilities of legacy systems are usually limited. In legacy IT systems, usually, products and options are presented on screen with not so good user experience. Product catalogues in legacy IT systems does not have enough functionality to support advanced user experience in applications. That leads to prolonged order entry time and increased number of errors, prolonged education for employees to learn how to enter the orders into system and the need to introduce additional product catalogue components for self-care and mobile applications.

III. PRODUCT STRUCTURE IMPROVEMENT

As products are one of the key elements in sales process, improvements in the product structure will lead to improvements in sales process through maximizing customer experience on all channels, easier product ordering and better customer retention. This could be achieved through easier searching for the right product that will fulfill the customer needs and easier order entry and product manipulation with less errors. IT system architecture has to enable shortening of time-to-market through introduction of centralized product catalogue.

Centralized product catalogue should enable product design in one place and product data mastership management across systems. This system could be a separate system (sometimes this system is called Enterprise Product Catalogue) or could be integral part of pre-integrated systems (best-in-suite approach). Centralized product catalogue should at least enable entering of basic product characteristics in one place and propagating it to specific systems. This will significantly shorten time to market as this will decrease the need for intensive product integration testing between systems. Some product specific definitions (for example some product related business rules) could still be entered directly in a specific system. As product definition is entered only in one system, this will reduce the need for involvement of IT experts for specific systems, and partially shift the product configuration work in systems towards the marketing department.

Product visual presentation capabilities should be supported by product catalogue to ensure consistency across all channels. This consistency should also be reflected on customer assets. Products ordered through self-care channel should be the same as in internal CRM and on the customer invoice. Product structure should be intuitive, simple, and enable order capture on small screens like mobile phones.

Product searching, comparison and recommendation tools should be supported and consistent across all channels. In today world, customers are used to research for products they want to buy, to compare them and read reviews from other customers. On the platforms provided by telecoms such functionalities should be also supported because this will help customers to choose the right product.

Order capture should start from ready-to-use configuration and should enable easy configuration with additional options and attributes. To achieve that, product catalogue should go beyond "Lego bricks" approach, and to put focus on the customer perspective of using products. Of course, flexibility provided by existing product catalogues should be somehow preserved. Order capture rules should be radically simplified. The system should not only check that order is correctly entered, it should recognize what customer wants to achieve and to manage that. For example, customer wants to buy a 3play bundle, and he already 2play and iptv service. In this case, customer only has to enter the order for 3play service, and the system has to recognize that he already have 2play and iptv and manage all necessary changes on existing services.

As telecoms are starting to experimenting more with new business models that include external partners (sell to, sell through, sell with models), this should also be supported from the product catalogue perspective. Product model has to provide all appropriate information needed to describe contractual relationship between telecom and his partners, and their roles for product provisioning and support [3].

IV. CONCLUSION

In this work we have presented analysis of existing telco product structures and proposals for product structure improvement. Increased orientation to self-care and mobile devices fundamentally changes order capture process and put new requirements on products in product catalogue. Visual representation of assets and products is becoming more important. It should be simple, intuitive and to provide help in choosing the right product for a specific customer.

LITERATURE

[1.] P. Kelly: Customer experience management: value-based delivery and service support : Analysys Mason, 2014.
[2.] TM Forum: Information Framework (SID) – Product Business Entities, Release 15.0.1., Morristown, NJ, USA, TM Forum, 2015.
[3.] TM Forum: Online B2B2X Partnering Step by Step Guide, Release 15.5.0., Morristown, NJ, USA, TM Forum, 2015.

Moving from Network-Centric toward Customer-Centric CSPs in Bosnia and Herzegovina

Nataša Banović-Ćurguz* and Dijana Ilišević *
* Telekomunikacije RS A.D., Banja Luka, Bosnia and Herzegovina
natasa.banovic-curguz@mtel.ba, dijana.ilisevic@mtel.ba

Abstract – Telecom sector is in the middle of a global transformation. Drivers for this are changes in customer behavior and in the telecom industry itself. Today CSPs need a more sophisticated, real-time view of performance that provides a deeper understanding of how customers are using and experiencing specific services. Moving from network-centric toward customer-centric operations and introduction of Customer Experience Management (CEM) are important steps in that transformation process. In this paper we have provided short overview of current state in performance & customer management segment for CSPs in Bosnia and Herzegovina. Afterwards, we have presented the review of necessary steps in the transformation journey from network-centric to customer-centric solution with the Service Operation Center (SOC) as enabler of CSP's CEM strategy.

I. INTRODUCTION

Communication service providers (CSPs) are facing with the crucial changes in the dynamics of the market: constantly increasing amount of mobile devices and applications that are being released almost daily, the explosion of overall traffic demand and changes in customer behavior and expectations. Today, customers are mobile, connected and informed. They can choose from many operators and can quickly and easily change their minds. Most CSPs still rely on network metrics provided by Network Operations Center (NOC) to understand how the network is behaving and monitor the end-user experience. But CSPs need to fix issues in advance before customers report them, which requires applying new predictive analysis techniques to avoid losing valuable customers to competition. In the future Quality of Experience (QoE) will be the most important factor affecting customers' loyalty towards CSPs. That is the reason that many CSPs are moving from a network-centric view to a service & customer - centric view of the network, Figure 1. This process is a step towards improving business Key Performance Indicators - KPIs (customer satisfaction, brand loyalty, reductions in churn, reductions in opex, faster time-to-market for new services and increases in service revenue). Service assurance investments are focused on improving customer experience more than ever before. The transition to Service Operations Centers (SOC) is creating a demand for unified service assurance solutions that combine performance, fault and service management functions.

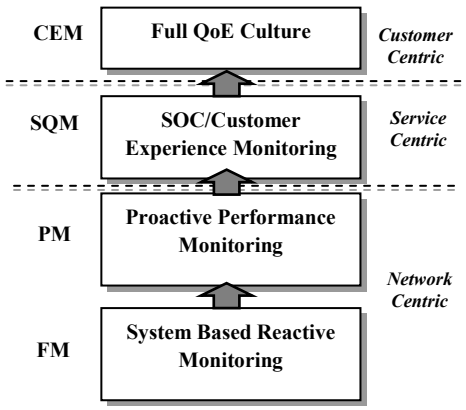


Figure 1. Transformation Process from Network-Centric to Customer-Centric CSP.

Customer experience has a direct impact on user loyalty and revenue and will decide the future of telecom services. Some progressive CSPs are already beginning to monitor and control end-to-end service quality and influence the consumer's perception of their services by implementing Service Quality Management (SQM) and Customer Experience Management (CEM) processes, [1].

II. CUSTOMERS & SERVICE ORIENTED ASSURANCE SOLUTION

A. Current state in B&H

In Bosnia and Herzegovina there are three national operators licensed to provide GSM&UMTS services: BHT Sarajevo, Mtel Banja Luka and HT Eronet Mostar. The global changes in telecom sector are reflected on the market in B&H, [2]. Even in conditions of complex political structure of B&H with weak economic growth and very specific market, official data of Communication Regulator Agency shows that there is a potential for growth, [3].

CSPs in B&H are too network-centric:

- Network operation departments have been traditionally more focused on the network infrastructure than on the customer.
- They use traditional network oriented operation and maintenance (OAM) concept with Performance Management (PM) and Fault Management (FM). The first one deals with

network performance metrics (KPI), and the second one – with network outages. But, as both systems manage the network only from the resource perspective, they can't directly provide information about the impact on customer services.

- Key performance indicators (KPIs) are based on alarms, performance and network configuration data. But, high network KPIs do not necessarily mean good QoE.

CSPs are focused on Customer Relationship Management (CRM) and Customer Care:

- CRM is in charge of managing the company's interactions with customers.
- The methodology employed by CRM mainly focuses on sales activities, marketing, and business procedures, not taking into account the customer needs on the service quality. As a result, there is a large gap between the approach of the company and customer's expectations.

Also, CSPs in B&H have not yet implemented automatic diagnosing of customer complaints.

- They are not equipped to handle service quality holistically across all network components, or to ensure the levels of service quality expected by their customers (no Service Level Agreements - SLA).
- They do not include third-party capabilities (downloads, content, applications, interconnection).

The fact is that the network performance is only a part of the service quality story. Customers are willing to increase business with companies in exchange for better customer experience and this is the real opportunity for companies in the region, [4]. In this context, there is an urgent need for CSPs to change the way they compete for new customers and revenue, and attempt to retain existing customers. Customer experience management has the potential to overcome the above listed limitations.

B. Service Quality Management

Customer experience is defined as the sum of all experiences a customer has with a company, over the duration of their relationship [5]. New processes have to monitor quality of experience based on individual customers. Quality of experience can indicate network and service experience and it can show the gap between the actual and expected quality of service and network. Many authors have researched quality of experience and proposed their own measurement criteria, models and design challenges, [6] and [7]. In general, the top-down CEI-KQI-KPI modeling approach is used in telecom sector. In this model KPI is a key performance indicator for a network or network elements; KQI is a key quality indicator for different services and it is closely related to user experience and CEI is a customer experience indicator which provides an objective measurement of customer experience. The traditional network/resource perspective, typical for fault and performance management systems, needs to be transformed. This process aims to translate network faults and metrics into customer service issues and metrics. A customer service assurance solution is given in Figure 2, [8].

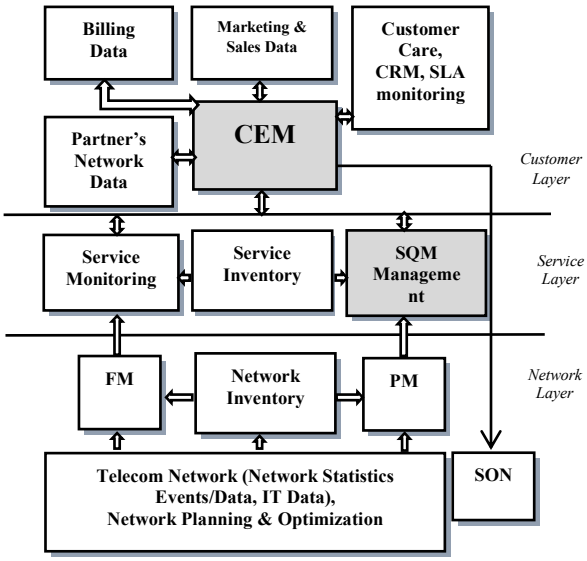


Figure 2. Customer service assurance solution, [8].

Service quality management (SQM) is the proper way to transit to CEM. It is the way to model and design a truly end-to-end service toward the purpose of understanding the user's experience consuming the service. SQM correlates network events with service performance metrics. SQM is enabling customer experience issues to be tracked down to the cause of network performance problems, which is essential for end-to-end customer experience management. Existing FM and PM modules in a CSP's IT environment can be used for such a transformation, with no need for a huge investment. Service Monitoring (SM) and SQM modules are integrated based on the concept of threshold alarms, generated by SQM modules, which feed the SM module. This way SM has a comprehensive view of service alarms. Network and Service Inventory modules provide service modelling information needed to align the resource domain with the customer service layer. The CEM module collects CRM-related data, end-to-end probing data, marketing & sale and billing data to assure that service alarms and KQIs can be trusted. SQM has an early detection capability. When a PM system detects a network performance problem, SQM calculates its impact on the customer service and initiates restoration before the customer actually experiences a problem. This may significantly limit customer complaints or at least shorten the time needed to fix a service, after the customer has reported a problem.

C. SOC as enabler of CSPs' CEM strategy

CSPs are evolving from a network-oriented view of their operations (Network Operations Center - NOC) to a more service oriented view (Service Operations Center - SOC). Typically, the SOC collects real-time data flows

from different sources, correlates data based on rules and service models to calculate KQIs, identifies the root cause and location of problems that are directly affecting services, coordinate measures to improve quality of service for the customer identified through CEM techniques. Also, the SOC collects information from other business departments, such as the call center and marketing, and feed information back to these departments to improve the customer experience. The SOC provides a real-time view of how customers are experiencing each service. In this way, operators can take a more proactive approach to providing a great experience, especially for the services and the subscribers that generate the most revenue. The SOC provides a greater understanding of how customers are using mobile services in general, which applications are most popular and how behaviors are evolving over time.

CSPs will adopt a mixture of approaches to introducing a SOC:

- Establishing an integrated SOC with an already developed Network Operations Center (NOC);
- Creating an independent one;
- Outsourcing the running of a SOC or keeping its management in-house;
- Using a SOC only for certain services or customer segments.

A CSP's strategy will depend on its maturity and size, attitude toward outsourcing and need for a speedy transformation. Some CSPs view the SOC as a natural evolution of the NOC and are integrating it directly into the NOC, [9].

III. PROPOSED CONCEPT FOR CSPS IN B&H

A. From NOC to SOC

The best way for shifting from network-centric to customer-centric organization is step-by-step transformation of existing NOC. Transformation process could be based on augmenting the existing PM and FM systems with Service Quality Management (SQM) and Service Monitoring System (SM) respectively in the service layer. The calculation of the network alarm's impact on customer services and translation of technical network metrics (KPIs) into customer service metrics (KQIs) should be enabled. KQIs can also be used to define service level agreements with customers. SOC is the most important step in a CSP's transition towards a customer-centric organization, and should be at the heart of a CSP's customer experience management strategy. It is an organizational department that monitors the quality of the overall service and takes the necessary actions in the case of service degradations and outages, [10].

The operators in SOC should continuously monitor SQM and coordinate with other departments. Key to the effective monitoring of any service is the integration of a broad diversity of raw data available to the CSPs:

- Performance data from network level Performance Management systems;
- Customer usage data (CDRs, xDRs);
- Network level Fault Management systems data;
- Passive and active probes for signaling monitoring data;
- Systems log files (e.g. from SMSC, MMSC or IT infrastructure);
- Blackberry enterprise server data (e.g. via SNMP and log files);
- IT server log files and application agent data sourced from the IT cloud;
- CRM systems data;
- Billing systems data;
- Product and service catalogue systems;
- IT and network inventory systems data;
- Service delivery platforms (SDP) data;
- Content management systems (CMS) data.

SOC should provide the capability to monitor services in an end-to-end context, enabling CSPs to take actions based on their impact on services and the customer experience. It may be deployed as a part of NOC, or as a separate entity, but in either case the SOC must be established at a higher level in the organizational hierarchy. An SOC plays the role of an intermediary between the customer-facing departments such as customer care and marketing and network-focused and back-ends teams such as network operations and network engineering, planning and optimization, Figure 3.

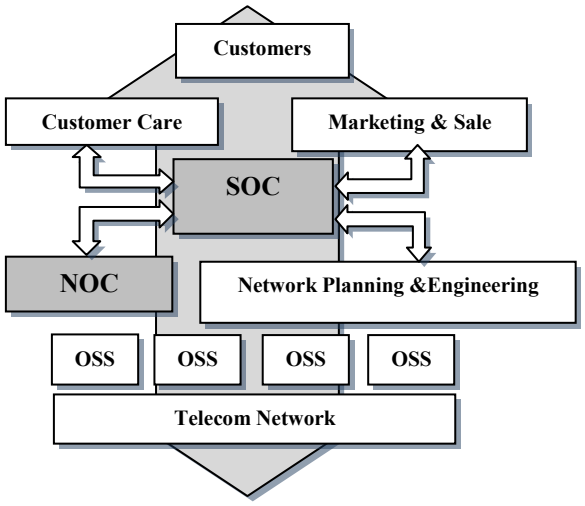


Figure 3. Service Operation Center, [11].

The people in SOC should include experts with skills in networking, IT and other necessary topics to streamline the troubleshooting activities. SOC should breakdown organizational barriers and encourage inter-departmental collaboration, transforming CSP into a highly agile and responsive organization. It enables CSPs to focus important resources, care processes and network improvement on high-value customers such as VIPs and enterprises.

A CSP should implement its SOC in phases, [11]. At every phase, the transformation should involve the overall organizational structure, processes, people and systems. In the first phase of SOC implementation, CSPs should deploy SQM for providing near-real-time service-level visibility. But in that phase, NOC is still central point for O&M activities, service management is still reactive and the correlation between service and network is established through service models and KQIs. In the second phase of SOC implementation, process changes shift the focus of operations from a network oriented reactive model to a service oriented proactive model. The customer-level visibility exists at all customer related touch-points. The CSP can use and share analytics data through inter-department interfaces and processes. Customer care, network operations and planning departments can access customer and network analytics data to prioritize tasks based on their service and customer impact. The CSP can analyze service-inventory and service-quality degradation information to proactively increase capacity in the network before it affects the customer experience. Most CSPs are either at the first or the second phase of this transformation journey. The move to the third phase will take some time because this requires a significant change in a CSP's approach to operations with processes that are started from the customer.

SOC implementation requires not only changes in system architecture and technology, but also changes in company culture, organization, process and people. As it is given in [11], a successful implementation will need to incorporate a number of principles:

- The development of key business objectives that drive customer experience improvements.
- The use of a plan that addresses organization, processes and change-management systems, SQM, network data collection systems, analytics, visualization and process automation systems.
- The adoption of an approach to transformation, to ensure alignment of business objectives, acceptance of changes and retraining of employees.
- The development of company culture where the customers have a priority position in the corporate set of values.

B. CEM consideration for CSPs in B&H

Introduction of CEM is a complex organizational transformation for all CSPs in B&H. The complexity of CSP's organizations is in fragmented systems, multi-vendor environments, multiple service/network layers and regulatory obligations. All of them present numerous challenges to CEM introduction. These challenges can be broadly grouped under the categories of defining, measuring and improving customer experience. Moving from NOC to SOC, predictive analysis, innovative billing plans, Voice of Customer (VOC), customer call back programs are parts of an effective CEM solution for telecom operators. Customer experience management means handling not only the technology and the people involved, but the processes as well, [12].

CSPs in B&H need to start looking at the network from the viewpoint of the quality of services delivered to the end customer. They need to identify which customers, products and channels are the most profitable and determine what changes would increase that success. When compared to the traditional approach of collecting vast amounts of system logs and data points, and then attempting to correlate this vast amount of data, carefully defined measurement and monitoring enables operators to be proactive and take corrective action before customers even notice. This approach makes it possible to identify potential revenue threatening situations as well as network abnormality. CSPs need to control all interaction channels along the customer lifecycle, from service delivery to post-selling and complaint management. The priority must be improving the communication flow between the technical department and the various business functions. Customer experience improvement program must be driven across the whole organization and involves multiple different processes, systems and domains. In according to [12] this can be achieved by making the technical department a key enabler in the process.

IV. CHALLENGES WITH IMPLEMENTATION

A. Barriers and business benefits

Moving from NOC to SOC is not an easy transition. The key challenges in that transformation are given in Figure 4, [9].

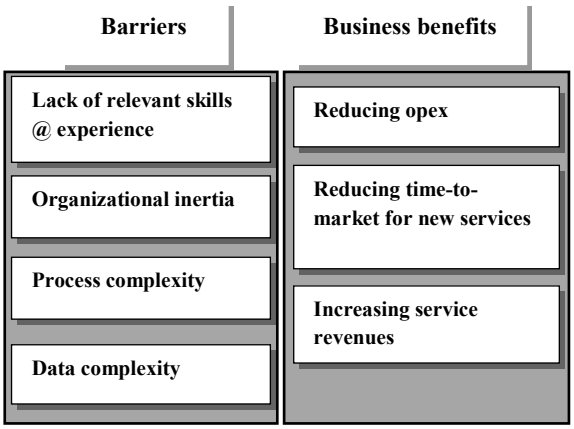


Figure 4. The key challenges in NOC to SOC transformation.

As the barriers in that process authors in [9] indicates next:

- Lack of relevant skills and experience in service modeling, service monitoring, interpreting alarms and correlating service quality data.
- Organizational inertia that includes high resistance to organizational changes, worries about job security, career development etc.
- Process complexity that means high risks in integration of new service quality processes with

network oriented processes, because SQM and CEM tasks are very complex.

- Data complexity that includes collecting, analyzing, correlating, understanding the complex relationships between huge amounts of service quality related data and taking actions in real-time.

Regardless of the above mentioned challenges SOC introduction gives important business benefits:

- Reduced opex (operational expense) – using the SOC to prioritize site visits and repairs on the basis of faults affecting services and customer experience and through proactive, planned maintenance.
- Reduced time-to-market for new services – KPIs and KQIs incorporation into service design, reduction the time for trials and redesign services.
- Increase service revenues – using the SOC offer opportunities for upselling and churn reduction, based on a rise in the customer's confidence in the CSP's quality.

Since the establishment of a SOC will be a challenge for CSPs, they must:

- Develop or acquire appropriate in-house skills, resources and tools in partnership with vendors who are well placed to help, given their knowledge, experience and ability to integrate SQM/CEM solutions.
- Put in place a strong set of processes that involves all relevant departments.
- Ensure that KPIs and KQIs are in accordance with business objectives.
- Ensure that performance metrics for personnel are aligned with customer-experience/service-quality metrics.

B. Segmentation

Monitoring all the customers in the organization is not feasible. That is why CEM initiatives rely on segmentation. An overall enhancement can be applied to all of the customer interactions. However features like real-time monitoring and analytics will require additional resources. Because of that, the most service providers who implement CEM start with enterprise and VIP customers. Customer Lifetime Value (CLTV) is using in CSPs' CEM strategy for better segmenting the customer base. By analyzing data consumption patterns and identifying customers who contribute the highest ARPU, CSPs can create personalized initiatives to prevent them from churning. Enterprise, VIP and roaming customers are significant sources of revenue for CSPs. By monitoring even simple failure scenarios, CSPs can obtain valuable insights which will help them find & resolve the problem.

CSPs should develop initiatives specifically for high-CLTV customers, [11]. Such initiatives could include: planning journeys and monitoring the touchpoints for customers with high CLTV to provide a consistently superior experience, monitoring the quality of experience in real time, monitoring the quality of experience of VIPs/VAPs, assessing the geo-location of network congestion hotspots where high - ARPU customers are concentrated, communicating with customer care and marketing department to set up dedicated care processes, deal with incoming complaints, give the care personnel the authority to proactively inform customers and recommend changes etc.

C. Expectations trends

In the upcoming period, customer experience will continue to grow in importance for CSPs. In this environment, as it is shown in Figure 5, in [13] is expecting next:

- Culture change intensifying because customer experience is a reflection an organization's culture and operating processes.
- Customer journey designing has become one of the most popular customer experience tools as it helps provide a customer-oriented viewpoint.
- Mobile continues to become a more dominant interaction channel and through increasingly varied types of devices and new smart applications.
- Predictive analytics personalizing.
- Metrics to action realigning. It is expected an increasing number of companies to shift their emphasis from tracking metrics to enabling action.
- Employee customer experience & empathy training.

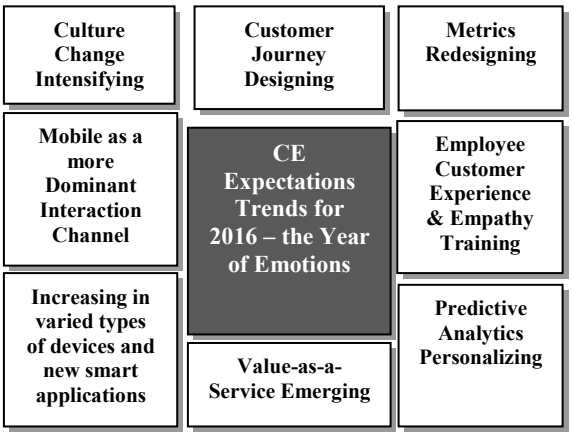


Figure 5. CE expectations trends, [13].

As we have seen, the transformation from NOC to SOC involves complex real-time monitoring of large amounts of data, collecting data from different data sources, their processing and correlating to provide a service-level view of the network, data mining over huge data sets and implementation of Big Data solutions to effectively manage a SOC. Also, a customer service assurance solution also enables to drive the

implementation of a Self-Organizing Network (SON). The idea is a model shift in network management towards network automation and its autonomy. It is essential that SON is driven by customer experience goals to really help CSPs in improving their business and reducing costs.

D. Case study of SOC implementation

A case study of Telecom Italia's SOC in which the operator is linking CEM with a service-level approach to managing its mobile network is given in [1]. Through this project they wanted to improve network monitoring & troubleshooting and to drive more customer-oriented quality management and trend analysis. The most important steps in process implementation were creating service models and the supporting data collection architecture. In the process of correlation of all this data, they use about seven hundred rules to update the service models and refresh their KPIs. Service models attributes include location and customer/offer type.

As the benefits of SOC implementation they recognize the positive returns from segmenting and dealing proactively with VIP customers, because SOC detects when such customers experience poor service performance and prioritizes their remediation process. Telecom Italia now has end-to-end service-level visibility of the network, helping operator to reduce management complexity and opex, and also to improve customer satisfaction and better targeting of network investment. Improved knowledge of customer usage of the network has led to a 75% reduction in customer complaints due to a major fault and increase network availability by around 20% in individual cells, [1]. In the upcoming period this operator will extend SOC concept to fixed network services and as a priority to model OTT (Over-The-Top) services so that it can monitor from SOC and analyze OTT-related customer experience.

V. CONCLUSION

As customer expectations and competitive pressures rise and as profit margins become smaller, there is an urgent need for CSPs to transform themselves from network-centric to customer-centric businesses. Customer experience will decide the future of telecom services. It is a top strategic priority to differentiate operators in order to have a competitive advantage. The introduction of SQM/CEM techniques managed via SOC is an important intermediate step and an enabler in the CSP's journey towards becoming a customer-centric organization. This transforming will be challenging for CSPs in B&H due to a lack of relevant skills, organizational inertia and technical complexities. Integrating CEM efforts throughout the organization will require a shift in company culture and the way of working.

REFERENCES

[1] C. Chappell, "Managing customer experience of the network: Strategies for Success", http://www.nsn.com, May 2014.

[2] D. Ilisevic and N. Banovic-Curguz, "Changing Face of Mobile Service Providers in Bosnia and Herzegovina", Proceedings of Telfor Belgrade, November 2014, pp. 51-54.

[3] "Annual report of Communication Regulator Agancy for 2014", CRA B&H, available http://www.rak.ba, 2015.

[4] N. Banović-Ćurguz and D. Ilišević, "Effective Strategies for Implementation of Customer Experience Management in MSPs in Bosnia and Herzegovina", Proceedings of MIPRO 2015/CTI, May 2015, pp. 429-434.

[5] A. Morais and other, "Managing customer experience through service quality monitoring", Conference Proceedings of Future Network and Mobile Summit, Berlin 2012.

[6] E. Liotou and other, "Quality of Experience Management in Mobile Cellular Networks: Key Issues and Design Challenges", IEEE Communications Magazine, July 2015, pp. 145-153.

[7] A. Cuada and other, "A global customer experience management architecture", Conference Proceedings of Future Network and Mobile Summit, Berlin 2012.

[8] L. Mendyk, "Moving from network assurance to customer service assurance", www.telecoms.comarch.com

[9] "The future role of the Service Operations Center", http://www.telecomsmarketresearch.com/, 2013.

[10] N. Banović-Ćurguz and D. Ilišević, "Towards Customers & Service Oriented Mobile Service Providers", Proceedings of Telfor Belgrade, November 2015, pp. 51-54.

[11] A. Rao and S. Ravindran, "How a service operations centre can enable the transition to a customer centric organization", www.analysismason.com, April 2015.

[12] V.Basile, J. Portal, A. Schlautmann and L. Dadhich, " Maximizing quality of experience", Arthur D. Little, 2013., www.adl.com/improving_cex, December 2014.

[13] https://experiencematters.wordpress.com/2015/12/15/11-customer-experience-trends-for-2016-the-year-of-emotion/, December 2015.

Future Communication Model: Challenges and Opportunities for Society as a Whole

D. Ilišević and N. B. Ćurguz
Telekomunikacije RS a.d., Banja Luka, Bosnia and Herzegovina
dijana.ilisevic@mtel.ba, natasa.banovic-curguz@mtel.ba

Abstract - Transforming into digital society requires the participation of a society as a whole. B&H is under pressure to accelerate social and economic development and its competetivnes. A key element to this process is broadband access. Broadband (BB) access is the key element to bridging the physical economy to the digital economy. Data center, cloud services, Big Data, Internet of Things (IoT) and Machine to Machine (M2M) are tehnologies that represents the targets in order the most efficiently transform economy for the digital future. Digital future comes with new challenges and first is to redraw the existing rules and bridge divides to create benefits of Information and Communications Technology (ICT) society. This article presents intelligent future and give roadmap to accelerate the adoption of digital services.

I. INTRODUCTION

Before fifteen years ago, ordinary phone calls were the basic remote communication between persons. What happened? SMS text messaging was the first change and ironically, text start to be alternative to voice. (Phone calls must be answered immediately, and text could wait and giving the user possibility to manage its time to answer). With smartphones and increased usage of Internet people can surf, watch, listen to, play and perform other activities.

Telcos spent its first phase providing voice calls and made a good profit on it. Long Term Evolution (LTE) is the first network generation designed with new user on mind, making it more suited for new communication model. New technologies (VoLTE) will help maintain voice as a one of sources for telcos revenue, but it will never restore voice as a key profit source. In today`s information society, it is important to be online, and this possibility will greatly improve people`s life quality, working efficiency. BB access is imperative for humanity and represents social issue of our time, and the UN has declared Internet access a basic human right.

The emergence of Big Data, M2M and IoT shift human centric communication to machine centric communication. The coexistence of human centric communication and machine centric communication will promote `stepping in the screen` model and to have the best of both worlds (real and virtual) since real world is no longer holding our attention like it once did.

II. THE FUTURE COMMUNICATION MODEL

Digital representation of reality will become important as reality itself and let people exist in parallel realities. Today, digital realities can matter as much as physical reality ever did, opening new teritories for new business model, and redefining the existing reality with new digital model. Most important questions in future communication model will be:

- Is real world having competitor, and can real problem be solved with virtual solution?
- Reality, how can I tweet that?
- New normal, where are the boundaries?

These questions are more focused on marketing of future digital society, but come with five dimensions of change, and every participants in this process must secure its role in future communication model.

A. Tehnological scenario

With the increase in traffic, content and applications, connectivity has become part of our work and life. From technology perspective competing standards requires much more complex networks scenario. The communication industry is approaching point where difference between fixed access (fast and reliable but static) and wireless access (slower and less reliable but flexible) are disappearing. In mobile world, existing 4G move toward 5G and LTE is converged in single LTE Advanced standard. It is possible to transmit data over high speed wireless networks (3G, LTE and in the future 5G) allowing Telco's to use wireless technology exclusively for access and fixed fiber for regional super networks, [1].

The future 5G networks are defined as ubiquitous, ultra-broadband with a true revolution in ICT field. With increasing complexity and associated costs, several concepts and technologies becoming relevant, currently working under European Telecommunications Standard Institute to define requirements and architecture. The future network infrastructure have to support complete redesign of services and service capabilities, architectures, functions, access as well as connection security.

The important challenges that must be addressed by 5G networks is higher capacity, lower E2E latency, massive device connectivity, reduced capital and operation cost and consistent QoE provisioning, as given in Figure 1.

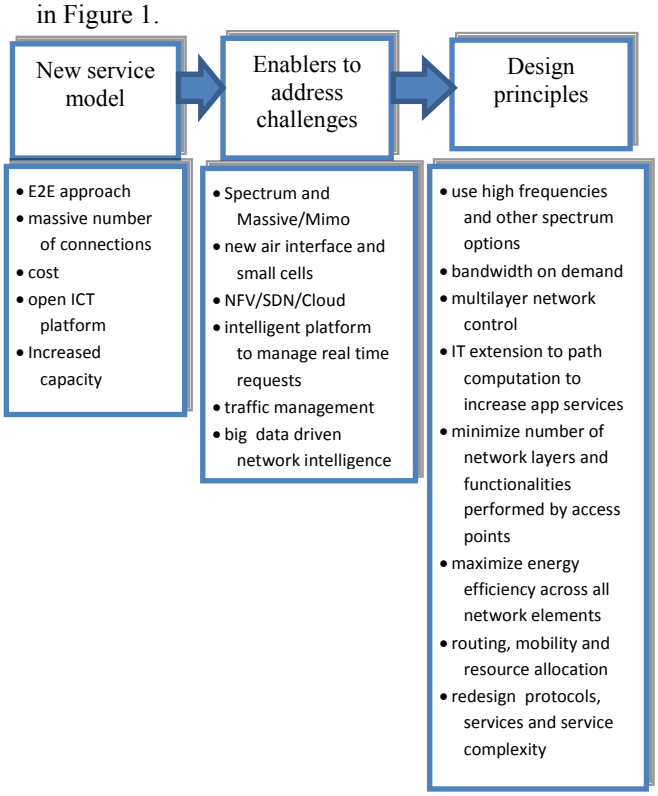


Figure 1, 5G Challenges and design principals

The main concept includes ultra-dense small cell deployments, intelligent use of network data to facilitate use of network resources for QoE for provisioning and planning and Network Function Virtualization (NFV) and Software Defined Network (SDN) to provide flexible network deployments, [2].

It is expected that this "5G" infrastructure will enable many new services in domains such as intelligent traffic systems, public safety, automation, e-health, transport, and logistics, and will help them to become economically viable. This trend will not only increase the overall spectrum demand, but will result in additional requirements on spectrum resulting from the high coverage, reliability, and availability requirements of the new services to be supported. Finally, a complex landscape of spectrum availability and access will emerge where multiple frequency bands, subject to different regulations including various forms of shared spectrum, are expected to be available to wireless communication systems. The pervasive and exponentially increasing data traffic present imminent challenges to all aspects of wireless system design, such as spectrum efficiency, computing capabilities, and fronthaul/backhaul link capacity. Review of the networking architectures and signal processing techniques adaptable for managing big data traffic in wireless networks is given in [3].

IoT will provide a lot of opportuinities to users, manufacturers and service providers in many sectors: environmental monitoring, health care, smart home and workplace, transport, security and surveillance and many others. In [4], three stages have identified that involve increasing levels of social involment of the objects composing IoT. *Res socialis* capability of objects lies in the fact that the social networks of objects are established between objects that are owned by humans who may have no connection with each other (Capability of building their own social network and Proficiency in building added –value complex services through collaboration in the object social network).

B. Human scenario

Most important is to start with analysis by looking at users of Future Communication (FC) model and the advanced services they are likely going to be used in the future (approach from service point of view or anything as a service). Once we understand demands and driving factors, we have possibility to identify the requirements for future communication model. Furthermore, we need to appoint what will become new reality in digital world? Every day, physical and digital worlds grow more connected, erasing the lines between online and off. A meaningful portion of our daily lives is being sent, trensferred, received, and stored in binary format thanks to anywhere-anytime online BB access.

How many realities are today? Virtual Reality (VR) refers to the creation of a completely artificial environment through digital means and with no connection to physical world. Another VR refers to digital reality that brings real world closer – remote reality. With high quality digital display, representation of real world objects through digital connection enable `stepping into the screen`. One of important use of VR will be in helping people better handle real worlds, both in terms of pathological or sociological conditions (digital representation of reality - fear of flying and creating empathy). VR is often considered as an escape or at least avoidance of the real world. Online social networks allows people using computers or mobile phones to continuously stay in touch with friends, relatives and other acquaintances wherever they are in the world, as long as there is access to the Internet. All of these realities have roots in what is already happening today.

The biggest challenge in FC model is that information is expected to be free but entertainment is surprisingly something that people are willing to pay for. Most people associate Internet and BB access with easy access to information, music video or in general entertainment and access to platforms, personal communication services like social networking. From the perspective of user the main demands are: personalized services, privacy, anonymity, software as service and improved availability, reliability and security. But in a new model (Content+Services+Managament), it is often trend that of users taking role of content and even service providers, and new security management standards must be established, [5].

In that case, user will be responsible for the security for customer and security/privacy issues must be offered by specialized personnel it is important to user to have control of private information, and possibility to track them. Service provider may be located in a different country and user must be able to redraw the usage rights at any time.

Online social networking (OSN) has become extremely popular all over the world. Use of OSNs from service perspective is important as well as from network perspective especially in mobile environment. Development of mobile platforms and application have important role in in understanding user behavior. Research in this area can guide Telcos to do action like adding traffic optimization in network elements and increasing user experience, [6]. How could a network become socially aware with knowledge of communities to facilitate efficient sharing of content such as video? Overview on the estimation methods for similarity in terms of demand, socialization, and mobility in defining and implementing a virtual community, that is, a socially aware mobile multimedia community, for sharing and delivering multimedia content is presented in [7].

C. *Business scenario*

In digital society is very difficult to predict how the competitive landscape will play out. Relatively rare phenomenon has now become a regular occurrence. Innovations that harness new technologies or business models, or exploit old technologies in new ways, are emerging on an almost daily basis (ankle-biters, sharing economy- UBER, HourlyNerds, Parkatmyhose etc.). The boundaries between industries are continuing to disappear, as experts in one sector apply their expertise to others – bringing previously separate industries together and in most cases redefining the way in which they're classified. These new competitors typically target a key part of the value chain, bypass the current business model, and making the existing rules irrelevant.

Technology is moving so quickly, and in so many directions, that it becomes challenging to track, and application concept convert mobile and similar devices into completely new and innovative tools that have potential to transform big business. In the same time, technology becoming greatest ally and greatest enemy, and building enterprise of the future will be conceivable with business leaders understanding influence of technology on new business model. The key advantage lies in understanding the new opportunities and threats before others.

The digital economy refers to an economy that is based on digital technologies (digital technologies as a sum of information, computing, communication and connectivity). Digital technologies are fundamentally transforming business strategies, processes and key relationships in business strategies. Accordingly, we identify that is time to rethink the role of ICT strategy (based on FC model) as a fundamental source of value creation in successful business strategy. (Apple has redefined the mobile entertainment model and Amazon has redefined the book selling system. In these two examples, there is no difference between business and digital strategy).

D. *Telcos future*

The development of the telecom sector has experienced a major process of transformation in terms of its growth, technological content, and market structure. With liberalization and subsequent demonopolization in the mobile telecom service segment the market scenario changed from that of single-player to multi-players. More than ever telecommunication sector need flexibility since is very difficult to predict the future, [8]. Change in business has been an either-or proposition: either quickly create economic value for shareholders or patiently develop an open, trusting corporate culture long term. But there is brutal fact: 70% of all change initiatives fail, [9].

Heavy investments into network maintenance and upgrade are not the only way to improve network performances and services, Figure 2. Another option is forcing more effective internal processes and their management in terms of monitoring, controlling and predicting service levels along all touch points in the customers' lifecycle, [10]. As technologies and services advances, networks are complex than ever before. Managed services have become possible solution to reduce O&M costs.

Network centric management can no longer accommodate demands, as Telcos are more concerned with service quality, user experience enhancement, revenue growth and value creation. For Managed Services Projects (part of larger Managed Services Provider organization), distinctive competitive advantage could be the balance between cost efficiency and effectiveness in meeting the Telcos network and customer needs. Careful organization design of Managed Services Project will contribute to establish a viable solution to achieve distinctive competitive advantage, [11].

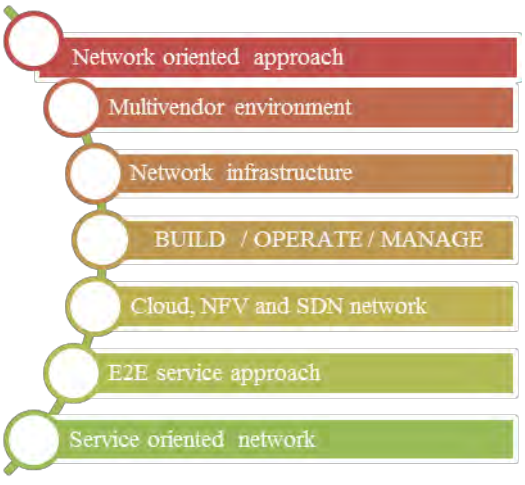


Figure 2, From network centric to service centric operator

Telcos will have to enhance their change management skills so that they can more effectively anticipate and manage continuous change. That means they will have to enable employees to participate in the innovation agenda and build dynamic business organization, to support new business models and new ways of working.

They will also have to manage risks more effectively. One way of doing this might be to create a modular organizational structure and isolate new ventures from the rest of their operations to minimize the effect of any failures. Collaborating with other companies would help to mitigate the financial and technological risks associated with moving into new areas of business, as well as stimulating innovation.

Telcos occupy a privileged position in the growing digital economy and must evolve in parallel with infrastructure to realize maximum efficiency. There will be a profit shift in the value chain, with the move from product to total solution. Furthermore, FC model is not new dimension; it is a new world of opportunities. For Telcos to succeed, new business model must be established, and it is also possible that OTT providers and network providers will need collaboration models where both parties can benefit commercial from these FC services.

E. Regulation scenario – government role

Broadband access will become new element of production alongside capital and technology. Challenges of ensuring a success deployment of future digital infrastructure are significant just as potential to improve economic development and quality of life. Government has the power to limit the adoption or progress of technology.

More importantly, society at large will comfort change on many fronts: in the way business organized themselves, how jobs are created, how technology is used to interact with world. Technology continues to advance on many fronts, and many developments bring new challenges for society and government. Governments cannot afford to be passive or reactive. Policymakers need to identify potentially disruptive technologies and consider their potential, before these technologies begin to display their disruptive powers in economy and society. Government should lead the way and must be more specific in pushing for development. E- Government is the simplest way and major catalyst for promoting BB access and improving digital literacy.

The Strategy for digital transformation must be carefully planned and executed. Spectrum, Digital Dividend, Digital Divide, education, health care, business, protection to children and improve care for elderly and people with disabilities must represent the targets to focus in order to most efficiently transform national economy for digital future.

Measures of economic activity such as GDP are important, but better efforts can be made to measure aggregate societal and economic advancement. Digital Economy and Society Index (DESI) includes five main dimensions: Connectivity, Human Capital, Use of Internet, Integration of Technology and Digital Public Services, [12]. New metrics (DESI) can influence policy decisions, so policymakers should select them carefully and measure to improve ICT development.

III. ROADMAP TO FC MODEL

Digital agenda for EU is given in Digital Agenda in the Europe 2020 strategy, [12]. War past, economic state and very complicated political arrangement represent bad environment to build digital society in B&H. Progress in digital divide represent important step to promote new face of B&H and the right conditions for an effective use of ICTs boost innovation, competitiveness, and greater social inclusion. The coexistence of human centric communication and machine centric communication will promote `stepping in the screen` model and every product will come with `virtual` layer such as information or services.

Government must began to study how to adapt B&H society into digital society (apply Digital Agenda). In addition, government has to meet the goal by redefining the responsibilities and include new participants. Academia, as needed new participant, should gravitate to make ICT benefits understood and to facilitate the spread of ICT services. Academia and policymakers needs to offer customized roadmap and services tailored for developing country like B&H (at affordable prices). In the area of Industry and Standards, Academia and Government should be important leader of FC oriented model, and quickly adopt its influence in new converged ICT fields. Roadmap to accelerate ICT development is given Figure 3.



Figure 3, Roadmap to accelerate ICT development

Effective e-government can provide a wide variety of benefits including more efficiency and savings for governments and businesses, increased transparency, and greater participation of citizens in political life adoption of digital services. For B&H, spectrum allocation must be shifted from spectrum revenues to socioeconomic enabler and important economy engine for job creation. One of spectrum issues is the 'Digital Dividend' (The Digital Dividend is the name given to the low-frequency part of the spectrum which is freed up as analogue TV broadcasting).

In B&H this process is going very slow and it is complicated since current General Manager is in technical mandate since 2007, and is strictly connected to political issue. (LTE License in B&H is not realised). It is important to maintain confidentiality of user data when they are either in storage or during processing. It is therefore important to develop secure yet efficient data processing and storage methods. Security measures may include privacy and secure aware distributed data storage and decentralized processing, which aim to maintain data confidentiality.

Environmental issues are becoming one of the central concerns in modern societies. Advanced green ICT deployments have to support environmental sustainability. Sharing of passive infrastructure can provide significant cost reduction and this process in B&H is established for more than ten years and ensures faster network deployments (Mtel, BH Telecom and HT sharing passive infrastructure for mobile sites). The rapid advancements in wireless communications are expected to increase the demand for radio spectrum resources by orders of magnitude during the next decade. This problem must be addressed using technology and regulatory innovations for significant improvements in spectrum utilization in cellular networks.

B&H authorities must consider how to support Digital Strategy with specific actions adopted to country circumstances. Absence of Strategy for digital development in B&H, including programs and projects resulted in negative consequences in field of education in ICT sector (especially in higher educational program for e- jobs). Without resources (human, organizational and financial) to overcome Digital Divide become fundamental process to make efforts in sustainable economic development and restructuring. New metrics must develop model reflecting future steps in measuring ICT development. With precisely defined milestones and the authorities responsible for the implementation, measure progress in ICT sector will be possible, as given in Figure 4. The development of human capital is inseparable from the development of the new digital society.



Figure 4, Model to measure ICT development

Digital society comes with new challenges and first is to create carefully planned and executed Strategy to achieve benefits of ICT development. By identifying ICT development trends, rules and directions, every participant in FC model can better formulate plans and strategies, and implement them.

IV. CONCLUSION

Development of digital economy has acted as a driver for economic growth in recent years. BB access is the key element to bridging the physical economy to the digital economy. National broadband strategy should take into consideration the requirements of different regions and provide policy support accordingly, and act as an important socioeconomic enabler. B&H have a unique opportunity to overcome the gap with EU region through development of a low cost BB Internet infrastructure. BB development will dramatically increase production and use of digital services bringing global knowledge in B&H, including growth in e- skills and local content application development. This process ensures bringing global jobs to local market, as an opportunity to be part of EU region. If the right conditions are set, B&H have an opportunity to create foundations for a digital society. This article provides an overview of an influence of future communication model on society as a whole, and gives roadmap to accelerate the adoption of FC model.

V REFERENCES

[1] D. Ilisevic and N. Banovic Curguz, "Evolution toward digital society in B&H", Telfor Belgrade, pp. 13-16, November 2015

[2] P. Kwadwo Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design Considerations for a 5G Network Architecture", IEEE Communication Magazine, Vol. 52, No. 11, pp. 65-75, November 2014

[3] S. Bi, R. Zhang, Z. Ding, and S. Cui "Wireless Communications in the Era of Big Data", IEEE Communication Magazine, Vol. 53, No. 10, pp. 190-199, October 2015

[4] L. Atzori, A. Iera and G. Morabito, " From smart `objects` to social `objects: the next evoluntionary step od Internet of Things" J IEEE Magazine., Communication, vol. 52, no. 1, pp. 97-105, January 2014

[5] J. Schonwalder, M. Fouquet and G. Dreo Rodosek and I.C. Hochstater "Future Internet = Content +Services+Management", IEEE Magazine., Communication, vol. 48, no. 11, pp. 27-33, 2009

[6] L. Jin, Y. Chen, T. Wang, P. Hui and A. V. Vasilakos, " Understanding user behavior in online social networks: A Surwey," IEEE Communication Magazine, Vol. 51, No. 9 September 2013

[7] C. Xu, S. Jia, L. Zhong, and G.-M. Muntean "Social-Aware Mobile Peer-to-Peer Communications, IEEE Communication Magazine, Vol. 53, No. 10, pp. 150-156, October 2015

[8] D. Ilisevic and N. B. Curguz, "Changing Face of Mobile Service Providers in B&H", Telfor Belgrade, pp. 51-54, November 2014.

[9] M. Beer, N. Nohria, "Cracking the Code of Change", Harvard Business Review, May-June 2000

[10] N. Banovic Curguz and D. Ilisevic "Toward Customer & Service Oriented Mobile Service Providers", Telfor Belgrade, pp. 9-12, November 2015.

[11] M. Banica, "Organization Design for Network Related Managed Services", Telfor Belgrade, pp. 5-8, November 2015

[12] Digital Agenda, available https://ec.europa.eu/digital-agenda/en/digital-agenda-europe-2020-strategy, January 2016

Some aspects of network management system for video service

O. Jukić, I. Heđi
Virovitica College, Virovitica, Republic of Croatia
{oliver.jukic | ivan.hedi}@vsmti.hr

Abstract - Network operators providing integrated service that includes fixed telephony, Internet access and television are faced with need of management system integration. It is needed in order to achieve complete picture of current service quality as well as service quality trend. Namely, while own access network as well as and HFC (hybrid fiber coaxial) network are built on some geographical areas, on other geographical areas operators are forced to use lines owned by other operators. On the other hand, customers use services transparently, without knowing anything about architectural or implementation details, perceiving service quality and availability only.
Hence, it is necessary to implement network management system that will integrate data from network parts controlled by other operators, in addition to own network monitoring, especially HFC part of the network. However, network parts controlled by other operators are visible as a "cloud", but data from cloud borders should be integrated into network management system. Management data from different sources should be integrated into video service quality management system through mediation layer. This paper describes some aspects of implementation of such system, obtained as a result of authors' work on real network management systems.

I. INTRODUCTION

Telecom operators offer the market a variety of services, which were implemented in the telecommunications network [1] (Figure 1):

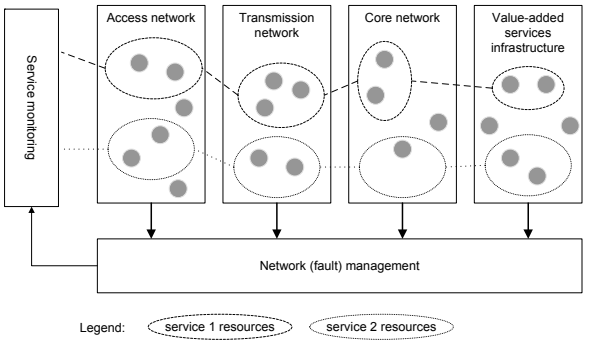


Figure 1. Service spreadng over network resources

In addition, each service is realized by using more network resources, depending on their complexity. Network operators monitor its own network and network resources. Knowing the structure of the network as well as the structure of the implemented services, it is possible to monitor potential problems on the network affect the quality of delivered services (e.g. downstream speed). Further, it is possible to monitor the quality of service by measurements of various service quality indicators.

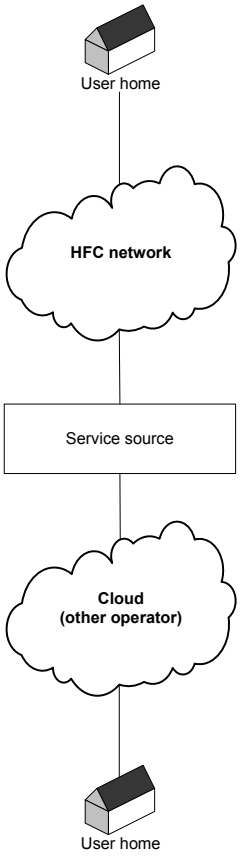


Figure 2. Video services – two different implemenations

The subscriber typically does not know and there is no need to know how the service is implemented on the network. As an example of such service, we took video service that can be provided to the subscriber in many ways. Here we consider two basic ways (Figure 2):

- The service is brought to the subscriber by using own HFC network,
- The service is brought to the subscriber using lines owned by other operators (i.e., leased lines).

This situation is very common in Republic of Croatia, because there is a substantial portion of the access infrastructure which was owned by the state telecom,

which is now rented to other operators, in ways that are regulated by law.

Operators on the same access networks typically deliver Internet access also; considerations related to video service apply for the Internet access service.

In this paper we will review the typical architecture that supports broadband data services over HFC network. Further, we look at the basic architecture that supports the delivery of such services through leased lines. We will show the need for the development of hybrid solutions of network management, which includes both of these architectures.

Further, some specific aspects will be touched. Service support and delivery requires inter-operator business communication as well as integration of geographical data related to specific customer. Those two aspects are briefly mentioned in third section.

Because one such network management solution is implemented in one telecom operator in Republic of Croatia, we will show certain aspects of implementation. Finally, we will specify the potential challenges and opportunities for further development.

A. Motivation

Although HFC network management is known term, described in number of papers ([6], [7]), there is lack of management integration in the case where service is delivered via leased infrastructure to the number of subscribers. Further, in our contact with telecomm operators, we have concluded that integrated management system will be welcomed in network and service operation centers. Hence, we have decided to research and analyze basic aspects of management integration, which resulted with integrated management system fully implemented at one telecom operator in Republic of Croatia.

II. TWO NETWORK ARCHITECTURES

In this section, two basic architectures providing video service will be described. First, basic HFC network architecture and then, architecture with access network implemented over the other operator's infrastructure will be shown. From network management point of view, both architectures have its own specific characteristics. It will be considered in third section.

A. HFC network

One typical network architecture of HFC network is shown in Figure 3.

The optical signal from the main station (headend) is transformed into an RF signal that is sent over the trunk and line amplifiers to the subscriber. Headend contains video signal receivers, which receive video signal from satellite or optical network, if possible; video-on-demand servers; routers for IP signal aggregation; equipment for telecom services (voice).

In addition to the sources of television signals, headend is connected to the Internet via high-capacity connections. On the side of the headend is CMTS (cable modem termination system), which enables communication with the customer's cable modem.

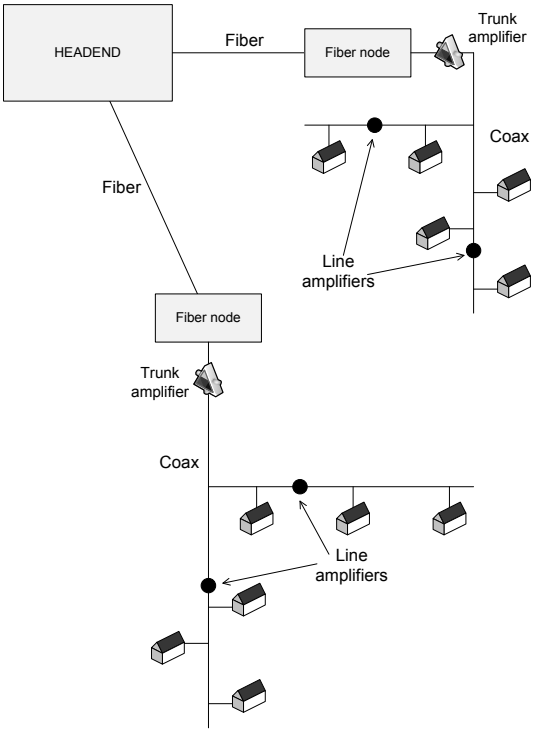


Figure 3. HFC network typical architecture

B. Access network over other operator infrastructure

Unlike access over the HFC network, in this case the subscriber is connected to the leased line of the other operator, where the signal is diverted from or to the parent operator, service provider (Figure 4).

If even intuitively clear here is very important to emphasize that there must be a complete synchronicity between these two operators, the first when provisioning their own services to the customer, but also later, when operators monitor the quality of service and finally when different repairs and network activities were performed.

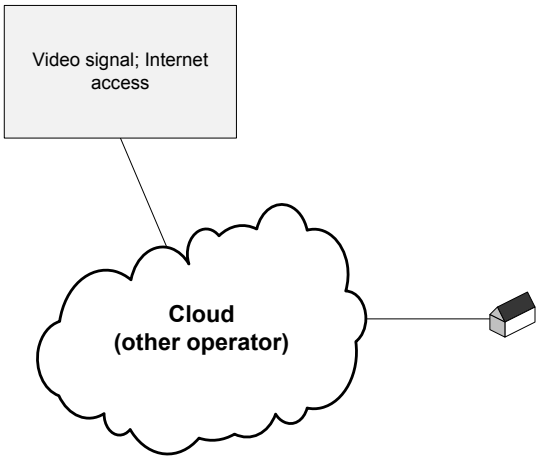


Figure 4. Access over other operator infrastructure

Each of these architectures have its own characteristics in terms of management.

III. NETWORK MANAGEMENT ASPECTS

In this section, basic network management aspects for both architectures described in second section will be discussed. Business communication between network operators as well as integration of geo coordinates will be also mentioned, as specific aspects of video service monitoring.

A. HFC network

Since the whole network infrastructure is the property of the network operator, it is possible to monitor all the network resources here, and hence monitor the service provided over those network resources. In that sense it is required to achieve the supervision and control over:

- Headend
- Fiber node
- Amplifiers

Further, it is required to supervise the sources of video signals, just like CMTS. Namely, it is possible to access the information about the state of the cable modem over CMTS, which can serve as an indication of eventual errors within the network. Knowing the habits of the subscriber is also required. For example, if the subscriber doesn't have a habit to turn off the modem overnight, each offline state of the modem can indicate an error.

B. Access network over other operators' infrastructure

In this case the main operator that provides the service has no direct access to the network resources. Because of that, the obligation of the provider over whose infrastructure the service is provided to provide information about the network measurements to the "main" provider is regulated by the law.

In this network architecture the data of other providers, provided typically in the form of files, is taken as the source of monitor data. It is required to process such files in real time, to aggregate the resulting data and based on it make control decisions.

C. Geo data integration

If we consider the functional architecture of the TMN (telecommunication Management Network) [3], shown on figure 5, it is obvious that the QAF (Q adapter function) and MF (mediation function) will play significant role in monitoring the above mentioned network architectures. These two functions enable the integration of data from multiple sources, in which we include the data from the other control systems (QAF). The mediation function will serve to uniform data from different sources, so that it could be used in the unique monitor system, and their aggregation [2].

Besides this, the business processes that ensure the data exchange between the operators and the connections between the CRM (customer relationship management) systems will have a significant role in the management.

Not only the need for the data of the network subscriber, but also the need for the GPS coordinates of each subscriber is imposed, so that the network could be properly visualized in the management tools. Namely, in city districts the areas covered by certain nodes of the HFC network can overlap, and it is necessary to know the exact area covered by a single node. This is especially important in the case of a failure of a particular node.

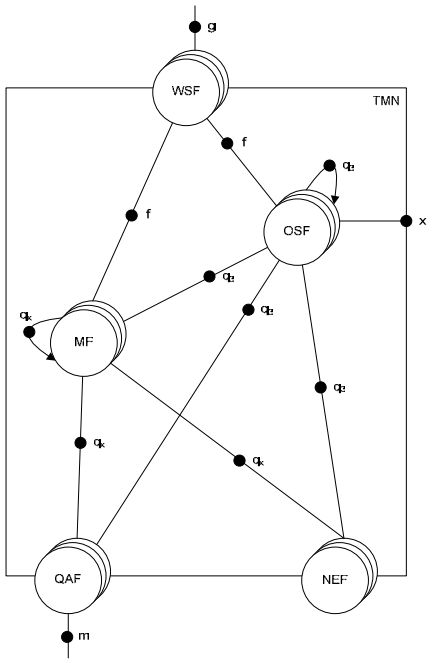


Figure 5. TMN functional architecture

IV. NETWORK MANAGEMENT SYSTEM IMPLEMENTATION ASPECTS

Based on management aspects mentioned above, we have implemented management system for video, voice and Internet services that are implemented over two different network architectures.

Management system is implemented and is fully functional at one telecomm operator in Republic of Croatia. Management system architecture is shown on figure 6.

For every particular network resource class, access module is developed: headend, CMTS, fiber node, trunk amplifier, line amplifier. These access modules collect management data using some of standard network management communication protocols, such as SNMP (simple network management protocol) [8].

Subscriber home data are monitored via CMTS. Hence it is presented as (logically) separated access module. However, subscriber home data as well as CMTS management data (e.g. alarms) are monitored by the same access module physically.

Data from HFC network are delivered to mediation and aggregation modules, preparing data for network management system for HFC network. On the other hand, files from other operators are collected continuously on

daily base, very similar to CDR (call detailed records) [4]. For every particular file type (e.g. file presenting performance data, physical measurements, inventory data, etc.) access module is developed also. These access modules act as parsers. Parsed data are inserted to management database where data are aggregated on daily basis. Aggregated data are delivered to network management system for other operators' infrastructure. Data from these two separated network management systems are joined using integrated management functions.

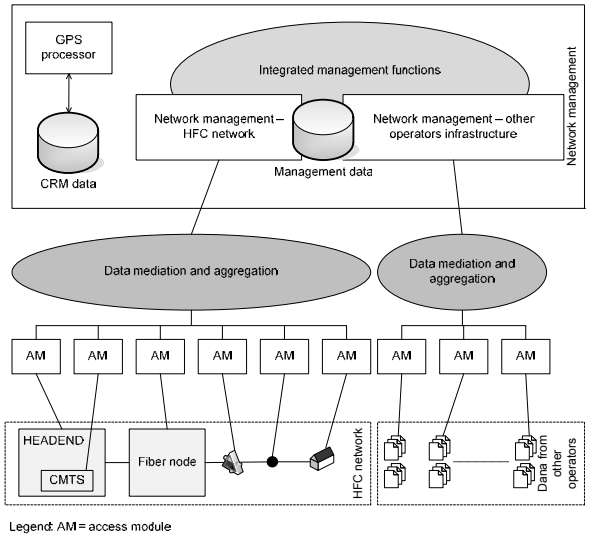


Figure 6. Network management system architecture

CRM data are integrated in management system. When customer data enter the CRM system, usually it is not know anything about its GPS (global positioning system) coordinates. Hence, GPS processor module is developed, resolving GPS coordinates for every new address entered in the CRM system.

For instance: with integrated management, it is possible to detect areas where subscribers exist that are connected to the network via other operators, even if own HFC network infrastructure exists.

Further, we can assume another scenario: help desk service receives customer's complaints without knowledge is customer connected directly to HFC network or service is delivered via other operator's infrastructure. User typically doesn't know anything about it. Integrating management data from two different sources in one management system, it is allowed to help desk personnel to react much faster [5], using one single tool.

V. CONCLUSION

In this paper we have pointed to and analyzed basic aspects of video (TV) service management. Service was implemented over two different network architectures. In one architecture, network operator has built own HFC network; on the other hand, in another architecture, service is delivered to the subscriber via leased other operators' telecommunication infrastructure.

For both architectures it is necessary to realize management system. However, it is crucial that those two systems finally should be integrated.

Architecture of such, integrated management system is described in this paper. As potential challenge we consider better integration with GPS coordinates which means improvement of CRM data quality: GPS data are crucial for high-quality of such management system.

Potential solution may be integration with public government's services carrying addresses database. Further, management data interchange mechanism between two operators should be improved (with finer granulation), but it is not technical issue only.

REFERENCES

[1] O. Jukić and M. Kunštić, "Integrated view on telecommunication network status", Proceedings of the 34rd International Convention MIPRO 2011, pp. 141-145, MIPRO, Opatija, 2011.

[2] O. Jukić and I. Hedi, "Service monitoring and alarm correlations", 4th International congress on ultra moder telecommunications and control systems, pp. 330-334, Saint Petersburg, 2012.

[3] D.K.Udupa, "TMN – Telecommunications Management Network", McGraw-Hill Telecommunications, New York, 1999.

[4] O. Jukić and I. Hedi, "The use of call detail records and data mart dimensioning for telecommunication companies", 20th Telecommunications Forum – Proceedings of papers, pp. 292-295, Beograd, 2012.

[5] S. Wallin and V. Leijon, "Telecom Network and Service Management: An operator Survery", MMNS 2009, pp. 15-26, 2009.

[6] Paff, A., Hybrid Fiber/Coax in the Public Telecommunications Infrastructure, IEEE Communications Magazine, April 1995.

[7] Perry, E. and Srinivas R., Network Management for Residential Broadband Interactive Data Services, IEEE Communications Magazine, November 1996.

[8] Uyless, B.: Network Management Standards – SNMP, CMIP, TMN, MIBs and object libraries, McGraw-Hill 1994.

Influence of OTT Service Providers on Croatian Telecomunication Market

I. Dražić Lutilsky*, M. Ivić**
* University of Zagreb Faculty of Economics and Business,
Department of Accounting, Zagreb, Croatia
** University of Zagreb Faculty of Electrical Engineering and Computing,
Department of Telecommunications, Zagreb, Croatia
idrazic@efzg.hr, marina.ivic@fer.hr

Abstract - The object of this paper is to express influence of OTT service providers on Croatian telecommunication market in comparison with some European countries and United States based on mobileSQUARED research. As part of the project "Looking to the future", the authors conducted a questionnaire about the impact of OTT service providers on the Croatian telecommunications market. The questionnaire has been sent to Croatian Telecom, VIPnet and Tele 2. The questionnaire was send in May and June 2015. All three operators have responded to the questionnaire. From the collected data, it has been concluded that decrease in traditional traffic since 2008 is around 10% and the operators believe that will remain the same for the next 5 years. Consequently, decrease in revenues from traditional services since 2008 is up to 20%. They are trying to generate revenues from charging data or from offering diversified service packages believing that LTE is offering new possibilities for generating revenues. They all hold that it is profitable to invest in infrastructure and new technologies so in 2016 they will invest from 10 to 40% of profit.

Keywords: ***OTT services, OTT providers, Croatian telecommunication market, empirical research***

I. INTRODUCTION

For some time Short Message Service (SMS) was safe source of income for mobile operators, and in 2012 represented 20% of total revenues [1]. However, from 2008, OTT service providers cause revenue reduction for operators jeopardizing SMS. With the increased usage of OTT services, especially those oriented to the text messaging, it started decrease usage of traditional telecom services such as SMS. Therefore, OTT services have become a serious problem for mobile operators because of the negative impact that they may have on the number of send SMS and MMS messages, and thus revenue. According to BEREC report, OTT services have a great impact on the telecommunication sector, regarding competition and consumer protection, net neutrality and current EU regulatory framework [2]. BEREC report defines OTT service as "content, a service or an application that is provided to the end user over the open Internet." Definition implies that what is provided can be either content, a service or an application, means that anything provided over the open Internet is an OTT service [2]. Therefore, they have broken down the OTT services as OTT-0: an OTT service that qualifies as an Electronic Communication Service; OTT-1: an OTT service that is not an Electronic Communication Service but potentially competes with an Electronic Communication Service; and OTT-2: other OTT services [2]. Due to the technological developments and blurred line between OTT service and Electronic Communication Service, the authors will give perspective over Croatian Telecommunication market and explain the influence of OTT service providers on Croatian operators.

The structure of the paper is as follows. In the first section, the Introduction, the authors are providing with the definition of OTT service and describing the problem for the operators. In the second section, the authors are providing literature review of some similar research which is showing the influence of OTT service providers on operators in some world regions and countries. In the third section, the authors are describing the results of conducted research in Republic of Croatia. In addition, in the fourth section, the authors are giving the overall conclusion of the paper.

II. LITERATURE REVIEW

Survey conducted in 2012 by mobileSQUARED called "Over - The - Top Services: How Operators can overcome the Fragmentation of Communication" was analysed [3]. The study included 68 countries and over 40 mobile operators. In addition, in the study WhatsApp and Skype were identified as the leading OTT service providers for voice and text services. According to the survey, Skype has over 900 million users, while WhatsApp in 2012 had 75 million users, and it is estimated that in 2016 will have over 250 million users. The main preamble of this study highlights that mobile operators begin to observe OTT service providers as well as the ability to generate incremental revenues, the only question is when. The point is that operators should facilitate the strategy of OTT service providers and thus participate in the distribution of income, rather than seeing them as services that threaten their.

What also stands out is fragmented communication because users can only communicate through one platform and there is no possibility of communicating outside the "walled garden". This weakness could be the possibility for operators to take a leading role in enabling OTT services and potential revenue, whereas the search for new business models in order to prevent reduction in revenue of traditional services such as calls and SMS.

According to estimates made by mobileSQUARED up to 2016 global penetration of smartphones will be 39%, which represents one-third of the user population of mobile devices (in 2012 users of OTT services on smartphones was 276.8 million and it is estimated that this number will grow in 2016 on 1.32 billion) [3]. In

addition, their study showed that 79% of mobile operators believe that OTT services are a threat to the traditional call and SMS services. Even 73.7% of operators have identified SMS as the most threatened by the OTT service providers. All surveyed operators believe they will feel the drop in the traditional traffic and revenue from call and SMS services over the next 5-10 years. Accordingly, 53% of operators believe that the revenues in the next 5-10 years will fall by more than 20% while 37% of operators believe that this decline will be up to 20%.

However, despite the concerns of operators, most operators believe that it is possible to make money on OTT services. Thus, 63% of operators believe that they will be able to make money on OTT services, but at the expense of call and SMS services. While 16% of operators believe that, they will be able to generate incremental revenue from OTT services while only 21% of operators believe that they cannot make a profit from OTT services.

The biggest reduction of SMS traffic volume per subscription through the years 2008 to 2013 was recorded in the Netherlands, Spain and Austria, while France and Belgium still recorded an increase [4]. These two countries are fighting for their territory offering free SMS messages and thus end users still use it for text messages. Figure 1 shows data for individual countries on reducing the volume of SMS traffic per subscription through the years 2008 to 2013 [5].

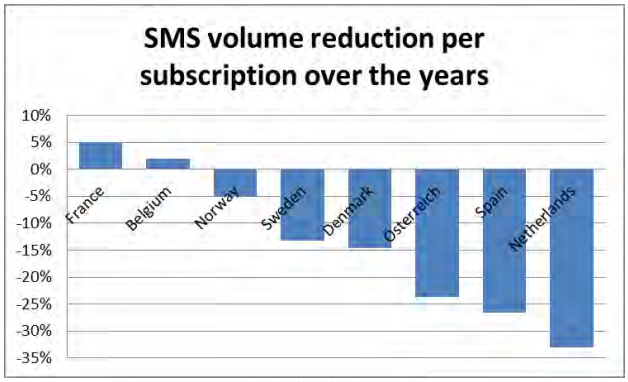


Figure 1. Data for individual countries in 2013

According to IDATE Research, operator revenues consist of revenues from fixed telephony, revenues of mobile calls and revenue from SMS [5]. Revenue made by OTT service providers consists of VoIP, IP text messages and part of revenues from social networks. They argue that the revenue made by OTT service providers is still marginal at world level in relation to the operators. In Figure 2 it is visible data at the global level in billions of euros of total revenues of operators and OTT providers [4]. They estimate that revenues from OTT service providers in 2016 will be 17 billion euros and in 2018 24 billion [5].

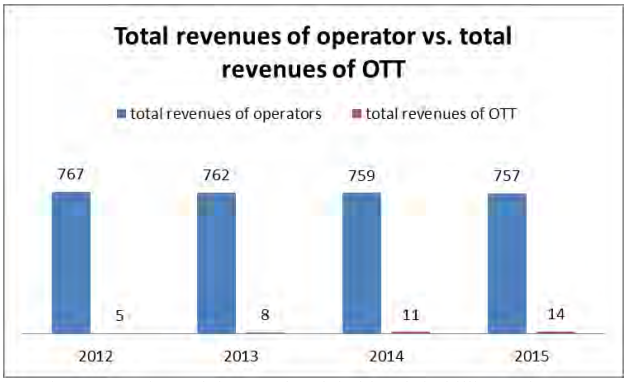


Figure 2. Estimated data on the global level (in billions of euros)

According to mobileSQUARED research, OTT market will be worth 166.5 million dollars in 2016. Both studies show different values of OTT market, which actually tells us how difficult it is to obtain the data for estimation the actual value of OTT services in the telecommunications market [3].

Moreover, towards mentioned research, operators in US, Asia and Europe (15%) have not yet been determined whether the OTT services influenced the decrease of traditional traffic. About 25% of European operators pleaded that in 2012 OTT services have not effected the reduction of their traditional traffic while 45% of them claimed that this reduction was between 1 and 10%. Almost 10% of European operators estimated that they have been affected by OTT services in 2012 which led to reduction of their traditional traffic between 11 and 20%. All of the surveyed operators believe that there will be a traffic reduction of call and SMS service in the next 5-10 years. 37% operators believe that this reduction will be up to 20%, while 53% of them believe that this reduction will be greater than 20%. Over 45% of European operators along with most of operators in North and South America are still expecting reduction in revenues due to OTT services. About 25% of European operators have confirmed the reduction in income up to 5% by 2012, while slightly more than 10% of European operators claim that more than 6% of revenue was lost by 2012 due to OTT services. 32% of surveyed operators believe that in the next 5-10 years the revenue decrease from traditional services will be up to 10% and even 43% of them believe that there will be a reduction greater than 11%.

A. United States of America

According to mobileSQUARED research and estimates in the United States will increase the number of users of OTT services from 155.6 million in 2012 to 265.8 million in 2016. According to the survey, Facebook is used by 37% of users, Skype 17%, Twitter 17%, 11% iMessage, BlackBerry Messaging 10% and 5% of WhatsApp users [3]. OTT text messaging uses a lot less users given to the traditional SMS services included in the package of services, which for customer is sunk cost and therefore they continue to use traditional SMS. Number of users of OTT services that send off-net messages will increase from 26.1 million in 2012 to 63 million in 2016, while the number of users who use off-net mobile and fixed calls will increase with 13.3 million and 8.5 million in 2012 year to 55.6 and 14 million in 2016. In

general, the assessment is that off-net traffic will generate income over $ 1 trillion in 2016, while in 2012 that income was 795.9 million. They find that revenue from OTT calls to land lines is negligible but that revenues from off-net text messages in 2016 to be around 756 million, while in 2012 were 643.1 million dollars [3].

B. Germany

Germany compensated initial poor penetration of smart phones. Total number of citizens in 2013 was 80.62 million. In 2012, there were 49 million users of smart mobile devices and it is expected to nearly double in 2016 on about 98 million users [3]. It is expected that the total number of OTT services from 14.9 million in 2012 to grow to 51.1 million in 2016. Users of OTT services who use off-net messages will generate revenue in 2016 around 20.4 million while the income in 2012 was 8.2 million. Number of OTT users using off-net fixed calls will increase from 2.7 million in 2012 to 5.1 million users in 2016. OTT customers who use off-net mobile calls will grow from 4.2 million users in 2012 to 20.4 million users in 2016. Total possibility of off-net services in Germany is estimated to be 366.7 million in 2016 while in 2012 this value was 250.5 million dollars [3]. Income from off-net messaging will dominate, generating revenues from 202.4 million dollars in 2012 to 276.8 million in 2016. They find that fixed and mobile off-net calls generate total revenue of 89.9 million dollars in 2016, while in 2012 it was 48.1 million [3].

C. Great Britain

Projections for OTT services are related to the strong growth of smart mobile devices. Total number of citizens in 2013 was 64.1 million. In 2012, the number of users of smart mobile devices was 43.1 million while it is estimated that in 2016 this number will reach 70.3 million [3]. The penetration of OTT services in 2016 is expected at 36.9 million users while in 2012 this number was 13.2 million. Number of off-net messages in 2012 was 7.2 million while it is estimated to reach 16.6 million in 2016. Number of OTT users who had off-net calls in 2012 was 3.7 million while the expected increase in 2016 is up to 14.8 million. Also, the number of users who had off-net fixed calls in 2012 was 2.4 million, while it is estimated to reach 3.7 million in 2016. The total value of off-net services in the UK was 220.6 million in 2012 while it is estimated to be 264.9 million in 2016 [3].

According to all of the above, it is undisputed that OTT services have a significant impact on the telecommunications market. However, the main question is the financial impact of OTT services on the telecommunications industry. In 2012 the mobile industry (calls, text messaging, VAS, and access devices, infrastructure and OTT) was worth about 1.5 trillion according to analyst Chetan Sharma mobile industry. According to this figure, mobileSQUARED estimates that it will be worth nearly 1.97 trillion in 2016 [3]. It is considered to be:

- Revenues from calls fall from 714 billion in 2012 to 573 billion in 2016;
- Revenues from the traffic and Internet access to nearly double in 2016 compared to 2012, with 222 billion dollars to 407 billion dollars.

Overall, it is considered that the income of operators of their "core" services to be 1.15 trillion in 2016, while in 2012 was 1.114 trillion. Revenues from traffic and Internet access will reimburse the loss of revenue from calls and SMS. These lost revenues will be allocated directly or indirectly to OTT service providers. Even 16% of operators believe that they will fail to generate incremental revenue from OTT services. But two-thirds of operators believe they will be able to earn from OTT services however only at the expense of revenues from voice and SMS services. Even 21% of operators believe that they will not be able to make money from OTT services or remain undecided. As far as European operators the situation is quite similar to the global level. About 12% of operators believe that they will be able to generate incremental revenue from OTT services. Even 75% of European operators believe they will be able to earn from OTT services but only at the expense of revenues from voice and SMS services. The remaining 13% of European operators do not believe they will generate revenues from OTT services. The situation is similar in Asia, North America and Latin America.

III. RESULTS FROM EMPIRICAL RESEARCH IN REPUBLIC OF CROATIA

As part of the project, "Looking to the future" (Croatian National Regulatory Authority) the authors conducted a questionnaire about the impact of OTT service providers in the Croatian telecommunications market [6]. The questionnaire has been sent to Croatian Telecom, VIPnet and Tele 2. The questionnaire was conducted in May and June 2015. All three operators have responded to the questionnaire.

According to the answers there are around 2 million users of smart mobile devices which use at least one OTT service. Figure 3 shows number of users of smart mobile devices that use OTT services. Regarding operator 1 and 2 in 2015, the number of users who will use OTT services will be much higher than 40%, while only operator 3 estimates that number will be up to 20%. From the given response, it is visible that Croatian operators marked text messages as the most used OTT service from all OTT services. That is most likely, given the popularity of WhatsApp and Viber in Croatia and the possibility of monitoring the reduction in revenues from traditional SMS services. Furthermore, an operator ranked voice services to mobile devices as second most used OTT service by their users, while other operator put video service at the same position. Operators are able to identify a portion of particular OTT services (such as WhatsApp, Viber, Skype, Facebook and YouTube) in relation to the total OTT services. Additionally, operators are able to distinguish the use of OTT services in relation to traditional services by comparing revenues from OTT services and follow-up revenue from traditional services. Also, ranking OTT services will depend on the preferences of end-users. Both operators that responded to this question believe that the voice OTT services to fixed devices are at

least used of the total used OTT services. One operator did not answer this question thinking it is impossible to assess, analyse and go into the content of the communication. All three operators believe that OTT service providers are threat to traditional SMS and voice services which they provide. When asked which of their services are most threatened by OTT service providers, two operators have responded that the most threatened are SMS service and voice service while one operator responded only SMS service. All this indicates that operators have become aware of the threat made by OTT services to traditional services resulting drop in revenues and profits.

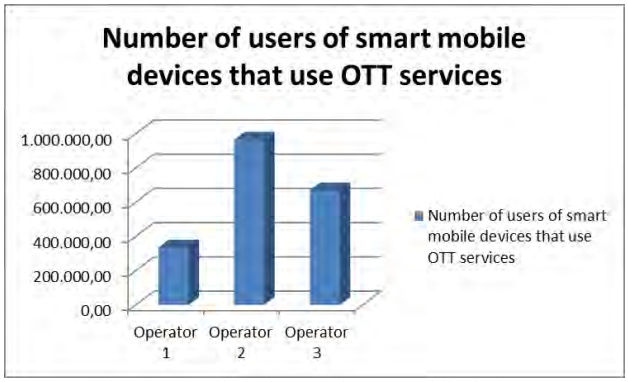


Figure 3. Number of users of smart mobile devices

Nevertheless, it is unlikely that overall operator costs will be reduced given the level of construction of infrastructure and capacity.

It is interesting that one operator answered negatively to the question whether the traffic decreased with the emergence of OTT service providers and applications on smart mobile devices since 2008, believing it does not affect in the same way to all kinds of services, and therefore it is not possible to clearly answer this question. However, at the same time in previous issues, aforementioned operator recognized that their SMS and voice calls are the most endangered by OTT service providers. Also, this operator estimates that there has been a revenue reduction with the use of OTT services on smart devices since 2008 up to 6-10%. Moreover, the operator predicts that in the future this reduction will amount between 11 and 15%. The revenue reduction is probably a partial reduction of the traditional traffic but also the reduction or change in service prices. The other two operators have responded that reduce in the traditional traffic is caused by emergence of OTT service providers. One operator cannot estimate the exact numeric range while other estimates that since 2008 the reduction was 1-10%. The subsequent operator believes that revenue reduction is not caused by increased usage of OTT services on smartphones since 2008, which is somewhat illogical, because it would imply price increase of their services to keep the revenue at the same level.

In the next question the operators were supposed to assess whether there will be decrease in the current level of traffic of traditional telecommunications services in next 5 years due to use increase of OTT services on smart mobile devices. One operator answered negatively because it considers that the use of OTT services is not affecting in the same way all kinds of services, and therefore cannot give an explicit answer. Other operator cannot assess the reduction but believes that it will happen while the third does not know whether the reduction will happen.

The next two questions are related to the assessment of revenue decrease in SMS service due to increased use of OTT services on smart mobile devices. The first question refers to the period from 2008 to 2016, and the second one on the following five years. On the first question, one operator estimated that there has been a revenue reduction from 16 to 20%, the second operator estimated the reduction from 11 to 15% and the third operator cannot assess the range. On the second question one operator estimated decrease in income by 16 to 20%, second one by 6 to 10% while the third operator cannot assess the range. The responses show that operators are aware of revenue decrease and the potential risks arising from it.

Accordingly, they will have to find new ways for achieving revenues and profit through the development of new technologies and offering various service packages at different prices.

Next question refers to a way of protection against falling income from traditional services due to increasing use of OTT services. Operators were offered the following answers: earn money from consumption of traffic; impose additional fees; offer their own similar services; creating partnerships with OTT service providers and more. Figure 4 shows answers of operators regarding ways to protect decline in revenue from traditional services due to increasing use of OTT services. According to the answers it is clear that the operators are still working on a business strategy. One operator tends to protect traditional service by charging traffic consumption; the other operator develops cooperation with OTT service providers while the third one sees the solution in offering a package that includes all kinds of services.

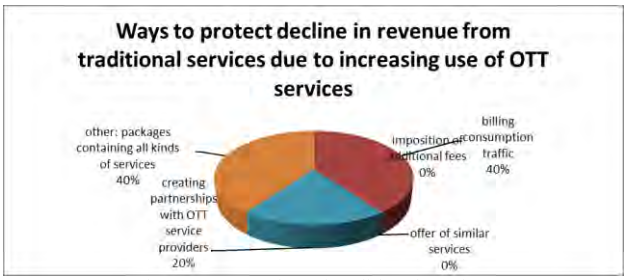


Figure 4. Protection of traditional services

Similarly, two operators believe that they cannot earn from OTT services or disproportionate value provided, but through fast and reliable Internet access and Internet services.

Asked whether LTE offers the possibility of creating some new business models that will generate revenue, the two operators have responded positively. One sees the possibility of price differentiation through speed data traffic while the other sees it through cooperation with OTT service providers.

In addition, all operators are considering cost-effective investment in the network, infrastructure and new technologies regardless the reduction in revenues.

One operator estimated investment up to 40%, the second operator up to 10% and the third pleaded that it is in the domain of business secrets.

All three operators are considering OTT service providers as a real threat to their business and below are their answers why.

Croatian operators are of the opinion that today OTT service providers exercise its direct relationship with end-users, as a rule, without entering into written agreements or ensuring the necessary level of transparency. Also they do not post General Terms and Conditions of Use that users, if they can access, generally accept. They also do not provide contacts and the procedures for filing a complaint. Unlike OTT service providers, the operators of public communications services have to fully apply the obligations of the relevant national regulations.

A significant problem is substantially reduced level of protection of users' rights, especially regarding the obligations in the field of data retention and protection of personal data. Stated allows OTT service providers to have greater flexibility and lower costs in the development and delivery of its services. They are not obliged to implement all the functionality that are nationally prescribed and are not faced with possible fines in case of non-compliance with regulatory requirements. They also believe that OTT service providers do not participate equitably in the distribution of investment risk and the resulting revenue.

At the same time OTT service providers use capital-intensive network capacities of Croatian operators free of charge. Furthermore, the profit earned through business income is not invested nor shared with key stakeholders of certain countries where infrastructure and end-users are located that present resource of revenue realization.

The existing OTT business model is generating the cost in the countries where users are located while income and thus tax payments are accomplished in the countries where the OTT service providers are registered.

IV. CONCLUSION

Thanks to the OTT services there was a decline of competitiveness of conventional telecommunications services such as SMS, MMS and voice. Simplicity, intuitiveness and cost effectiveness offered by OTT services to their end-users ensure their positioning in the telecommunications market and the trend continues exponentially. In order to protect its revenues, operators of traditional services are changing the business strategy and adapt to new trends in the market. Current answer to OTT services in the Croatian telecommunications market is in the charging the traffic consumption, in possible cooperation with OTT service providers and offering a package that includes all kinds of services.

Strategies that have direct negative impact on the financial perspective for long-term users are unsustainable and should be avoided. Other strategies require investment in telecommunications infrastructure, or the long-term sustainability, and the question is who should and to what extent participate in the venture. The telecommunications mobile network operators are facing a great challenge because mobile networks are simply too slow and overburdened due to the accumulation of new services by OTT service providers that directly causes the explosion of data traffic. Unavailability of services shows that the existing network reached its maximum strength, and only network of fourth generation - LTE corrects the identified problems.

Looking from a financial perspective to the global telecommunications market, the total income of operators fell by 10 billion euros by comparing 2012 and 2015, while OTT service providers almost tripled their total revenue (in 2012 was 5 billion euros, and in 2015 as much as 14 billion euros). The reason is the explosion of data traffic and a drop in turnover of traditional telecommunications services. The Croatian telecommunications market is following world trends. The research on Croatian Telecommunications market showed a decrease in traditional traffic around 10% since 2008 and remains the same for the next 5 years. Decrease in revenues from traditional services since 2008 is up to 20% and remains the same for the next 5 years. Croatian operators are trying to generate revenues from charging data or from offering diversified service packages while believing that LTE is offering new possibilities for generating revenues. They all believe that is profitable to invest in infrastructure and new technologies and in 2015 they will invest from 10 to 40% of profit.

REFERENCES

[1] Popović, Ž, (2012): OTT usluge – prijetnja ili poslovna prilika za operatore, Komunikacije, 12/2012. http://www.ericsson.com/hr/etk/novine/kom0512/30.pdf

[2] BEREC – Body of Euroepan Regulators for Electronic Communications, "Report on OTT Service", Draft Version, October 2015. http://berec.europa.eu/eng/document_register/subject_matter/berec/reports/5751-berec-report-on-ott-services;

[3] Report by Mobile Squared „Over - the – top (OTT) Services: How operators can overcome the Fragmentation of Communication" , 2012. http://www.schwartzpr.de/de/newsroom/tyntec/2012-08tyntecOTTWhitepaperSummary.pdf

[4] IDATE Research, „Telecom & Over the Top – Communication Services, VoIP – IP Messaging – Social Networking: OTT vs Traditional telco markets", 2014.

[5] "Netherlands telecommunication report", Business Monitor International, 2014.

[6] Croatian Natinal Regultory Authority, http://www.hakom.hr/default.aspx?id=902; project "Looking to the future".

Implementing Shared Service Center in Telecom Environment as More Efficient and More Cost Effective Business Model

T. Žilić and V. Čošić
Hrvatski Telekom/IMS Shared Service Center Hub Department, Zagreb, Croatia
tomislav.zilic@t.ht.hr; vladimir.cosic@t.ht.hr

Abstract - Seeking for more efficient and more cost effective business model is a constant issue for managers in IT industry, especially in companies like big telecom providers that are incumbents. This paper is dealing with one of those new business models, implementation of Technology Shared Service Center in one of the biggest telecom providers in Europe. To have such an approach, and even more to have it within several national companies, is a huge organizational change. Advancing telecom technology is an additional motivation for moving to new business models. Therefore, implementation of the Shared Service Center in telecom, or any IT business, could be both organizational and technological challenge. Besides the challenges faced while implementing Technology Shared Service Centers, this document is describing the main benefits of this model with special focus on telecom environment.

I. INTRODUCTION

In today's business landscape, seeking for new business models and implementing solid strategies are crucial governance issues in all big organizations. The goal is to find more efficient and more cost effective business model by which organizations make value.

An example of such a big organization is a telecom provider that consists of several companies as affiliates in different countries. In this organization, or corporation, many companies share common business processes and require a corporate role to ensure their most effective use. Corporate role stands for the actions mandated at the corporate level that create synergies at the business unit level [1]. Organizations can also create synergies by aligning the internal units that provide shared services. These shared service units or centers are created in a corporate or division level because of economies of scale and the advantages of specialization and differentiation that these units can create.

This paper describes Shared Service Center concept with its benefits and reasons for implementation. Furthermore, the main objective of document is to present the implementation of Technology Shared Service Center in NT/IT (Network Technology / Information Technology) environment, in an international telecom provider to be more precise. Challenges faced during the process of implementation, both organizational and technological, will be described as well. At the end, the benefits of such implementation will be seen in conclusion as possible motivation to use this business model.

The paper is organized in four chapters. After introduction as the first one, in the following chapter of this paper will be described the definition of the Shared Service Center and its main benefits. Further focus, in the third chapter, will be on implementation of the Shared Service Center model in telecom, or any IT business related environment, with challenges faced from the organizational and technological point of view. Finally, as a conclusion in the fourth chapter, main reason and benefits of this model will be summarized to see whether Shared Service Centers are more efficient and more cost effective business model for telecom providers.

II. SHARED SERVICE CENTERS

A. Definition and characteristics

There is no unique understanding of Shared Service Centers (SSCs) in science and practice. The different definitions vary in many ways. Nevertheless, some characteristics, which are defining the term SSC, been mentioned in most publications are: the consolidation of processes within the group, delivery of support processes, cost cutting as a major driver, focus on internal customers, alignment with external customers, separated organizational unit, and operation like a business [2].

Another definition says that the Shared Service Center is an accountable entity in the internal organization tasked to provide specialized services to operational entities (divisions, business units) on the basis of a service level agreement, and full charge out of costs on the basis of a transfer price system. A Shared Service Center is a cost center, thus it is not a profit center, although it has to contribute to the profit of the organization in a measurable way [3].

Shared services are a collaborative strategy in which a subset of existing business functions is concentrated into a new semi-autonomous business unit that has a management structure designed to promote efficiency, value generation, cost savings, and improved service for the internal customers of the parent corporation, like a business competing in the open market [4].

Some of the most frequently mentioned characteristics and features are saying that SSC:

- is the result of the consolidation / concentration process within the company,

- is engaged in support services / staff functions / internal services,
- is aligned with external competitors,
- is a separate operated and steered organization within the group,
- reduces costs / competitive costs,
- is operated like a "normal business unit",
- has a focus on internal clients / delivery to internal customers,
- delivers competitive service quality,
- makes use of "Best Practices",
- uses resources group-wide,
- has established a supplier-customer relationship,
- is controlled by business units,
- provides services tailored to customer needs,
- delivers its services to several units,
- charges back to business units,
- focus on continuous improvement.

Type of activities, mostly performed by SSC's are: accounting services, HR services, IT services, facilities management, legal and insurance services, purchasing. Later on in this paper, SSC for NT/IT services in telecom environment will be described in more details.

As seen in the definitions above, the biggest challenge while implementing SSC in big organizations is organizational change. Activities from existing business entities are being transferred to newly established units of SSCs. These new units are created almost like an independent company within an existing organization. New processes, contractual agreements and even employees are a major part of this change.

B. Drivers for SSC introduction

In the current global economic environment, which is characterized by downsizing, mergers, acquisitions, and uncertainty, managers are trying to find a way to improve while increasing competitiveness. Thus, the main driver in seeking new business models like SSC is about increasing effectiveness and value of organizations. Additionally, drivers for SSC introduction are fundamentally about optimizing people, capital, time, and other corporate resources.

By standardization and consolidation of common functions across multiple organizations in the form of introducing SSCs, management of such organization seeks to reduce operational costs and to increase information and knowledge sharing. However, to successfully implement SSC, many organizations discovered that it requires considerable efforts to make their shared services model suitable. It is obvious that shared services require significant transformation of business processes. On the other hand, this change is needed if organizations would like to dramatically improve their business performance and achieve benefits like lowering the cost of business processes, improving services, forming a long term strategic alliance with other entities to share information and knowledge, and establishing leadership through focus on core business functions [5].

Shared services are often mistakenly understood as outsourcing, even though there is a fundamental difference. For shared services, the shared service provider is formed and governed by the parent organization. For outsourcing, the relationship between the service provider and the outsourced organization is usually defined by a bilateral contract. However, sometimes the implementation of SSC is the preliminary stage of outsourcing. Processes and resources are consolidated company-wide in a SSC to be potentially outsourced to an external provider afterwards.

III. IMPLEMENTING SSC IN TELECOM ENVIRONMENT

A. Technology/Telecom business environment

Shared services in technology environment are becoming important and often utilized approach to organizing NT/IT services. Similar like drivers for SSC introduction in general, that were named in the previous chapter, the motivator for establishing Technology Shared Services Centers typically include quality of service improvement, competitive prices, comparability increases, use of synergies with economy of scale, etc. In addition, they can generate additional revenue by providing services to external customers as well.

Characteristics of the Technology SSCs that are commonly accepted in practice and literature [6] are presented below. Technology Shared Service Centers:

- aim to consolidate the supporting processes within the company,
- focus on the organization's internal customers,
- are semi-autonomous, independent organizational units,
- provide supporting processes for the core processes of the company,
- supply multiple divisions, respectively customers,
- aim to optimize the use of resources within the organization, and
- orient themselves towards external competition.

Figure 1 shows how Technology SSCs can be organized within the business group.

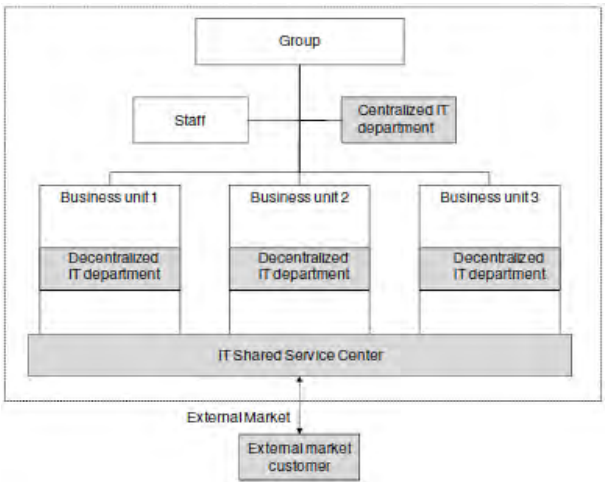


Figure 1. Technology Shared Service Center organization

To explain how Technology SSCs are competitive in providing services for the internal organization, compared to alternatives such as outsourcing, following enablers could be identified [7]:

- same corporate culture,
- knowledge of the parent organizations and its processes,
- lower transaction costs (due to internal market of parent organization),
- lower average cost,
- no dependence on external NT/IT service provider,
- better data protection (in terms of data security),
- improved technical opportunities (creation of new opportunities for the parent organization), and
- obligation to provide service.

An example, described later on, of Technology SSC business model implementation is a big telecom corporation that consists of several companies. Those companies are big corporation affiliates that are legal entities in different countries. Implementing SSC model in such an environment was a huge challenge. It was an effort not just from organizational and process point of view, but technological aspect as well is very important as telecom technology is advancing and foreseeing some new opportunities in near future.

B. Organizational challenges in implementation

As presented in previous chapters, organizational impact while implementing the SSC model is probably the biggest issue. A couple of years ago, one of the biggest telecom providers in Europe and big international company decided to implement Technology Shared Service Centers across its organization. This decision was impacting a dozen of national companies, separate legal entities, as parent company affiliates in a dozen different countries. Drivers for such an approach are obvious, as it was already described in this paper, to centralize and concentrate common activities and achieve certain benefits as efficiency and cost effectiveness. Even though there were different services in scope at global company level (HR services, accounting, etc.), the subject of this paper are technical services that are specific to telecom business.

A new virtual entity of Technology Shared Service Center was established, including a subset of several Shared Service Centers for each technical area, or parts of telecom network/service to be more precise. Examples of such newly established entities as telecom network functions are IMS (Internet Protocol Multimedia Subsystem) SSC, IPTV (Internet Protocol Television) SSC, RAN (Radio Access Network) SSC, EPC (Evolved Packet Core) SSC, etc. The challenge was how to organize such activities across several countries. After thorough analysis, it was decided to appoint certain SSC unit to certain national company as an affiliate of the corporate group. That meant that new business units as SSC departments were established in national companies and were providing services to other affiliates of the group. SSCs were divided across the whole group, according to certain national company references in a certain area of expertise, and to avoid concentration of SSCs in one or two of the affiliates. SSCs were responsible to the parent organization in global, but still remaining as business units within its national companies and bringing them additional value. On the other hand, SSCs were supposed to be cost effective at global group level. Besides establishing new SSC entities, a lot of processes needed to be developed within a global organization. Each of SSCs had to develop its service catalog to be precise on demarcation lines between SSCs and national companies. It had to be clearly understood, which services should be provided by SSCs so national companies could reorganize their internal activities of the same services. Furthermore, new contractual agreements were to be aligned and signed within all legal entities included in the process.

C. Technological challenges in implementation

Putting aside challenges from an organizational point of view, technical ones should be a matter of concern as well. After all, this paper is dealing with SSCs of specific technical services in global technology company. It had to be analyzed, from technical prospective, what services are common to certain organizations with the potential to be moved to SSCs. Furthermore, it had to be agreed which of those services could be centralized and in what way. On top of that, everything had to be aligned across all national companies and global organization to achieve common understanding.

As already seen, Technology SSCs were organized according to network functions. All of this network functions are in some way connected and dependant on each other. Since the object of analysis is a big telecom provider, there are hundreds of telecom network elements and telecom services in scope. Therefore, a lot of other parties like telecom vendors in the first place, were significant stakeholders in the outcome of this process. So, it was a complex mixture of different network functions, different telecom services, different vendors and different national companies with specific network setups to be aligned with the process of implementing the Technology SSC model.

Another technological challenge that is coming towards is centralization of network functions. This is common to all NT/IT services in general, as advancing technology is making network functions virtualized and centralized. As a matter of fact, processes of NT/IT virtualization and centralization are ensuring even more effectiveness for Technology SSCs and making their services even wider.

D. Benefits

After implementing Technology SSC, as described previously, with its challenges faced during the process of implementing it should be seen what are the benefits of such an approach. There were several drivers as hypothetical benefits for the organization and its stakeholders. Since the SSCs are now up and running within the organization, those potential benefits have to be confirmed as really beneficial and indicate the outcome of SSC implementation as successful. Anyhow, it should be measured in a certain way, whether the objectives of this process were met. The main stakeholders in this process,

that are entitled to evaluate, are: national companies as Technology SSC service receivers, the parental organization at the corporate group level, and Technology SSC units themselves.

National companies, now as Technology SSC customers and service receivers, can find several benefits that could influence their business in a better way:

- they don't have to focus anymore on part of activities that were moved to SSC,
- efforts and resources from their prospective could be minimized in a matter of SSC activities,
- since SSCs are using synergies at group level, for example to have common contracts with vendors or any other stakeholders, by economy of scale it is clear that cost savings are a major argument for SSC success.

At the corporate group level, parental organization can find also following benefits:

- major benefit as cost savings at the group level,
- centralization of common activities within SSC is another major advantage that makes this business model rather efficient as opposite to previous organization, and
- another global benefit is that SSC can now represent a whole group to external stakeholders.

Technology SSC units themselves, as service providers, have several positive findings of shared services approach as well, as follows:

- they can now focus just on activities moved to their units and become highly specialized, thus more effective,
- resources are better utilized due to decrease of common and repetitive tasks,
- possibility to achieve additional revenue, and
- as they are now representing whole group, SSCs became more visible and respected entities within the group and towards external partners at the same time.

It should be clear now that the implementation of the SSC business model is beneficial to all three crucial parties involved, national companies, corporate group organization and SSC units. In spite of all the challenges in the implementation stage, SSCs are proven to be the successful strategic direction for big international organizations like telecom provider in this example. Finally, a benefit as an outcome of SSC implementation in telecom environment is more efficient and more cost effective business model that works.

IV. CONCLUSION

Shared services have risen in both importance and visibility and have become ubiquitous among large international companies. In this paper, the concept of shared services was described, together with characteristics and drivers for SSC implementation. Focusing on a specific example of Technology SSC model introduction in telecom environment, challenges both organizational and technological, were described together with the benefits of such an implementation. While describing Shared Service Centers, aim of this paper was to see them as a new business model that is more efficient and more cost effective suitable for big organization as global telecom service provider.

To conclude, SSCs have many benefits, as shown in this paper, and from an economic prospective are undeniable. In terms of human relations and alignment across multiple companies there could be quite a few challenges while implementing and supporting this business model. Advancing NT/IT technology is another beneficial factor in supporting this model. The shared service model is not without its limitations, however. Shared services can live up to its potential only when the parent organization is above a minimum size. In addition, achieving a successful SSC implementation requires attentiveness to the corporate culture, appropriate technologies, and investing the time and resources to manage all the expectations. According to the described example of Technology SCC implementation in this paper, it can be concluded that most of the expectations have been successfully met.

LITERATURE

[1] R.S. Kaplan, D.P. Norton, "The strategy focused organizations", Harward Business School Press, 2001.

[2] V. Schulz, W. Brenner, "Characteristics of shared service centers", Transforming Government: People, Process and Policy, Vol. 4 Iss: 3, pp.210 – 219, 2010.

[3] J. Strikwerda, "The Shared Service Centre: Change, Governance and Strategy", Business School, University of Amsterdam, 2006.

[4] B. Bergeron, "Essentials of Shared Services", John Wiley & Sons Inc., Hoboken, New Jersey, 2003.

[5] H. Wang, "Performance analysis for Shared Services", Communications of the IIMA, Vol. 7 Iss:2, 2007.

[6] V. Schulz, A. Hochenstein, F. Ubernickel, W. Brenner, "Definition and Classification of IT-Shared-Services-Center", AMCIS 2009 Proceedings, Paper 265, 2009.

[7] V. Schulz, T. P. Herz, M. A. Rothenberger, W. Brenner, "IT Shared Service Center and External Market Activities", AMCIS 2010 Proceedings, Paper 532, 201.

ICT APPLICATIONS

Usluga mobilnog plaćanja računa m:pay

Vesna Žlof*, S.Salapura, PhD*

* Mtel a.d. Banja Luka, BiH

vesna.zlof@mtel.ba, sasa.salapura@mtel.ba

Sažetak - U eri masovne upotrebe mobilnih komunikacija u poslovne svrhe mobilno plaćanje i mobilna trgovina postaju sve dominantniji načini obavljanja finansijskih transakcija. Moderni ljudi svoje poslove obavljaju ubrzano pri čemu u pomoć pozivaju moderne tehnologije. Korisnici mobilnog plaćanja imaju mogućnost izbora više platformi i informacionih sistema koji su prisutni na tržištu. Za velike kompanije su posebno atraktivni EBPP sistemi (electronic bill presentment and payment) jer uvode revoluciju u proces naplate računa nudeći online i u realnom vremenu uvid u sadržaj računa a potom izbor načina i vremena obavljanja transakcija. Potencijal EBPP iskazuje se kroz B2C (Business to Consumer) i B2B (business to business) procese koji mogu biti unificirani. Na ovaj način se značajno poboljšava briga i način upravljanja odnosa s korisnicima (CRM), te kreira dodatna vrijednost od strane izdavaoca računa.

U ovom radu biće predstavljena usluga mobilnog plaćanja m:pay koja krajnjim korisnicima kompanije Mtel omogućava plaćanje računa za telekomunikacione usluge te dopunjavanje prepaid računa. Biće opisan način rada servisa, arhitektura sistema, veze sa drugim sistemima/servisima te tehnologije razmjene podataka i izvještavanja.

Ključne riječi - m:payment, mobile payment, EBPP sistemi (electronic bill presentment and payment), B2C (Business to Consumer), B2B (business to business)

Abstract - In the time of mass use of mobile communication for the business purposes, mobile payment and mobile commerce become dominant way of performing financial transactions. Modern business people conduct their tasks fast with the help of modern technology. Mobile payment users are given a possibility to choose between several platforms and information systems present on the market. Large companies find EBPP (electronic bill presentment and payment) systems especially attractive because they introduce revolution in the process of bill collecting offering online insight at real time into the content of the bill as well as the selection of the manner and time of conducting transactions. The potential of EBPP is presented through B2C (Business to Consumer) and B2B (Business to Business) processes which can be unified. In this way the care for and management of the customer service (CRM) are significantly improved, and additional value is created by the bill issuer.

This paper presents the mobile payment service m:Pay which enables end-users of Mtel company to pay their bills for telecommunication services and to top up their prepaid accounts. It will provide description of the service performance, system architecture, connection to other systems/services and technology for data exchange and reporting.

Key words: m:payment, mobile payment, EBPP systems (electronic bill presentment and payment), B2C (Business to Consumer), B2B (business to business)

I. UVOD

U skladu sa svjetskim trendovima, vođena prvenstveno preferencijama sopstvenih korisnika, kompanija Mtel je u saradnji sa partnerima uvela uslugu mobilnog plaćanja računa za telekomunikacione usluge kao i dopunjavanje prepaid računa upotrebom mobilne aplikacije. M:pay, kako je usluga nazvana, je potpuno novi servis iz domena finansijskog poslovanja koji korisi sve prednosti koje donose moderne tehnologije za prenos podataka implementirane u telekomunikacionoj mreži Mtela te omogućava korisniku da iskoristi svoj mobilni telefon u punom kapacitetu i za potrebe izvršavanja finansijskih transakcija, pored raznolikosti drugih postojećih servisa različite namjene.

Iako neke studije slučaja prikazuju povećani opticaj sredstava (cash flow) i do 30% i to samo smanjivanjem vremena proteklog od izdavanja do naplate računa, te dodatne uštede ostvarene uvođenjem 'bespapirnog' načina poslovanja, ne treba zapostaviti očekivano drastično smanjivanje broja grešaka nastalih pogrešnim popunjavanjem podataka u nalogu za plaćanje [9].

Servis mobilnog plaćanja omogućava korisnicima Mtela jednostavnu, sigurnu i pouzdanu proceduru plaćanja računa za telekomunikacione usluge kao i dopunjavanje prepaid računa korisnika direktno sa bankovnog računa koji korisnik mpay usluge posjeduje u odgovarajućoj finansijskoj instituciji sa kojom je Mtel realizovao ovaj servis.

Plaćanje mobilnom aplikacijom omogućeno je samo korisnicima Mtela i to isključivo putem smartphone aplikacija koje na ovaj način preuzimaju jedan segment poslova plaćanja koje tradicionalno nude poslovne banke, te tako prenose kompletno iskustvo elektronskog bankarstva na mobilni terminal. Da bi komunikacija

između svih poslovnih subjekata bila potpuna i funkcionalna, Mtel je za potrebe mPay usluge implementirao mpayment sistem koji služi kao gateway gdje se terminiraju zahtjevi korisnika inicirani preko mobilne aplikacije s jedne strane, dok sa druge strane pomenuti sistem ima realizovanu komunikaciju sa bankarskim sistemom plaćanja. Komunikacija se obavlja razmjenom unaprijed usaglašenih specijalizovanih tipova poruka u XML formatu a korištenjem WEB servisa podignutnih na obe strane, kako na strani sistema Mtela tako i na strani sistema banke.

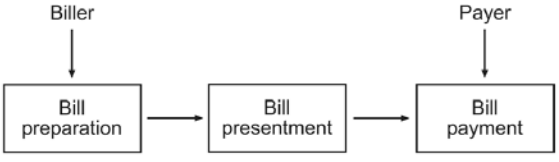
Upravo ovim modelom razvijenog poslovnog procesa kao i primjenom mobilne aplikacije u pomenutom poslovnom okruženju, Mtel je izašao u susret jednom novom izazovu u cilju proširenja spektra postojećih usluga kako bi na tržištu bio prepoznat kao operator koji je svojim korisnicima omogućio lakše, brže i efikasnije poslovanje upotrebom mobilnih aplikacija.

Zato se u ovom radu razmatra značaj mobilnih aplikacija u postizanju konkurentnosti institucija različitih djelatnosti, kako na zadovoljstvo svojih korisnika, tako i na povećanju efikasnosti zaposlenih uz ostvarivanje korporativnih ciljeva kompanije.

II. USLUGA EBPP (ELECTRONIC BILL PRESENTMENT AND PAYMENT)

EBPP je termin koji se koristi za predstavljanje mogućnosti prezentovanja računa krajnjem korisniku i plaćanja tih računa putem elektronskih alata. Korištenje EBPP usluge je u mnogim kompanijama prepoznato kao mogućnost smanjenja troškova za štampanje računa i dostavljanje krajnjem korisniku s jedne strane, odnosno povećanja efikasnosti poslovnih procesa bržom i sigurnijom dostavom računa elektronskim putem s druge strane, omogućenoj preko pouzdanog kanala komunikacije između same kompanije i krajnjeg korisnika. Pored samog dostavljanja i prikazivanja računa, EBPP obuhvata i obezbjeđuje takođe i metode plaćanja računa, tako da se korisniku, nakon što se obezbijedi uvid u pregled računa, ponudi i mogućnost da taj račun i plati sa istog mjesta bez potrebe odlaska na šalter date kompanije ili finansijske institucije preko koje bi uradio datu transakciju.

Jedan takav model poslovanja je prikazan na slici:



Slika 1. EBPP elementarni proces; Izvor [1].

Osnovni elementi EBPP poslovnog modela su priprema računa, prezentacija računa i plaćanje. I pored toga što je EBPP termin uveliko korišten u billing industriji, postoje i sljedeći termini korišteni od strane različitih kompanija koje generišu račune, distributera, agregatora pa i krajnjih korisnika:

- IBPP (Internet bill presentmentand payment)
- EBP (electronic bill presentment)
- OBPP (online bill presentment and payment)
- ED&EP (electronic delivery and electronic payment)
- BIPS (bill and invoice presentment and settlement)

Za proces kreiranja računa, izdavanja računa, njihova prezentacija kao i proces poravnanja uplata postoji više načina koji omogućavaju direktnu vezu između kompanija koje izdaju račune i njihovih krajnjih korisnika koji te račune plaćaju. Jedan od načina je putem kontakt centara datih kompanija, customer-care rješenja i prodajnih tačaka, dok sa druge strane imamo moderniji pristup zasnovan na korištenju mobilnih i web aplikacija koje se podrazumijevaju kao integrisani paket CRM (Customer relationship management) modela.

Značajan interes za uvođenje EBPP usluge imaju telekom kompanije, komunalne kompanije, osiguravajuće kuće kao i bankarski sektor. Pored ovih, značajan interes će u budućnosti pokazati i ostali segmenti industrije, kao npr. sektor zdravstvene zaštite, državne administrativne službe, veliki prodajni lanci i slično. Konkretno u slučaju telekom operatora, EBPP je atraktivan zbog mogućnosti integrisanja podataka na jednom mjestu, kao npr. online računi za telekom usluge, računi za razlićite komunalne usluge i slično.

III. M:PAY MOBILNI SERVIS ZA PLAĆANJE TELEKOMUNIKACIONIH RAČUNA – ARHITEKTURA I OPIS SISTEMA

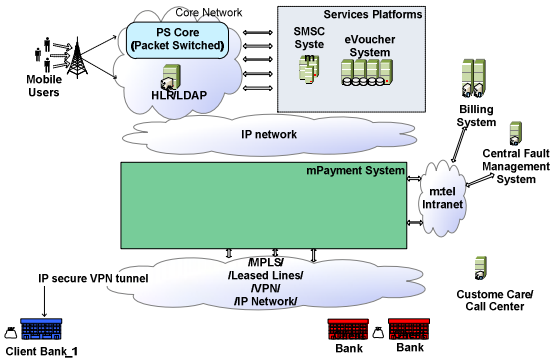
M:pay mobilni servis za plaćanje omogućava korisnicima kompanije Mtel, sigurno i pouzdano plaćanje telekomunikacionih usluga putem mobilne aplikacije na Android i iOS mobilnoj platformi. Pored plaćanja telekomunikacionih usluga, ovim novim kanalom je korisnicima omogućena i usluga dopunjavanja prepaid računa Mtelovim prepaid korisnicima.

Osnovne funkcionalnosti koje mobilna aplikacija omogućava su aktiviranje m:Pay usluge, prijem i plaćanje vlastitih korisničkih računa, plaćanje korisničkih računa za treća lica, dopunjavanje prepaid broja, pregled urađenih transakcija i osnovna podešavanja korisničkih parametara, kao što je lični identifikacioni broj (PIN - Personal identification number) za pristup aplikaciji.

Uzimajući u obzir kompleksnost implementacije rješenja koje bi omogućilo ovakve funkcionalnosti, Mtel je identifikovao različite subjekte koje je bilo potrebno uključiti u realizaciju:

- Poslovni/rezidencijalni korisnici Mtela i odgovarajuće finansijske institucije
- Mrežna infrastruktura Mtela
- Integrisano mpayment rješenje unutar Mtela
- Finansijske institucije sa kojima je Mtel razvio poslovno-finansijsku saradnju,

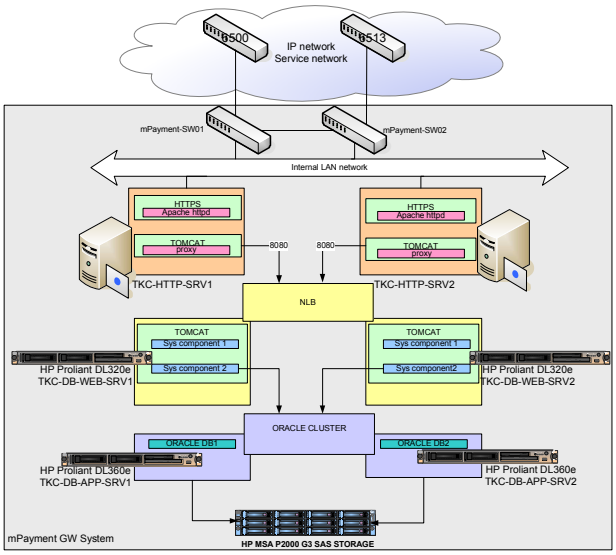
kao što je predstavljeno na slici 2.



Slika 2. Integracija mpayment rješenja u Mtel mreži i eksterni subjekti

Zahtjevi koje korisnik inicira putem mobilne aplikacije, preko mobilnog interneta odnosno mreže za paketski prenos saobraćaja u Mtel mreži, dolaze na centralizovani mpayment sistem, koji ima ulogu gateway-a koji omogućava direktnu komunikaciju između krajnjeg korisnika i njegovog računa u odgovarajućoj finansijskoj instituciji.

Detaljna arhitektura mpayment centralizovanog mrežnog elementa je prikazana na sljedećoj slici:



Slika 3. Arhitektura mPayment GW sistema

Arhitekturu mPayment gateway sistema čine dva switching mrežna elementa, dva HTTP (Hypertext Transfer Protocol) proxy servera, dva web servera, dva database servera i storage sistem. Za serverske komponente sistema izabrane su odgovarajuće komponente proizvođača HP (Hewlett-Packard Company) sa pouzdanim karakteristikama, dok su za switching komponente odabrani Cisco pouzdani switch mrežni elementi.

Kompletna komunikacija mpayment sistema, nevezano da li se radi o internim sistemima unutar Mtel mreže, ili eksternim sistemima kao što je informacioni sistem banke, odvija se preko switching komponenti, odnosno preko IP (Internet Protocol) servisne mreže Mtela.

U cilju zaštite korisničkih podataka i kompletne baze, http funkcionalnost je od ostatka mreže mpayment sistema izdvojena u DMZ (demilitarized zone) segment.

Uloga web servera je prihvatanje, obrada i prosljeđivanje zahtjeva, kako od strane krajnjeg korisnika, tako i u komunikaciji sa finansijskom institucijom. Na web serverima je podignuta softverska funkcionalnost NLB (Network Load Balancing) kako bi se postigla pravila raspodjela zahtjeva i omogućila redudansa na aplikativnom nivou.

Iza web servera u arhitekturi nalaze se database serveri na kojima je podignut klaster (Oracle cluster) koji omogućava pouzdanu i sigurnu bazu za smještanja korisničkih podataka, podataka o svim transakcijama, istoriji i slično.

Kompletna struktura je redudantna što u slučaju ispada pojedininh sistemskih komponenti omogućava neometan rad i nastavak pružanje servisa, a sve u cilju zadovoljenja potreba krajnjeg korisnika.

IV. INTERFEJSI SA EKSTERNIM SISTEMIMA

Kao što je prikazano na slici 2, mPayment sistem je uvezan sa različitim elementima i sistema unutar vlastite Mtel mreže ali i vanjskim sistemima.

Postojeći sistemi unutar vlastite mreže koji se koriste u mpayment strukturi su LDAP, SMSC, mail server, eVoucher, Central Fault Management System, Billing, O&M clients i biće detaljnije objašnjeni u nastavku ovog rada. Vanjski sistemi sa kojima je ostvarena veza su postojeći poslovni sistemi komercijalnih banaka.

Komunikacija sa sistemom banke je ostvarena preko IP mreže i odvija se preko VPN tunela (Virtual private network) koji je podignut i konfigurisan uz najveće IP standarde sigurnosti direktno između servera na kojem je podignut WEB servis na strani banke te mpayment sistema Mtela. Razmjena podataka ostvarena je prema unaprijed usaglašenoj specifikaciji razmjene XML (EXtensible Markup Language) poruka između web servisa na obe strane [3].

Komunikacija sa LDAP (Lightweight Directory Access Protocol) bazom je neophodna u cilju provjere pripadnosti Mtel mrži. Provjera se vrši ldapsearch upitom u bazu podataka koja je smještena na HLR (Home location register) mrežnom elementu. Ovo je neophodno jer je regulator za telekomunikacije omogućio neometan prelazak korisnika iz mreže jednog operatora u mrežu bilo kojeg drugog operatora na području BiH, a uspostavljeni servis je ipak namijenjen isključivo korisnicima kompanije Mtel.

Komunikacija sa SMSC (Short Message service center) sistemima je ostvarena u cilju obavještavanja i slanja korisniku odgovarajuće notifikacije i jednog dijela koda za autorizaciju. Komunikacija sa SMSC sistemima je ostvarena putem SMPP (Short Message Peer-to-Peer) protokola. Drugi dio koda za autorizaciju se iz sigurnosnih razloga šalje putem SMTP protokola

(Simple Mail Transfer Protocol) na korisnikovu elektronsku poštu.

Komunikacija sa eVoucher[1] sistemom je ostvarena u cilju realizacije usluge dopunjavanja prepaid računa sredstvima koja se nalaze na korisnikovom računu u poslovnoj banci. Komunikacija se vrši preko HTTP protokola sa prema internoj specifikaciji za interfejs sa ovim sistemom.

Komunikacija sa billing sistemom je ostvarena u cilju preuzimanja podataka o prispjelim i plaćenim računima, kako bi se oni na adekvatan način prikazali korisniku kroz mobilnu aplikaciju. Dohvatanje informacija o korisničkim računima se vrši preko web servisa razvijenog specijalno za ovu namjenu.

Komunikacija i izvještavanje O&M segmenta (Operating and Maintenance) ostvarena je u cilju redovnog praćenja, održavanja i eksploatacije servisa i to na isti način na koji je realizovana i komunikacija sa operaterima u Kontakt centru kompanije Mtel, a su cilju bržeg rješavanja zaprimljenih korisničkih žalbi te brže reakcije na probleme pri plaćanju računa ili dopunjavanju prepaid broja.

Komunikacija sa CFM sistemom (Central Fault Management System) je ostvarena u cilju neprekidnog praćenja i identifikacije događaja i alarma generisanih od strane sistema kako na nivou grešaka na samom hardveru, tako i u slučaju grešaka na aplikativnom nivou.

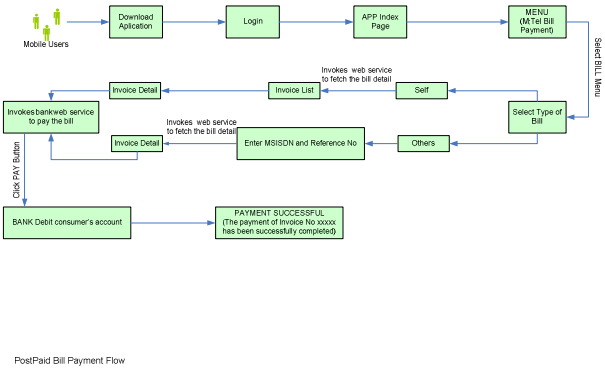
V. PRIMJER TOKA SERVISA I RAZMJENE PORUKA ZA PROCEDURU PLAĆANJA

Komunikacija između sistema banke i mPayment sistema Mtela vrši se razmjenom specijalizovanih poruka u XML formatu, koje su unaprijed dogovorene i usvojene od strane inženjera i razvojnog tima banke i Mtela [3]. Poruke su podijeljene u posebne grupe i markirane odgovarajućim identifikatorima, kako bi se znalo na koju konkretno proceduru se odnose. Tako imamo poruke za proceduru registracije korisnika, poruke vezane za evidenciju i plaćanje faktura, poruke vezane za dopunu prepaid broja, specijalizovane sistemske i finansijske poruke.

Nakon što se korisnik registrovao za korištenje mpay usluge, sistem banke šalje odgovarajući zahtjev u formi xml poruke sa osnovnim podacima korisnika, prema mpayment sistemu, koji nakon provjere da li je korisnik Mtelov, vrši registraciju korisnika u svojoj bazi. Korisniku se šalju kodovi za autorizaciju i link sa kojeg može da izvrši preuzimanje mobilne aplikacije u zavisnosti od operativnog sistema njegovog mobilnog terminala. Nakon što korisnik instalira aplikaciju, unese svoje kodove koje je dobio od sistema i na taj način izvrši proces personalizacije, on biva potpuno registrovan i spreman za korištenje mpay usluge.

Na sljedećoj slici je prikazan algoritam za proceduru plaćanja računa:

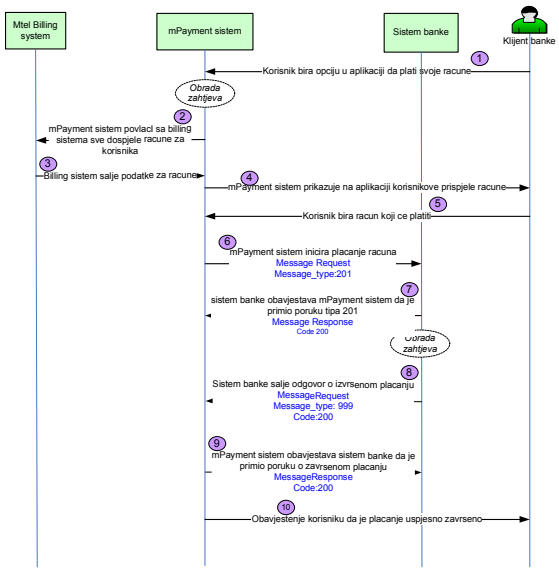
[1] TopUp - Sistem elektronske nadoplate računa



PostPaid Bill Payment Flow

Slika 4. Algoritam za proceduru plaćanja računa

Korisnik ima mogućnost plaćanja sopstvenih računa i računa trećih lica na osnovu referentnog broja. Nakon što korisnik pokrene mobilnu aplikaciju te unese svoj PIN za ulaz u aplikaciju, njemu se prikazuje glavni meni sa opcijama. U glavnom meniju korisnik bira opciju računi, a potom bira podmeni u kojem postoje opcije dospjeli računi, plaćeni, neplaćeni i plaćanje po broju (opcija plaćanja računa trećih lica). Nakon izbora vlastitog računa pod opcijom Dospjeli računi, korisnik klikom na prikazani račun ima uvid u podatke tog računa, kao što su iznos računa, opis, dospijeće, datum izdavanja računa i slično, gdje samo jednim klikom na opciju Platite inicira plaćanje izabranog računa. Time se inicira komunikacija sistema sa web servisom banke. Tok servisa i razmjene poruka između uključenih strana u ovoj proceduri, prikazan je na slici:



Slika 5.Dijagram toka servisa za proceduru plaćanja računa

Nakon što korisnik na mobilnoj aplikaciji potvrdi plaćanje izabranog računa, mPayment sistem šalje zahtjev banci, u vidu xml poruke označene sa tipom 201:

```
<?xml version="1.0" encoding="utf-8"?>
<soap:Envelope
xmlns:soap=http://www.w3.org/2001/12/soap-envelope
soap:encodingStyle="http://www.w3.org/2001/12/soap-encoding">
          <soap:Body xmlns:m="http://www.mtel.ba/pgw">
          <m:MessageRequest>
          <m:MessageID>MT123456789012345678900</m:MessageID>
                    <m:MessageType>201</m:MessageType>
          <m:UserID>1234567890</m:UserID>
                    <m:AccountID>1234567890</m:AccountID>
```

```
                <m:PB>1234567890123</m:PB>
                <m:MSISDN>38765xxxxxx</m:MSISDN>
                <m:Amount>12.10</m:Amount>
                <m:Currency>BAM</m:Currency>
          <m:PurposeOfPayment>Uplata racuna fiksne
        telefonije</m:PurposeOfPayment>
          <m:BeneficiaryAccount>55500712xxxxxxx</m:Beneficiary
        Account>
          <m:BeneficiaryName>Mtel</m:BeneficiaryName>
          </m:MessageRequest>
          </soap:Body>
</soap:Envelope>
```

Sistem banke na ovaj zahtjev odgovara statusnom porukom tipa 200, kao indikacija da je poruka uspješno primljena. Nakon što banka izvrši odgovarajuće provjere na svojoj strani (npr. korisnik ima dovoljno sredstava na računu) šalje mpayment sistemu Mtela xml odgovor u formi zahtjeva označenim sa tipom poruke 999, koja sadrži kod odnosno identifikaciju da li je plaćanje uspješno završeno (kod 200) ili nije (neki drugi kod):

```
<?xml version="1.0" encoding="utf-8"?>
<soap:Envelope
xmlns:soap="http://www.w3.org/2001/12/soap-envelope"
soap:encodingStyle="http://www.w3.org/2001/12/soap-encoding">

<soap:Body xmlns:m="http://www.mtel.ba/pgw">
        <m:MessageRequest>
<m:MessageID> MT12345678901234567890</m:MessageID>
          <m:MessageType>999</m:MessageType>
          <m:MessageRef>NB12345678901234567890</m:MessageRef>
          <m:Code>200</m:Code>
          <m:Comment>Sve je OK</m:Comment>
           </m:MessageRequest>
</soap:Body>
</soap:Envelope>
```

referencirajući se na ID poruke za koju je Mtel u prethodnom koraku poslao zahtjev za plaćanje.

Na sličan način, samo sa drugačije usvojenim identifikatorima poruka, vrši se razmjena poruka za sve ostale procedure u komunikaciji sa sistemom banke.

VI. ZAKLJUČAK

U ovom radu je opisan servis mobilnog plaćanja m:pay koji Mtel korisnicima omogućava jednostavnu, sigurnu i pouzdanu proceduru plaćanja računa za telekomunikacione usluge te dopunjavanje prepaid računa korisnika direktno sa bankovnog računa koji korisnik mpay usluge posjeduje u nekoj od poslovnih banaka sa kojom je Mtel realizovao ovaj servis.

U ovom radu je opisano više izazova sa kojim su se projektanti suočili pri donošenju odluka koje su osim tehničkih obuhvatale i pravne i poslovne aspekte poslovanja kompanije.

Pokazano je da je moguće i opravdano da kompanija u procesu upravljanja računima osim formiranja i slanja računa krajnjem korisniku, preuzme i svojim korisnicima ponudi i izbor metode plaćanja kao i dinamiku plaćanja.

Kako je ovaj rad pisan u vrijeme komercijalnog početka servisa, u ovom trenutku nije moguće dati detaljniju informaciju o načinu na koji su korisnici doživjeli ponuđeni servis. Takođe, ne zna se ni koje osobine servisa su dobro planirane a koje će biti mijenjane te prilagođene ocjeni krajnjeg korisnika. Sve to treba biti obuhvaćeno u narednim radovima.

Nakon što servis bude proširen za poslovanje sa ostalim poslovnim bankama koje su prisutne na tržištu, a posebno dodavanjem komunalnih preduzeća čiji računi će biti raspoloživi za plaćanje, možemo očekivati veću prihvaćenost servisa od strane korisnika koji će u konačnici dobiti nove načine informisanja i plaćanja.

LITERATURA

[1] K.Terplan, "Electronic Bill Presentment and Payment", Copyright © 2003 CRC Press, LLC

[2] „Supply and Installation of the mPayment/ePayment Solution in Mtel network“, Interni dokument, Mtel, Banja Luka, 2013

[3] “Protokol razmjene poruka između mpayment sistema i banke”, Interni dokument, Mtel, 2013

[4] Personix, "Electronic Bill Presentment and Payment - The Gateway to One-on-one Customer Relationships", Whitepaper, Personix, 2007

[5] ISACA, "Mobile Payments: Risk, Security and Assurance Issues", Emerging Technology, 2011

[6] "eVoucher Interface Specification", Interni dokument, Mtel, Banja Luka, 2007

[7] Hayes B., Judy H., "Privacy in Cyberspace", Computer Security Handbook, 4th Edition, 2002

[8] „Zakon o zaštiti ličnih podataka“, Službeni glasnik BiH, 2002

[9] Prodapt, EBPP solution for a leading South African Telecom Operator - Case Study, 2011

PRAĆENJE IMOVINSKO PRAVNIH POSLOVA NA ELEKTRONIČKOJ KOMUNIKACIJSKOJ INFRASTRUKTURI PUTEM WEB GIS APLIKACIJE

Dražen Salopek [1], Tomislav Đigaš [1], Franjo Ambroš [2], Mario Štimac [2]:

[1] HT Hrvatski telekom d.d., Zagreb, Hrvatska

[2] GEOprem d.o.o., Osijek, Hrvatska

drazen.salopek@t.ht.hr; tomislav.digas@t.ht.hr; franjo.ambros@geoprem.hr; mario.stimac@geoprem.hr

Sažetak - Složenost imovinsko pravnog uređenja elektroničke komunikacijske infrastrukture posljedica je neuređenih evidencija o nekretninama, nepostojanja podataka o položaju dijelova trasa elektroničke komunikacijske infrastrukture te o velikom broju titulara nekretnina u Hrvatskoj. Praćenje radnih procesa koji se odvijaju na području cijele države, s većim brojem izvođača radova te dislociranim korištenjem informacija ne može se efikasno organizirati bez odgovarajuće informatičke podrške. U Hrvatskoj postoji više vlasnika elektroničke komunikacijske infrastrukture i značajan broj infrastrukturnih operatora koji obavljaju telekomunikacijske djelatnosti na zakupljenoj infrastrukturi. Poslovni procesi zahtijevaju korištenje analognih i digitalnih baza podataka nastalih u raznim vremenskim periodima. Ovo pred informatičare postavlja specifične probleme - kako od nestandardnih podataka dati pouzdanu informaciju. Podaci potrebni za ustrojavanje aplikacije za praćenje radnih procesa u imovinsko pravnom uređenju infrastrukture sastoje se od temeljnih podataka katastra i zemljišne knjige, podataka katastra vodova i podataka sadržanih u prostornim planovima koji se odnose na namjenu korištenja zemljišta. Pravni režimi kojima podliježu nekretnine često su posljedica raznih zakonskih ograničenja a manje volje vlasnika. U javne evidencije o nekretninama se pravna ograničenja upisuju pojedinačno rezultat čega je velika ovisnost točnosti podataka o vremenu njihovog korištenja. Rješenje WEB GIS (geografski informacijski sustav dostupan putem mreže), otvorenog tipa, može pomoći u vođenju poslovnih procesa i u situacijama kada izvore podataka čine razne dislocirane baze podataka.

I. UVOD

Područje elektroničkih komunikacijskih usluga spada u veoma intenzivno razvojno tehnološko područje. Karakteriziraju ga brza izmjena tehnologije i velik utjecaj dostupnosti usluge na bruto društveni proizvod (BDP). Izgrađena elektronička komunikacijska infrastruktura (EKI) države uvjet je njenog razvitka. Hrvatska baštini dugu tradiciju primjene telekomunikacijskih usluga, prvenstveno javne telegrafske i telefonske usluge te radio i televizijske usluge, baziranih na analognom prijenosu signala. Svaka od ovih usluga imala je više vlasnika infrastrukture koja je služila za prijenos usluge od točke A-B. Dio infrastrukture je bio u funkciji javnih telekomunikacijskih usluga, a dio je namijenjen funkcionalnim sustavima veza. Izgradnja složenih sustava kao što su telekomunikacijske mreže fokusiranih na javne telekomunikacijske usluge počela je 80-tih godina prošlog stoljeća i intenzivno se odvijala više od dvadeset godina. Ove mreže su u početku projektirane bakrenim vodičima (koaksijalni i simetrični kabeli), da bi pred završetkom ciklusa izgradnje svjetlovodi zamijenili bakrene vodiče. Digitalizacija mreže obuhvatila je komutacijske čvorove, dok je za prijenos korištena postojeća bakrena infrastruktura. Nesklad javnih evidencija o nekretninama je usporavao reguliranje imovinsko pravnih odnosa s vlasnicima privatnih nekretnina, dok se korištenje nekretnina u sustavu društvenog vlasništva samo po sebi podrazumijevalo kao izgradnja infrastrukture od interesa za državu na „ničijoj nekretnini“. U to doba dobiti telefonski priključak je bila stvar prestiža pa, u pravilu, vlasnici privatnih nekretnina nisu tražili obeštećenje za služnost. Ukoliko je tempo investicija i postupak ishođenja građevinskih dozvola bio sporiji uspjeli su se riješiti i imovinskopravni odnosi s vlasnicima nekretnina. Tragovi ovih aktivnosti nalaze se u nekim gruntovnica u Hrvatskoj. WEB GIS interpretacija uknjiženih prava za katastarske općine Osijek i Novigrad vidljiva je na slici 1. Brzo uređenje imovinskopravnih odnosa uvjetovani je novim zakonskim odredbama [1,2] i interesom da se osigura daljnja modernizacija ove djelatnosti. Praćenje ovih aktivnosti raspršenih na cijelom prostoru Hrvatske nezamislivo je bez WEB GIS podrške.



Slika 1. K.O. Novigrad i K.O. Osijek uknjižene služnosti u korist PTT (prednika HT Hrvatskog telekoma d.d.)

II POVIJEST PODATAKA O NEKRETNINAMA

U Hrvatskoj postoje dvije evidencije o nekretninama katastar i zemljišna knjiga (gruntovnica). Ove evidencije baštinimo iz vremena Austrougarske monarhije. Osnovna pravila o evidentiranju nekretnina nisu se bitno mijenjala. Tehnološki napredak je pomogao da se te dvije evidencije vode sve suvremenijim metodama, uz neznatno mijenjanje zakonskog okvira koji je utjecao na njihovo održavanje. Katastar kao tehnička evidencija nekretnina prikazanih kao katastarske čestice sa svojim granicama i brojem te prikazom građevina na njima bio je namijenjen fiskusu (naplati poreza). Organizacijski brigu oko katastra vodi Državna geodetska uprava, jedna od osam državnih upravnih organizacija. Ministarstva nadležna za tu upravu su se mijenjala ovisno o ulozi geodetske djelatnosti na društvo (poljoprivreda, graditeljstvo, zaštita okoliša i dr.). Zemljišna knjiga je uvijek bila pod nadležnošću Ministarstva pravosuđa, brigu oko njene funkcionalnosti vode općinski sudovi. Registri koje vode te institucije spadaju u javne registre države i u neku ruku vlasnicima garantiraju njihovo vlasništvo. Katastar je kroz svoje evidencije usmjeren na posjed, odnosno registraciju svakodnevnog korištenje nekretnina od zainteresirane osobe-posjednika. Pri čemu su dvije bitne evidencije katastarskog operata: grafička, katastarski planovi i pisana, posjedovni listovi. Digitalizacija pisanog dijela katastarskog operata započela je sedamdesetih godina prošlog stoljeća i spada u prve informatizirane registre države. Od 2003. godine javni registri o nekretninama se sustavno uređuju pod nazivom Projekt sređivanja zemljišnih knjiga i katastra. Projekt je započeo digitalizacijom zemljišnoknjižnih uložaka prepisivanjem ručno vođenih preko 4,2 mil. zemljišnoknjižnih uložaka, u kojima je bilo evidentirano više od 14,4 mil. katastarskih čestica. Analiza o broju upisanih titulara nije rađena. Procjena je da je to oko 60 mil. titulara u raznim suvlasničkim i samovlasničkim kombinacijama. Digitalizacija pisanog i grafičkog dijela katastarske evidencije provodila se također sustavnim preoblikovanjem katastarskih planova na papiru u digitalni oblik (DKP). Oko 75% tih planova nastalo je u devetnaestom stoljeću. Usklađenost tik planova sa stanem na terenu po pitanju položaja i granica katastarskih čestica veoma je upitna. Razlozi tome nisu predmet ovog rada, ali se ne mogu zanemariti u interpretaciji podataka na temelju kojih se ostvaruju neka prava ili reguliraju odnosi među građanima. Od 2011. godine je zainteresiranim subjektima na raspolaganju digitalni ortofoto plan (DOF), koji može poslužiti za kontrolu stanja na terenu i DKP. Danas su javni registri o nekretninama dostupni preko web servisa: katastar.hr, pravosudje.hr i geoportal.hr. U Hrvatskoj je u implementaciji projekt: Zemljišnoinformacijski sustav (ZIS) koji će objediniti podatke katastra i zemljišne knjige. Blagodati ovih napora tek se očekuju, iako već danas zvanične zemljišnoknjižne izvatke koje vodi jedan općinski sud moguće je dobiti od drugog općinskog suda ili javnog bilježnika. Također je svakom zainteresiranom omogućeno da vidi zemljišnoknjižno stanje za bilo koju česticu u Hrvatskoj. Za pretraživanje služi identifikator, katastarska općina i broj katastarske čestice. Pretraživanje javnih registara nije moguće po imenima vlasnika.

Svim ovim akcijama metodom iteracije povećavamo kvalitetu evidencija o nekretninama. Broj katastarskih čestica u odnosu na broj stanovnika Hrvatske ukazuje da će novim naporima trebati pronaći motivaciju okrupnjavanja nekretnina odnosno isknjižavanja nepotrebnih suvlasnika iz javnih evidencija.

III OPSEG PODATAKA

Ovo sve ukazuje na složenost i dinamiku evidencija o nekretninama. One se u realnom svijetu geodetskim izmjerama preoblikuju, mijenjaju identifikatore, a pravnim poslovima mijenjaju titulara, prostornim planovima ili zakonskim rješenjima uvode se ograničenja raspolaganja, preuzimaju ulogu jamca u kreditnim aražmanima. Ako tome dodamo da je na 8,72% [3] svih parcela u Hrvatskoj izgrađena podzemna EKI u vlasništvu samo Hrvatskog telekoma i tome dodamo zračnu mrežu za koju ovog trenutka nemamo statistička istraživanja vidimo da se procesi imovinskopravnog uređenja moraju sustavno pratiti. Potrebno je ukazati i na još jedan identifikator koji se koristi za objekte izgrađene na katastarskim česticama, a to je adresa. Adresni sustav ukazuje na lokaciju potencijalnog korisnika EKI. Čine ga naziv ulice/trga i dodijeljeni kućni broj. Logika dodjele kućnih brojeva je prilagođena tipu naselja i u funkcije je lakšeg snalaženja i pronalaženje potencijalnog odredišta za razne usluge (poštanske, komunalne i dr.). Adresni sustav nema matematičku logiku, ali se digitalizacijom adresa mogu koristiti sve funkcije sadržane u GIS logici prezentacije podataka. U Hrvatskoj, prema popisu stanovništva 2011. [4] imamo 1,923 mil. stalno nastanjenih i 334 tisuća privremeno nastanjenih stambenih jedinica, koje imaju adresu. Malo detaljnijom analizom adresnog sustava baziranog na alfanumeričkom identifikatoru zgrade uočavamo nelogičnosti sustava. Slijedom te logike adresa zgrade može biti samo atribut u bazi. Digitalizacijom zgrade i pridodavanje koordinata centroidu ili poziciji ulaza zgrade osiguravamo korektnu interpretaciju adrese kao identifikatora lokacije.

Kad govorimo o opsegu digitalnih podataka o trasama EKI prvenstveno govorimo o uređenim poli-linijama kao poveznicama digitalnih čvorova i korisnika telekomunikacijskih usluga. Procjena Hrvatske regulatorne agencije za mrežne djelatnosti za 2013. godinu je 81.100 km podzemnih i nadzemnih trasa. Od toga 48.100 podzemnih (kabelske kanalizacije 21.800, kabela direktno položenog u zemlju 26.300 kilometara) i 33.000 kilometara trasa zračnih kabela.

Elektroničku komunikacijsku infrastrukturu u fiksnoj mreži čine i povezani objekti i oprema smješteni na području obavljanja telekomunikacijskih usluga. Ovi se objekti u bazi vode kao točkasti ili poligonalni elementi. Ukoliko je njihov tlocrt u naravi takav da ukazuje na značajno zauzimanje prostora koje treba posebno imovinsko pravno regulirati (obveze proizašle iz Zakona o državnoj izmjeri i katastru nekretnina) njihov prikaz u bazi tada mora biti poligonalan.

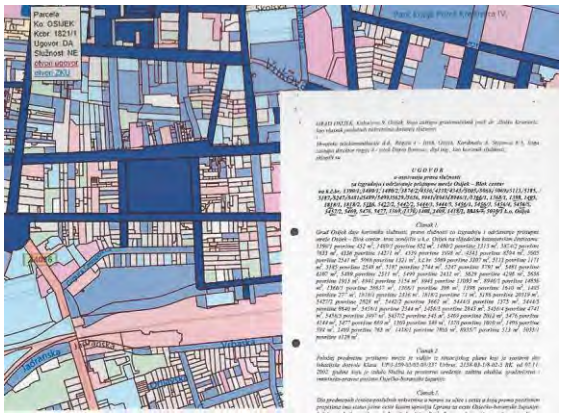
IV GENERALIZACIJA PRIKAZA U BAZI PODATAKA

U prirodi je čovjeka da saznanje o nekom događaju, pojavi ili objektu stvara na temelju raznih podataka. Pri

tome podatke koje skupi sintetizira i generalizacijom reducira/sažima. Ljudi se ovako ponašaju nesvjesno, ono je plod racionalnog pristupa i potrebe da se saznanja u vidu informacija prenesu drugima u realnom vremenu. Nesvjesno se ova logika uvukla i informatizaciju poslovnih procesa, a u geografskim informacijskim sustavima (GIS) ona je temelj interpretacije podataka. U GIS-u koristimo točku, liniju ili poligon kao nositelja informacije za lokaciju. Koliki nivo generalizacije ćemo primijeni ovisi o potencijalnim korisnicima. Korištenjem boja, tipova linija i oblika (u poligonalnoj interpretaciji), neke informacije posebno naglašavamo, dok druge potiskujemo, odnosno sugeriramo da imaju manju važnost za odluku o nekoj pojavi.

Dokument na temelju kojega temeljimo formiranje baze imovinskopravnog stanja EKI (ugovor o služnosti, potvrda o pravu puta i dr.) nastao je sustavom koordinirane djelatnosti tvrtke koja je započela pregovorima i završila potpisivanjem ugovora. Ugovor sadrži podatke interesantne za cijelu tvrtku (pravna služba, financije, služba održavanja). Svako gleda ugovor iz svoje perspektive. Pravnicima je interesantno ostvareno i po mogućnošću publicirano pravo, Službi za ekonomske odnose trošak tog ostvarenog prava, prikaz izdatka za to pravo jednokratno ili trajno, pronalaženje modela povrata kroz model amortizacije, praćenje rokova plaćanja. Službi održavanja je neophodno znati kada se može doći na nekretninu radi otklanjanja kvara, realizacije novog priključka, mjerenja i slično. Bez obzira na težnju unificiranja ugovora, imamo puno finesa koje se u ugovoru nalaze. Tome doprinose i zakonske odredbe koje u određenom vremenskom razdoblju egzistiraju, kao i jezične odredbe koje su posljedica lokalizama.

Stvaranjem baze koja bi sadržavale detalje ugovora dobili bismo mogućnost statističkog praćenja pojava, ali se postavlja pitanje svrsishodnosti takovog posla u ovoj fazi. U slučaju osporavanja vlasnika da operator koristi njegovu nekretninu iako postoji ugovor, potpisan od prednika, treba osigurati vidljivost tog ugovora. Jednostavno treba formirati link/vezu prema bazi ugovora i na taj način ugovor učiniti vidljivim (slika2).



Slika 2 Veza prema bazi ugovora

Često se za dokaz legalnosti traži građevinska dozvola (potvrda glavnog projekta) ili uporabna dozvola te je uputno da se i tu načini link na sve dokumente temeljem kojih je izvršena izgradnja elektroničke komunikacijske infrastrukture i povezane opreme.

Izgradnja EKI na nekoj parceli može biti u više navrata, s ugovorenim raznim pravnim ili ekonomskim obvezama. Proširenje elektroničke komunikacijske infrastrukture temelji se na dokumentima prostornog planiranja i građenja. O tome se mora voditi računa. Link na vidljivost ugovora, uporabnu dozvolu i sl., rješava ovaj problem jer čini datum pojedinih aktivnosti vidljivim.

V. TAJNOST PODATAKA

Poslovne procese usmjerene na elektroničku komunikacijsku infrastrukturu ponekad je potrebno voditi uz ograničenu transparentnost informacija o njima. Prije svega ono što se odnosi na podatke o krajnjem korisniku telekomunikacijske usluge kao i podatke iz ugovora sa vlasnicima nekretnina po pitanju identiteta i ugovorenoj naknadi za korištenje nekretnina. U digitalnom okruženju identitet osoba se lako može zloupotrijebiti pa se u postupku stvaranja baze podataka o tome mora voditi računa da se ne povrede odredbe Zakona o zaštiti osobnih podataka [5].

Podaci o pojedinoj nekretnini preuzeti su iz katastra zemljišta/nekretnina i nadležnog zemljišnoknjižnog suda. U skoroj budućnosti to će biti jedinstvena baza ZIS-a. Podaci sadržani u tim bazama javnog su karaktera iako sadrže osobne podatke vlasnika, kao što je ime, prezime, adresa prebivališta i osobni identifikacijski broj (OIB). U okviru te baze su i grafički podaci o nekretnini, oblik katastarske čestice i odnosi sa susjednim katastarskim česticama.

Podaci o trasama EKI preuzeti su iz katastra vodova i interpretiraju se zajedno sa digitalnim katastarskim planom koristeći logiku koordinatnog sustava, putem kojega geodeti interpretiraju podatke o položaju objekata prikupljenim u raznim vremenskim razdobljima. Katastar vodova je javna evidencija po pitanju položaja u prostoru.

VI TOČNOST PODATAKA

Hrvatska baštini evidenciju o nekretninama zasnovanim na katastarskoj općini kao teritorijalnoj jedinici na temelju koje se radi evidencija svih katastarskih čestica. To je u pravilu područje naselja s okolnim zemljištem. Takovih teritorijalnih jedinica u Hrvatskoj ima 3.346. Valja naglasiti da svega oko 25% teritorija Hrvatske izmjereno suvremenijim numeričkim geodetskim metodama. Zbog neadekvatnog održavanja ili manjkavosti tadašnje mjerne opreme, a prije svega propustima u izjednačenju geodetske osnove morat će se organizirati nova izmjera. Tom poslu se sustavno pristupilo prije dvadeset godina. Danas možemo reći da je na svega 3% teritorija Hrvatske provedena suvremena katastarska izmjera. Deklarirana točnost je 0,10-0,20 m unutar građevinskog područja, 0,40 m izvan građevinskog područja.

Analiza koja je rađena za potrebe Studije uspostavljanja nacionalnog integriranog geoiformacijskog sustava infrastrukture vodova [6] ukazuje na stupnjeve točnosti EKI: do 10 cm – 16,7%; od 10-30cm – 15,3%; od 30-50cm – 16,7%; od 50-100cm – 18,2%; od 1-5m – 31,3%; preko 5m – 1,8%. Obzirom da su u tijeku

geodetska snimanja vodova za očekivati je znatno više vodova čija će točnost zadovoljiti objektivne kriterije naplate korištenja nekretnina s osnova služnosti/prava puta.

Na ovo treba upozoriti kako ne bismo nerealno očekivali da se obračun naknade prema vlasnicima odvija s stopostotnom točnošću. Infrastrukturni operator neće biti zakinut jer se radi o koridoru trase, koja je geodetski snimljena u prostoru. Jedino vlasnici će biti zakinuti zbog toga što granice svojih nekretnina u katastru nisu uskladili sa stanjem na terenu. Kako to u stvarnosti možemo očekivati vidimo na javnom web servisu geoportal.hr (slika 3).

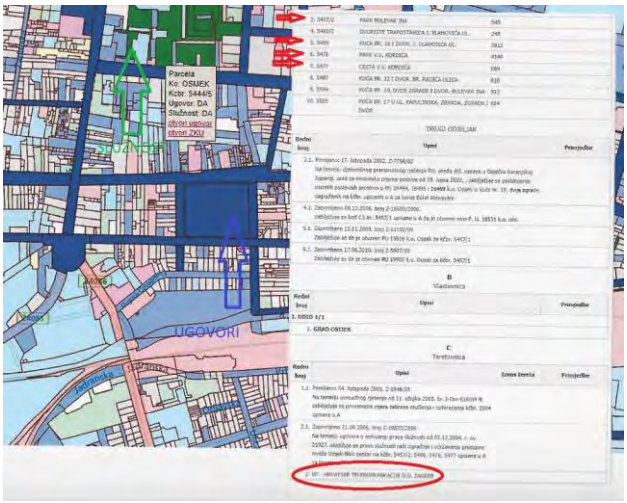


Slika 3. Neusklađenost digitalng katastarskog plana sa stanjem na terenu

VII. ORGANIZACIJA IMOVINSKOPRAVNOG UREĐIVANJA ELEKTRONIČKE KOMUNIKACIJSKE INFRASTRUKTURE I POVEZANE OPREME

Nabrojeni javni registi, zemljišna knjiga, katastar i katastar vodova imaju svoju zakonitost. Nihovom digitalizacijom se povećala mogućnost kontrole svakog vlasnika da vidi kako je njegova nekretnina smještena u prostor. Od 2011. godine postoji digitalni ortofoto (DOF) georeferencirana fotogrametrijska snimka, načinjena za potrebe legalizacija zgrada. Preko 830 tisuća zahtjeva za legalizacijom ukazuje na težnju da imaju legalno uređene nekretnine. Legalizacija je bila moguća bez obzira na čijoj se katastarskoj čestici zgrada nalazila. Što ćemo poduzeti danas kada na vidjelo izlaze razne uzurpacije susjedskog zemljišta (u vlasništvu fizičkih osoba ili općina, županija ili Republike Hrvatske). U ovakovom okruženju Hrvatski telekom provodi postupke imovinsko pravnog uređenja EKI. Posao je organiziran kao multidisciplinaran. Procesi se vode po internim procedurama. Velik dio poslova je organiziran kao outsourcing djelatnost. To su djelatnosti na katastru vodova postojeće mreže, izradi elaborata izvlaštenja, izradi elaborat za pravo puta i održavanju katastra vodova pri novoj izgradnji sve to obavlja dvadesetak geodetskih tvrki kontinuirano već nekoliko godina na cijelom području Hrvatske. Paralelno se istražuju podaci o upisanim služnostima nad vodovima koje je proveo prednik Hrvatskih telekomunikacija d.d., istražuju se svi analogni elaborati predani u područnim katastarskim uredima, organizira njihova digitalizacija. Provodi se inventura kabela uvućenih u distributivnu kabelsku kanalizaciju (DTK), dokumentiraju zauzeća cijevi i njihova dogradnja mikrocijevima sve u svrhu efikasnog iskorištenja kapaciteta. Poslije izrade elaborata za imovinsko pravno uređenje (elaborat za upis služnosti/pravo puta) slijedi niz susreta s vlasnicima nekretnina (lokalna zajednica, fizičke osobe, institucije, tvrtke…) kojima se omogućava kontrola elaborata i s njima se provode pregovori o svim elementima ugovora. Slijedi postupak uknjižbe svih ugovora za služnost i kontrola tog postupka. Za elaborate za koje se ishođuje potvrda o pravu puta, također je potrebno regulirati plaćanje naknade.

Upisana služnost vidljiva je na zemljišniknjižnom izvatku i tereti cijelu katastarsku česticu (slika 4). Podacima katastra i zemljišne knjige za sada nije moguće online pristupiti pa je neophodno svoj web servis ažurirati.



Slika 4. Upisana služnost na više katastarskih čestica u okviru istog zemljišnoknjižnog uloška

Vizualno praćenje potvrda o pravu puta preko ovog weba mora se drugačije organizirati. Preko trase kabela postavlja se bafer određene boje (npr. u postupku, dobivena potvrda).

VIII. OBNOVA ŠIROKOPOJASNE INFRASTRUKTURE

Imovinsko pravno uređenje EKI i inventurizacija koja to prati u funkciji je obveza iz Direktive 2014/61/EU Europskog parlamenta i Vijeća o mjerama za smanjenje troškova postavljanja elektroničkih komunikacijskih mreža velikih brzina [7]. Postavlja se dilema koja proizlazi iz tumačenja regulatora da se pravo puta kao oblik uređivanja odnosa između infrastrukturnog operatora i vlasnika nekretnine može koristiti jedino u postojećoj infrastrukturi. Ukoliko se mreža obnavlja zamjenom bakrenih kabela položenih u zemlju imovinskopravne odnose će trebati urediti principom

ustanovljavanja služnosti u toku ishođenja građevinske dozvole. Mreže velikih brzina moguće je realizirati na više načina. Prelazak na svjetlovodne prijenosne medije, napuštajući bakar kao medij prijenosa sigurno je budućnost. Kako je oko 73% kabela s bakrenim vodičima (kabeli položeni direktno u zemlju i zračni kabeli) nemoguće zamijeniti bez većih investicijskih zahvata postavlja se dilema kojim smjerom usmjeriti aktivnosti fokusiranih na budućnost tehnološkog napretka i potreba za telekomunikacijskim uslugama. Prisutna je stalna dilema kod planiranja budućnosti. Kome to treba? Imamo li mi dovoljno korisnika za usluge koje svjetlovodni medij omogućava. Povijest telekomunikacija nas uči da ne smijemo sumnjati u tehnološki razvoj. Utjecaj ove djelatnosti se odražava na bruto nacionalni dohodak. Ona povezuje ljude, štedi vrijeme i transport, doprinosi zaštiti okoliša. Sve te indirektne koristi koriste svi. Očekivati da operatori, koji se ovim poslom bave zbog profita snose sav teret investiranja u tehnologiju čija se posredna dobit disperzira na sve građane nije realno.

IX. ZAKLJUČAK

Praćenje poslovnih procesa koji se odvijaju paralelno, kod kojih je prijelaz na višu fazu uvjetovan okončanjem prethodne faze, uz puno međuovisnosti o održavanju sustava. Djelatnost koja je prostorno disperzirana i multidisciplinarno intonirana bez vizualizacije procesa i uvida u stanje predmeta i realiziranih zadataka i to 24 sata dnevno nije realno očekivati da će se brzo intervenirati na mjestima zastoja procesa.

Web pristup nudi rješenje. Moraju se uvažiti svi podaci koji utječu na sustav. U kontrolu procesa u fazi ustrojavanja baze moraju se uključiti razni specijalisti i svojim zapažanjima doprinijeti što bržem ustrojavanju sustava. Edukacijom svih zaposlenika fokusiranih na korisnike usluga i osiguranjem povratne veze prema vlasniku baze za očekivati je da će se produktivnost zaposlenika i zadovoljstvo korisnika povećati.

Nusprodukt ovakve baze može biti prikaz elektroničke komunikacijske infrastrukture za potrebe zaštite kabela od kidanja ili izbjegavanje polaganja EKI od strane drugih infrastrukturnih operatora u ulicama u kojima postoji kabelska kanalizacija HT.

U radu je opisana djelatnost fokusirana na imovinskopravno uređivanje elektroničke komunikacijske infrastrukture. Potrebno je napomenuti da u prostoru postoje i drugi vodovi (elektrovodovi, plinovodi, toplovodi, parovodi, naftovodi, vodovodi, kanalizacija) za koje važe slična pravila imovinskopravnog uređenja korištenja nekretnina u vlasništvu drugih. Duljina tih vodova je 4 puta veća od EKI.

Za pretpostaviti je da će uređivanjem svojih evidencija ostali vlasnici vodova postati davatelji informacija, bilo dislociranim dijelovima svoje tvrtke bilo drugim korisnicima pri čemu će koristiti infrastrukturu nekog od infrastrukturnih operatora.

Mogućnost uvida u položaj voda, koji je registriran u katastru vodova, preko web servisa omogućit će izbjegavanje kidanja radi neznanja i izbjeći značajne štete koje na taj način nastaju.

Za sada je mrežni pristup zamišljen kao vizualizacija svih prikupljenih informacija i dokumenata značajnih za održavanje elektroničke komunikacijske infrastrukture. Po istom principu moguće je sustav proširiti i podacima ostalih vodova. Na taj bi se način početna zamisao, vođenja procesa, proširio i na svojevrsnu inventuru izgrađene infrastrukture čime bismo ostvarili obvezu iz Direktive 2014/61 EU. Ujedno će pregled vlasništva nekretnina ukazivati na eventualne usporivače procesa za izgradnju nove svjetlovodne mreže.

LITERATURA

[1] Zakon o elektroničkim komunikacijama, Narodne novine, 73/08, 90/11, 133/12, 80/13 71/14

[2]Pravilnik o potvrdi i naknadi za pravo puta;Narodne novine, 31/09, 89/10 i 152/11, 151/14

[3] F.Ambroš, O.Antoni, P.Božičković: Javne evidencije katastra i gruntovnice kao temelj geodetskih elaborat izvlaštenja za elektroničku komunikacijsku infrastrukturu, VI simpozij ovlaštenih inženjera geodezije, Neiskorišteni potencijali geodezije – napuštena i nova područja djelovanja, rad u zborniku (181-191), Opatija 2013.

[4] www.dzs.hr

[5] Zakon o zaštiti osobnih podataka, Narodne novine 103/03, 118/06, 41/08, 130/11, 106/12

[6] Studija uspostavljanja nacionalnog integriranog geoiformacijskog sustava infrastrukture vodova, Državna geodetska uprava RH, Zagreb, 2015.

[7] Direktiva 2014/61/EU Europskog parlamenta i Vijeća o mjerama za smanjenje troškova postavljanja elektroničkih komunikacijskih mreža velikih brzina

Fieldbus Diagnostic Online Solution Program Establishment at Rijeka Oil Refinery

B. Žeželj * i H. Hajdo**
*Rafinerija nafte Rijeka, Automatsko vođenje procesa, Urinj 53, 51221 Kostrena
boris.zezelj@ina.hr
**Emerson Process Management, Process Systems and Solutions, Selska 93 10 000. Zagreb
haris.hajdo@emerson.com

Abstract - Rijeka refinery has built in year 2010. new hidrocracker complex unit as a part of refinery modernization project. This was time when complete new approach in term modern communication protocols was introduced in a field of instrumentation equipment communication. Protocol name is FOUNDATION Fieldbus which by very common definition represent an all-digital, serial, two-way communications system that serves as the base-level network in a plant or factory automation environment. It is an open architecture, developed and administered by the Fieldbus Foundation. Most beneficial cost advantage is reflected in the amount of cabling required which is much lower in Fieldbus than in 4-20 mA installations, several parameters can be communicated per device in a Fieldbus network and it provides a good foundation for the creation of a predictive and proactive maintenance strategy. First experiences show that plant-wiring concepts, shielding, training and the right tools are essential to gain the benefits of fieldbus technology. This paper describes online troubleshooting possibilities and diagnostic tools which can be used to address issues with devices before they become critical problems.

I. INTRODUCTION

Fieldbus is an industrial network system for real time distributed control systems management. Fieldbus is a generic term that describes the modern industrial digital communications networks that replace the standard 420mA analog signal. The network is digital, two-way, serial, communication network that connects remote components eg. field devices such as controllers, transducers, actuators and sensors. Each component has a simple computer possibility, which make it a "smart device“. Each device is able to independently perform simple functions such as diagnostics, control and maintenance, and provide two-way communication opportunities. With these devices not only operator able to access remote device, but they themselves are also able to communicate with other devices. Potentially fieldbus can replace centralized management with distributed control networks. From this side of the fieldbus is much more than a replacement for the 4-20 mA analog standards. Fieldbus technology provides improved quality, reduced price or an increased efficiency. These opportunities fieldbus technology resulting in part from the fact that information that the device accepts or transmits can be transmitted digitally. It removes the need for the DA conversion on the device and then the AD conversion on the side controller, which greatly increases the accuracy. Each device can also be a "smart" device and can take care of own control, maintenance and diagnostics. As a result we can get information if the device is defective or requires manual calibration. This increases system efficiency and reduces the scope of required maintenance. Fieldbus devices are more flexible than the older devices, as a result of the inclusion CPU. One fieldbus device can be used as a substitute for many devices that are used the old 4-20 mA analog standard.

Purpose of this paper is to show how it is carried out diagnostics of the entire foundation fieldbus infrastructure in Rijeka oil refinery at hydrocracker complex which was part of INA modernization project back to 2010. Reasons is very simple, to carried out and determine a possible non-compliance executed state installations in compared to internal recommendations Emerson defined in the "Fieldbus Installations in a DeltaV Digital Automation System" and "Site Preparation and Design for DeltaV Digital Automation Systems ".

II. TESTING PROCEDURE

Testing procedure was based on internal "Fieldbus Segment Checkout" protocol described in the "Fieldbus Installations in a DeltaV Digital Automation System "and" Site Preparation and Design for DeltaV Digital Automation Systems" documents using the following diagnostic tools

-Fluke 125 oscilloscope with "Fluke View" application used to transfer the oscillogram on PC, Fluke 287 digital multimeter and DeltaV diagnostics and Pepperl+Fuchs Mobile Advanced Diagnostic Module DM-AM-KIT.

The oscilloscope is used to record the waveforms of signals and measuring capacity, while the resistances and DC voltages of segments measured with Fluke 287 digital multimeter. When measuring the resistance between the plus and minus conductor side of fieldbus segments

certain variation of resistance was expected due to the charging and discharging of capacitors in the terminator and the because the measured value of the variable and not were expressed in absolute amounts. The main objective of this test is to determine that the resistance is not less than 50 kΩ, while among other conductors infinite resistance was expected, ie greater than 30 MΩ. The expected values of the capacity between the positive and negative conductors are about 1 µF. The values significantly below the 0.5 µF usually indicate a lack of terminator while values greater than 2 µF usually indicate the presence of an additional (third) terminator. There may be variations depending on the length of the cable and barrier presence, so that the specified parameters best be observed in combination with oscillograms.

The expected values of capacity among other conductors are less than 300 nF, where 500 nF represents the maximum with the correct oscillogram. Measuring the capacity of the trunk connector directly on the barrier or power conditioner should be 1 µF, and any significant deviation from this value indicates a potential failure or deactivated terminator. Before conducting measurements it is necessary to disconnect the connector segment with P + F MB-FB-4R module that goes to the plant, while the upper connector which is a module associated with DeltaV H1 card (marked with a "host") remains connected.

Measurement position in the P + F MB-FB-4R module is shown in Figure 1.

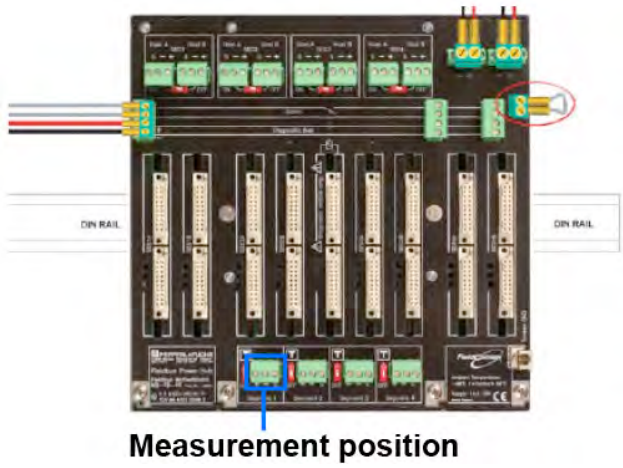


Figure 1. Segment connector P+F MB-FB-4R module

The above allows the isolation of the wiring to the field of H1 cards and DC power supply, isolate "shield" conductor of ground and allows measuring the resistance and capacity. Reference values of resistance and capacity are shown in the followed tables.

Table 1. Resistance reference values

Position	Expected result
between plus side and minus conductor side	>50 kΩ
between plus side and conductor shield	open circuit, >30 MΩ
between minus side and conductor shield	open circuit, >30 MΩ
between plus conductor side and bus grounding	open circuit, >30 MΩ
between minus conductor side and bus grounding	open circuit, >30 MΩ
between conductor shield and bus grounding	open circuit, >30 MΩ

Table 2.Capacity reference values

Position	Expected result
between plus side and minus conductor side	about 1µF, between 0,6 µF and 2 µF is acceptable
between plus side and conductor shield	<300 µF
between minus side and conductor shield	<300 µF
between plus conductor side and bus grounding	<300 µF
between minus conductor side and bus grounding	<300 µF
between conductor shield and bus grounding	<300 µF

After measurements of resistance and capacity it is necessary to determine the correct position of the internal switch for activation/deactivation of the terminator, which should be in active position and then reconnect cable to the segment connector and determine the value of DC voltage that for P + F module should be between 28 and 30 V. The final step involves recording AC oscillogram using the Fluke 125 oscilloscope and storing recorded values in "Fluke View" application. The following things should be taken into account, the oscilloscope settings set to 200 mV scale with a time base of 20 ms. In order to kept recorded signal on the oscilloscope screen "HOLD" or "SINGLE TRIGGER" functionality can be used. The first option retains the currently displayed signal on the oscilloscope screen, while the other allows individual triggering the oscilloscope in accordance with predefined settings (eg. 50 mV with falling or rising edge).

The equivalent circuit for generating fieldbus signal is shown in Figure 2. Power supply circuit is provided by using the DC rectifier in combination with a power conditioner device whose purpose is to ensure sufficient output impedance for the frequency range of fieldbus communication (31.25 kHz). Standard DC power adapter cannot be used for this type of communication because the low output impedance will cause short circuit for fieldbus signal.

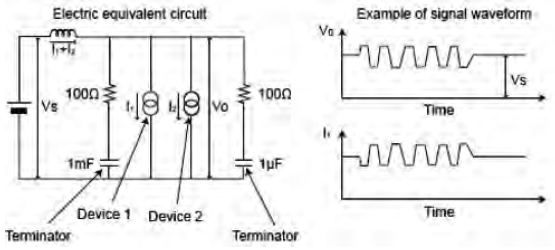


Figure 2. Electric Equivalent Circuit

At the ends of each fieldbus segment signal terminator device is installed which is composed of resistor (100Ω) and the capacitor (1 μF). Terminator is necessary to prevent signal reflection and provides a specific impedance of the connected instruments in the diagram shown as a current source.The currents I1 and I2 flow through the devices 1 and 2 (DeltaV H1 card, FF instrument). When the current of the first device increases by 10 mA, the potential difference of 0.5 Volts occurs between the conductors of the fieldbus trunk cable, since the total resistance of two resistors connected in parallel equals 50 ohms (10 mA x 50 ohm = 0.5 V). After this, the first device decreases its current by 20 mA which causes the voltage potential difference of 1 V. In former case this causes discharging and in the latter case charging of the capacitor. In this way, the current oscillation of +(-) 10 mA generates an AC signal of 1 V of peak to peak amplitude. The data is encoded by oscillating the voltage level in the center of the time interval required to generate a signal for one bit (32 microseconds at 31.25 kbps). The logical one is encoded as a voltage drop in the middle of 32 microsecond interval, while the logical zero is encoded as a voltage rise (Manchester encoding). The typical waveform generated by the described encoding scheme is shown in Figure 3. Every message begins with a preamble used for receiver synchronization and ends with a delimiter.

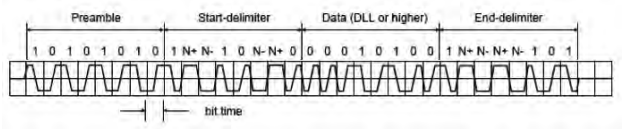


Figure 3. Fieldbus signal encoding

The above shows that the typical expected value of the amplitude of the AC signal around 1 volt AC, where more voltage levels indicate an increase in impedance of the circuit (eg, missing terminator), a lower to decrease. All values below 0.5 V and above 1 are considered to be critical and require additional diagnostics. It is important to point out that the control system even in such situations is capable for operating without any communication errors but significant deviations from normal value certainly require a more detailed diagnosis given that any mistake reduces overall system tolerance to failures. Since fieldbus signal is not grounded, ie. floating, system still works smoothly in case of short circuit one of the conductors to "shield" conductor since the differential voltage remains the same, but partial short circuit and another guide to "shield" the guide then leads to a complete standstill communication which in some cases may cause failure plant forcing removal failure in such cases it is necessary.

Since the works carried out during scheduled outages plant is expected that the presence of Forests in oscillograms be minimal which is why this type of measurement desired repeat during operation of the plant. Also, since the data part of a fieldbus frame can be up to 266 bytes of which the individual oscillogram shows only a few bits, there may be occasional errors that will not be recorded using an oscilloscope (eg. PVBAD alarm occurs several times a day). Therefore during operation of the plant was the best use diagnostic tool that provides continuous diagnostics and monitoring the fieldbus communication which would allow the detection and the kind of failures (eg. the occasional short circuit caused by the penetration of water into a measuring instrument). For this report will also include print port diagnostics of all DeltaV H1 card for a period of at least 24 hours. It is important to point out that System error number just in case completely illegible message of which cannot be detected partial degradation of the signal. All diagnostic counters will be reset before the start of the test and the alleged after at least 24 hours. DeltaV uses a series of diagnostic counters that can be found in port statistics and detailed port statistics which is part of continuous fieldbus diagnostic tool for control systems.

III. FAULTS AND RECOMMENDATIONS

The diagnostics measurements uncovered local ground loops within the DCS and marshaling cabinets. After disconnecting the trunk connector from the P+F MB-FB-4R module infinite resistance is expected between the shield conductors and the instrument ground bar. However, majority of the measurements indicated a closed loop due to previous grounding of the incoming trunk cable directly on the marshaling cabinet instrument ground bar, instead of the P+F module grounding connection. This should not cause any communication issues, but causes difficulties with the ground loop detection since it requires identification of the individual grounding points (not documented), disconnecting them, and then repeating the measurement. After disconnecting the instrument ground bars from the rest of the grounding system infinite resistance was determined between the Foundation Fieldbus trunk shields and the rest of the grounding system, while with the instrument ground bar connected the resistance was less than 0.5 Ohms which proved that the bus is properly grounded.

The majority of diagnosed problems manifested itself as signal reflections caused by inadequate impedance of the fieldbus segments usually due to missing termination or bad connections inside of junction boxes or individual instruments. This was determined from the increased scope meter waveform voltage peak to peak level combined with significant signal distortion. Since DeltaV Fieldbus implementation is extremely robust and the communication protocol itself is quite resilient to interference, none of the diagnosed faults has led to

complete loss of communication. However, it is still very important to fix all the diagnosed faults since any additional fault could cause total communication failure. In case of such an event, the affected control valves would go their fail safe state which could lead to a plant disturbance or even a trip.

IV. MEASUREMENT RESULTS

Example on Figure 4.shows increased amplitude of fieldbus signal waveform caused by a missing terminator. In this case it is necessary to check the status of the terminator dip switches on the P+F Fieldbus barriers. The terminator dip switch should be in the enabled position on the last barrier in the segment, as specified in the fieldbus segment layout documentation.

If the switch is in the expected position on both sides of the segment, it may be necessary to replace the barrier (in case of internal terminator fault).

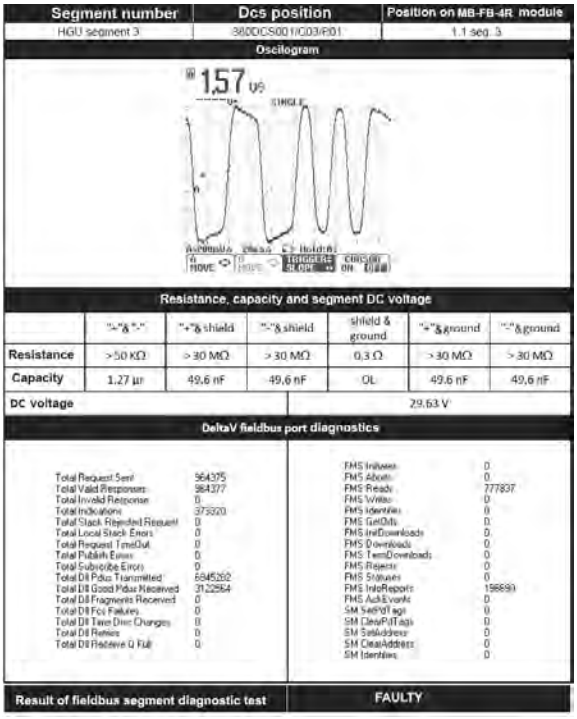


Figure 4. Diagnostic measurement results (increased amplitude of fieldbus signal)

Figure 5. illustrates an example of a short between the "+" conductor and the shield as shown by the ohm measurement reading of 9.5 ohms. This was further confirmed by the scope meter waveform where half of the signal amplitude is expected between the "+" and "-" conductors and the shield, while in this case the signal between the "+" and the shield was only slightly above 50 mV. One shorted conductor doesn't lead to a loss of communication since a differential signal between the "+" and "-" conductors is used, but the segment will be more sensitive to noise and additional short of the "-" conductor and the shield would lead to a complete loss of communication..

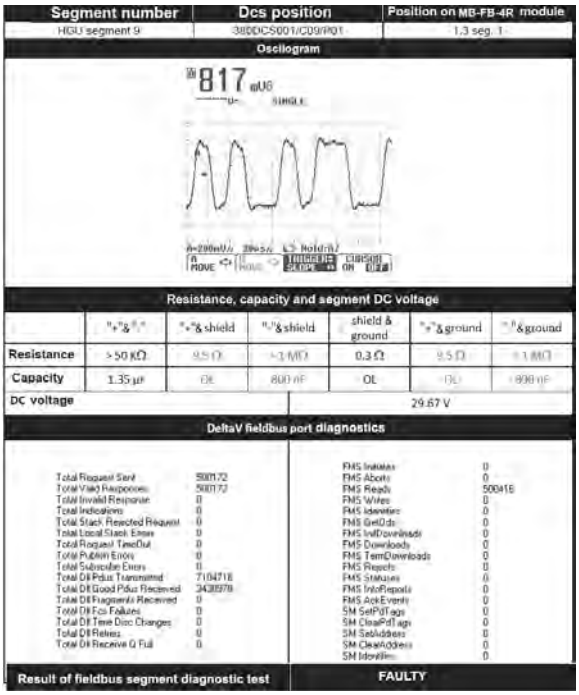


Figure 5. Diagnostic measurement results (short circuit between plus side bus conductor and shield)

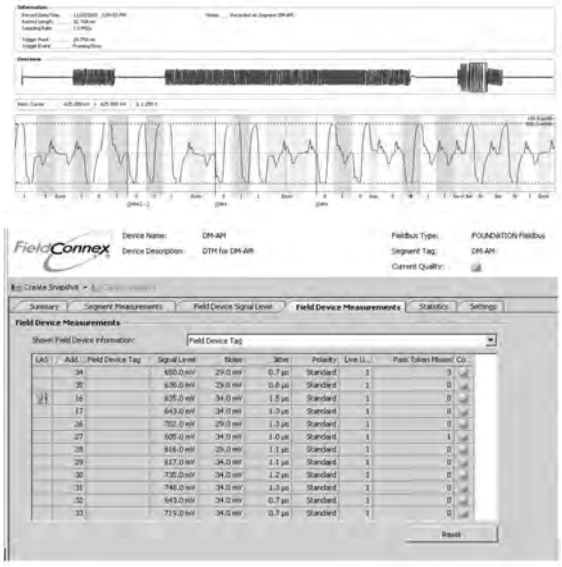


Figure 6. Diagnostic measurement results
(Pepperl+Fuchs Mobile Advanced Diagnostic Module DM-AM-KIT)

Figure 6. illustrates an example of a token passing timeout on specific foundation fieldbus segment belonging to address 34 which represents specific instrument. Previously it was determined a growing number of communication errors in the DeltaV diagnostics. An inspection of specified instrument was needed.

V. CONCLUSION

The described diagnostic measurements have certain limitations. The digital multimeter resistance and capacitance measurements on the Foundation Fieldbus trunk cables can only be carried out while the plant is in shutdown, with the fieldbus segments unpowered and disconnected. Furthermore, the length of a Foundation Fieldbus data frame can be up to 266 Bytes, while the scope meter can only display a few individual bits unless the time base is increased up to a point where it is difficult to assess the shape of the waveform. Because of this it is very unlikely that transient faults which may occur a few times per day could be detected by using these diagnostic methods. Another scope meter limitation is a lack of the ability to decode individual messages and display device addresses which can be highly useful when trying to narrow down the possible sources of the fault. Also, the likelihood of communication noise is much higher in case the plant is in operation which implies that a tool based on continuous, online monitoring of the Foundation Fieldbus segment communication parameters would be the superior diagnostic tool which could diagnose even these types of faults. The existing diagnostics module installed in the P+F motherboards are of the basic type and support detection only of critical errors (power loss, trunk shorts, etc.).

It is suggested to replace those existing modules with the "advanced" type which can monitor all the relevant diagnostic parameters as specified by the 61158 standard including jitter, signal unbalance, noise ratio, signal fall and rise times and can decode fieldbus messages and device addresses. Those modules can be combined with a gateway which allows all the diagnostics parameters to be integrated into the DeltaV DCS via the AMS Alert Adapter application which would allow detection of even subtle fieldbus signal degradation long before it is manifested in the DCS. This paper undoubtedly displays the first application of advanced communication diagnostic for instrumentation equipment in Croatian process industry.

REFERENCES

[1] 501-380 Fieldbus Physical Layer Troubleshooting Guide Rev B.1,Relcom Inc.

[2] Emerson Process Management: IZVJEŠTAJ O DIJAGNOSTICI FOUNDATION FIELDBUS INSTALACIJE HIDROKREKING KOMPLEKSA EPM-SERV-2014-1 HGU, SRU, HCU ,Revision F1–2014-Nov-11.

Automatic communication system ship to shipping terminal, for reporting potential malfunctions of a ballast water treatment system operation

Goran Bakalar, Ph.D.*, and Myriam Beatriz Baggini**
*International Maritime Technology Consultancy, President, Split, Croatia
**Bonum Mare Consultancy, Principal, Cocoa Beach, USA
Goran.Bakalar@xnet.hr

A**bstract - This research paper reviews ballast water treatment systems in two parts regarding data and methods: multi-criteria analysis and reliability study. The problem is how to timely recognize failure or operation stoppage of a ballast water treatment system. In the first part the multi-criteria analysis parameters and scenarios have considered the functional and operational problems of the systems. Aim of the study was to determine the most optional system from the maintenance point of view related to data transfer amount. The best score was 62% and it was the system that uses filtration, electrolysis and UV generator. Ballast water treatment systems could be in failure or under repair. That was confirmed and mathematically verified in the second part of the study. Reliability of the systems was calculated and total reliability value was 0.7 for each of the two computing systems individually, where one is redundant. New invented volumetric sensors should indirectly automatically report malfunctions through wireless and wire communications and that was suggested and ascertained in this paper.**

Key words: Ballast water; Multi-criteria analysis; Reliability; Sensor, Communication

I. INTRODUCTION

The problem analyzed in this article is unsuitable function of automated ballast water management systems on board the ships that could cause contamination of domestic waters. That contamination could be of harmful algae, bacteria or any other substances dangerous for humans or the environment. The problem is that the ability of modern ships to cover long distances in a short amount of time, provides a means for non-native marine species, including algae, invertebrates, and pathogens to arrive in ports via ballast water **[1]**.

The International Maritime Organization (IMO) commenced problem solution legally. At first, in 1991., it was launched *Guidelines for Preventing the Introduction of Unwanted Organisms and Pathogens from Ships' Ballast Waters and Sediment Discharges*. Then, finally, it was launched the *International Convention for the Control and Management of Ships' Ballast Water and Sediment*. The International Convention for the Control and Management of Ships' Ballast Water & Sediments was adopted by consensus at a Diplomatic Conference at IMO in London on Friday 13 February 2004**[2].** The Convention requires certified inspectors to do their best efforts to avoid unnecessary ship stoppage due to potentially required sampling. That means the advantage is usually given to the party of transportation instead of the sea protection. The inspectors could give an order to a designated institution or certified laboratory to take samples and to analyze ballast water that any crew intents to discharge into the sea. Laboratory analyses need to find specific microorganisms in the ballast water, and to determine salinity in order to find the origin of the ballast water. If the result of the laboratory analyses doesn't detect dangerous microorganisms or any other reason that would prevent ballast water discharge, all the expenses would be paid by the Ministry that the inspector represents. Those expenses are very high and could even cause serious problems with the cargo operations planning in the next port of call for the incriminated ship. There comes the need for cheaper and faster inspections. New technologies could help to find better ways for that. The experience has shown problems of the repairs that happen on the ships **[22]**. The ship's movement in the storms combined with the salt exposure brings up a very big risk of damages or improper functionality of any device or parts of the systems of ballast water treatment **[18]**.

Ballast water treatment systems could have clogged filters (UV ballast water management systems), burned UV bulbs, improper function of de-chlorination unit (chlorination ballast water management system) and many other possibilities **[22]**. The aim of this study is to determine a solution that will lead to avoid laboratory analysis and to provide environmental sustainability in the future related to this problem.

II. MATERIAL AND METHODS

MULTI CRITERIA ANALYSIS

The overall aim of this analysis was to arbitrate which ballast water treatment system had optimal performance from the maintenance point of view related to communication, operational data transfer, software sub-systems and ecological point of view. Multi-criteria analysis can be defined as a decision-making model that consists of a set of solutions (variants to rank or sort by

the decision-maker), a set of criteria (quantitative and qualitative, economic and ecological, using different measures) and the set value (score) of each variant for each criteria **[3]**. The PROMETHEE (*Preference Ranking Organization Method for Enrichment Evaluations)* method was designed to analyse criteria parameters including an alternative choice of specific criteria. The PROMETHEE II method is particularly well-fitted to this research as it can make decisions from incomparable criteria. PROMETHEE II has been used with success to solve many problems **[4]**.Compared to the well-established ELECTRE decision making method (also based on partial aggregation), the PROMETHEE II method is easier to use and able to get more robust results **[5][6]**.

Four different effective criteria with type approvals were chosen to be ranked using the PROMETHEE II multi-criteria decision making method:

1. Filtration + electrolysis + UV generator
2. Filtration + electrolysis + electro-chlorination + waste control
3. Electrolysis + oxidant
4. Deoxygenation

The importance of ecological protection is dealt with in three of the five criteria parameters. All criteria alternatives were analysed according to the following criteria parameters, as shown in Table 1.

TABLE I. PARAMETERS DATA **[7]**

BALLAST WATER TREATMENT SYSTEM	Environment pollution footprint 200/2000 m3/h	Efficacy	G8 or G9	Treatment in port or during the voyage	Communication data transfer (methods)
Filtration + electrolysis + UV	0.6 / 3	4	G9	port	2
Filtration + electrolysis + electrochlorination + waste control	8.7 / 12.4	4	G9	port	3
Electrolysis+ oxidant	4.5 / 11	5	G9	port	2
Deoxygenation	5 / 9	1	G8	sea	2

Environmental pollution parameter refers to the contaminants of natural environment that could cause adverse change. The ecological footprint used for this criterion parameter was ascertained by measuring the ballast water treatment system operation demand on the Earth's ecosystems, i.e. the amount of natural resource capital used per annum. Efficacy was a testing score parameter, i.e. data collected from the test performance results of ballast water treatment systems recorded via the United States Marine Invasive Species Program **[8]**. If a ballast water treatment system belongs to the G8 or G9 group depends on the test results. Tests are done during the certification process. If a certain ballast water technology uses an active substance, then, to comply with the Convention, it should be approved by the IMO in accordance with the *Procedure for approval of ballast water management systems that make use of Active Substances – G 9*, adopted by the MEPC (Marine Environment Protection Committee) in session 53 **[9][10]**. The number of treatment methods and whether treatment is done in the port or during the voyage is an important parameter from the maintenance point of view [21]. Number of methods is related to large amount of information if the system uses more methods for neutralisation. Neutralisation methods were in development from the year 2000 **[26] [27]** till today. There are more than 60 treatment systems listed in year 2014 **[7]** and some of them were invented at this year 2016 **[24]**. Some of the systems are using more methods and that means more complicated computational sub-systems. The failure index for those systems is higher than for the systems with less treatment methods. The analysis has shown the system that would be of best choice in timely manner for durance of cargo operations beside to important ecological point of view**.**

RELIABILITY AND AVAILABILITY OF THE SYSTEMS

Performance reliability of certain ballast water treatment systems on board ships have indicators which should be observed, followed and analysed over a longer period of system exploitation. Reliability is the ability of the system for operational work without interruption **[11]**. Predictions of potential failures caused by communication, software or hardware errors, as well as potential failures in mechanical part performance, can only be predicted once findings from the experience in handling technologically-similar systems on board ships are found. External factors on ships are one of the most important factors, because the dynamics of movements are constantly under external influences. Likewise, operator reliability and exposure to high temperatures are also very important in assessing the reliability of the proposed system **[12]**.

The reliability of these types of operation and communication systems can be shown as an exponential function of the time interval if the time interval is considered to be the useful lifespan of the device. For electronic systems such as sophisticated devices for the on-board ballast water treatment of ships and ship's communication systems , the failure density function has a form of an exponential distribution, so the failure frequency function is the same by definition **[13]**:

$$\lambda(t) = \frac{f(t)}{R(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda \tag{1}$$

In the equation (1.1), we can see that the failure frequency function has a constant value (λ). As such, the equation for the reliability function *R(t)* (exponential law of reliability) has the form of:

$$R(t) = e^{-\lambda t} \tag{2}$$

Mean time to failure (MTTF) is used to describe the reliability of elements, modules or communication systems, and it is the mean time of proper operation of a

system. It is determined as a mathematical expectation of proper performance duration. MTTF is expressed in hours, while the malfunction frequency λ is expressed by the number of malfunctions per hour or the number of malfunctions over a period of million hours of function (*Failures per Million Hours* - FMPH). Mean Time To Repair (MTTR) is a quantitative indicator for repairable systems. These systems include ballast water treatment systems and its communication on board the ships. MTTR is very difficult to assess in cases of failure caused by maintenance or ballast water treatment systems on board ships. On board ship experiences are unavailable, but MTTR can be determined by the formula:

$$MTTR = \frac{\sum_{i=1}^{M} t_i}{M} \tag{3}$$

where t_i is the time required for one individual repair, while M is the total number of repairs. Mean time between failures (MTBF) is a quantitative indicator of repairable systems and provides indispensable information regarding modern technical systems, as well as ballast water treatment systems on board ships and its communication system. Availability is defined in several ways. The one that is defined by the manufacturer (inherent, A_i) can be calculated using the formula:

$$A_i = \frac{MTBF}{MTBF + MTTR} \tag{4}$$

For ballast water treatment systems on board ships, MTBF can only be determined during the testing process when obtaining an operating permit or from on board ship experiences which are at present unavailable. Differences between MTTF and MTBF exist, which is in most cases numerically very small, but this difference is conceptually very significant. Since demand for a device which is the most reliable is always higher, for example 0.99 for 1 hour of operation, the mean time between failures amounts to 100 hours. As such, the reliability of the mechanics of the ballast water treatment systems on board ships which functions for 10 hours amounts to 0.951 or 95.1%. The reliability of certain ballast water treatment systems on board ships from a propriety aspect (where P_c is the proper function state) or the state of failure can be presented as:

$$R(t) = P_c(t) = e^{-\lambda t} \tag{5}$$

The computing and communication subsystem of some of the ballast water treatment systems on board ships can be found in three states: proper function, procedural failure and non-procedural failure **[14]**. Computing sub-system reliability depends on the probability of proper function and failures in a specified time. The probability of a computing subsystem of certain ballast water treatment systems on board ships going from a state of proper function to a state of non-procedural failure is:

$$P = (t+\Delta t) = \lambda \Delta t P_c (1-F) + (1-\mu\Delta t) P_{i.o}(t) \tag{6}$$

where:

$P_{i.o}$	probability of non-procedural state
P_c	probability of functioning state
μ	frequency of repairs
F	redundancy failure detection ratio
Δt	time lag
λ	malfunction index

The Malfunction index is the relation between malfunctioning components and operational components:

$$\lambda = \frac{1}{P_c} \frac{d P_f}{\Delta t} \tag{7}$$

where P_c is the number of components which remained operational after a specified time, and P_f is the number of components which malfunctioned after the specified time of functioning. Redundancy is a characteristic of the quality of the computing system that ensures failure avoidance when one part of the system fails. This is generally ensured through the use of additional software, through the reliability of the two redundant systems working in parallel and with a known malfunction index for these types of devices in predefined ship conditions. It can be assumed that the computing part of the subsystem in the system with two parallel subsystems, one of which is redundant, will be regularly maintained by the operator and through the self-diagnosis function for error removal. In this way, the possibility of malfunction of some ballast water treatment systems on board ships is reduced

III. Results And Solution

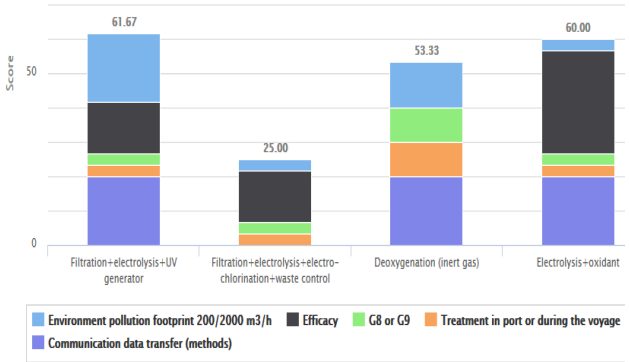


Figure 1. Multi criteria analysis results

The analysis has determined the most optional system for the scenario used in this research. The most weight was given to parameters of an ecological point of view, efficacy (in black color in Figure 1) and number of methods (in purple color). In green color is analysis result of the parameter which determines G8 or G9 group for the systems. Weighting related to G8 and G9 group and in purple color "number of methods" (amount of data to be transferred via communication system) parameter was

given in order to assign a system easier for the maintenance. Treatment in port or at sea (colored in light brown color in Figure 1) was related to maintenance point of view. The most optional system with best score in this analysis was the system that uses electrolysis and UV generator. According to the result, this system would have need less time for maintenance in port during the cargo and ballast water operation. Anyway this system would be controlled and automatically monitored to avoid potential by-passing of the system by crew members **[21] [22]**. The malfunction index or maintenance failure which cannot be diagnosed and automatically eliminated for these types of systems ranges from $\lambda = 0.5$ to $\lambda = 0.2$ **[15]**. If the mean is taken as being $\lambda = 0.35$ for each of the serially attached subsystems and this value is included in the equation of reliability for 1000 hours of work, we can obtain a reliability value for each of the two computing systems individually, where one is redundant:

$$R(t) = e^{-\lambda t} = e^{-0.35} = 0.703218 \quad (8)$$

The bad performance or operational shutdowns interfered with some of the parameters analysed in the multi-criteria analysis. A malfunction or a breakdown of ballast water treatment systems are serious problems that will occur more often in the future since more ships will have ballast water management systems installed in them **[7]**. The detection of treated ballast water quality would have to be implemented worldwide in purpose to control result of ballast water treatment **[16]**. The detection solutions were analyzed and compared in previous studies and flow cytometry was suggested as most optional detection method **[17]**. None of suggested solutions were timely indicators since detection results of these methods confirm ballast water quality after the treatment. It is needed to have an information immediately after some failure happen in a ballast water treatment system during the operation. One of the solutions could be volumetric flow monitoring system. There are many of flow monitoring options. One of them is Laser Doppler Anemometer (LDA). Differential pressure method with different diameter of pipes could be a solution as well. All those metric systems are sensitive electronic tools. Figure 2 shows possible location for installation of a flow metric unit.

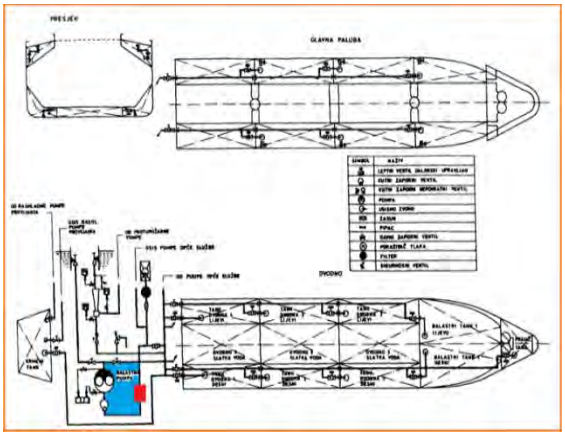


Figure 2. Blue color: monitoring stern unit. Red color: volumetric flow meter

Flow measurement is the quantification of fluid movement. The flow measurement data has to be available in shipping terminal control room immediately and permanently during the ballast water treatment system operation. If the system breaks down due any reason or the maintenance stoppage happen, the shipping personnel should have be alerted. Any stoppage should be inspected by shipping terminal duty operator. In that way it would be always possible to know whether the ballast water was treated or not. This is an indicative method for potential improper operation of a ballast water treatment system. On Figure 3 is a scheme of the automatic communication volumetric system.

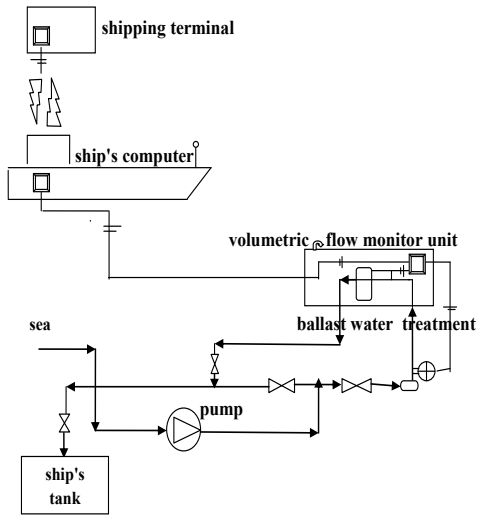


Figure 3. Scheme of automatic communication for volumetric flow meter data transmission

This is not a remote system. Remote systems were already suggested by some scientists **[19] [20] [25]**. Remotely operated automatic communication system would not be a communication option for this solution, since the communication has to be permanent and consistent.

IV. CONCLUSION

Multi-criteria analysis of ballast water treatment systems considered operational performance, communications and systems failures. According to the multi-criteria analysis results, the electrolysis + UV methodology treatment system had a success rate of 62%. The total calculated reliability of the entire computing and communication system was 0.7. That is a risk for ballast treatment system operation in ports and for the environment. If no good follow-up methodology is agreed upon, then the solutions stated in the BWMC (*Ballast Water Management Convention*) of 2004 cannot fully protect the marine environment. The purpose of the BWMC 2004 was to protect the marine and port environment and to diminish economic loss caused by

invasive microorganisms. Suggested automatic communication system ship to shipping terminal with flow measurement sub-system improves this purpose. The multi criteria analysis results from this study can be seen from a technical, environmental or ship owners' profit point of view**.** Functionality problems occurring while the ballast water treatment system was in operation should be considered from all these points of view. If the ballast water treatment systems do not function properly no profits will be made as active substances will be harmful for both humans and nature **[23]** and ultraviolet radiation will become dangerous for crewmembers **[28]** as will acids, chlorine, hydrogen peroxide and other substances from the G9 group. Crewmembers could by-pass the systems due to lack of the time for cargo operations because of their loyalty to their shipping companies. Suggested system of volumetric indicative method and its communication system from a ship to shipping terminal would prevent wrongful operation of ballast water treatment systems. Whenever the flow meter automatically measures suspicious flow amount through treatment system, the terminal operator has to inspect and record either a failure or maintenance of a ballast water treatment system. Mathematically confirmed reliability of software communication system was quite low. That means the system needs to be considered as a point of obliged control and monitoring. If the system does not operate well, or not at all, the price of treatment is higher, as is air and water pollution, and the active substances used, e.g. oxidant residue, become unmanageable and out of control.

REFERENCES

[1] http://hawaii.gov/dlnr/dar/ballast.html., Accessed: 8 February 2016.

[2] http://globallast.imo.org/the-bwmc-and-its-guidelines/, Accessed : 8 February 2016.

[3] S. Hajkowicz, and K. Collins, "A review of multiple criteria analysis for water resource planning and management,". in Water resources management, vol. 21(9), pp. 1553-1566, 2007.

[4] M. Behzadian, R. Kazemzadeh, A. Albadvi, and M. Aghdasi, " PROMETHEE: A comprehensive literature review on methodologies and applications," Eur. J. Oper. Res., vol. 200(1), pp. 198–215, 2009.

[5] T. Al-Shemmeria, B. Al-Klouba, and A. Pearman, "Model choice in multicriteria decision aid," Eur. J. Oper. Res., vol. 97(3), pp. 550–560, 1997.

[6] P. Thaillandier, and S. Stinckwich, "Using the PROMETHEE multi-criteria decision method to define new exploration strategies for rescue robots," IEEE International Symposium on Safety, Security, and Rescue Robotics, Proceedings, pp. 321-326, 2011. https://hal.archives-ouvertes.fr/hal-00688925v1., Accessed: 14 January 2016.

[7] Lloyd's, "Ballast water treatment technology - current status," http://www.lr.org/en/_images/213-35824_Understanding_Ballast_Water_Management_0314_tcm155-248816.pdf, 2014. Accessed: 8 February 2016.

[8] N. Dobroski, L. Takata, C. Scianni, and M. Falkner, "Assessment of the Efficacy, Availability and Environmental Impacts of Ballast Water Treatment Systems for Use in California Waters," California State Lands Commission Marine Facilities Division, 2007.

[9] MEPC: g/2012/Individual Guidelines for reference/G4.pdf, 2005. Accessed : 31 January 2016.

[10] E. Tsolaki, and E. Diamadopoulos, "Technologies for ballast water treatment: a review," J. Chem. Technol. Biot., vol. 85 (1), pp. 19-32, 2010.

[11] H. Pham, Software Reliability, Springer Google Books, 2000.

[12] D. R. Siewiorek, and R. Swarz, The Theory and Practice of Reliable System Design, Digital Press, Bedford, Maynard,.MA., USA, 1982.

[13] E. Turban, R. K. Rainer, and R. E. Poter, Introduction to Information Technology., John Wiley & Sons Inc., 2003.

[14] M. L. Shooman, Reliability of Computer Systems and Networks: Fault Tolerance, Analysis, and Design John Wiley & Sons Inc., 2002.

[15] G. Bakalar, "Efforts to develop a ballast water detecting device," Global IMO R&D Forum on Compliance Monitoring and Enforcement, Proceedings, Istanbul, Turkey, pp. 117-126., 2011. http://globallast.imo.org/wp-content/uploads/2015/01/RD_Turkey_2011.pdf, Accessed: 7 January 2016.

[16] G. Bakalar, "Automatic Control System for Ship Ballast Water Treatment by Using Flow Cytometry and Satellite Communications Technologies." Dissertation,. Rijeka University Croatia. 170 pp., 2013.

[17] G. Bakalar, "Review of interdisciplinary devices for detecting the quality of ship ballast water," SpringerPlus, vol. 1(3), pp. 468. doi: 10.1186/2193-1801-3-468., 2014.

[18] G. Bakalar, V. Tomas, and A. Buksa, "Monitoring of chemical pollution from the ships in coastal areas," International Conference on Climate Friendly Transport "Shaping Climate Friendly Transport in Europe, Key Findings & Future Directions", Belgrade, Serbia, 2011. https://bib.irb.hr/prikazi-rad?lang=en&rad=520608, Accessed: 7 February 2016.

[19] G. Bakalar, V. Tomas, and Z. Sesar, "Remote Monitoring of Ballast Water Treatment System Quality by Using Flow Cytometry and Satellite Communication Technologies," 54th International Symposium ELMAR 2012, Procceedings, pp. 290-295., 2012.

[20] G. Bakalar, and V.Tomas, "Possibility of using satellite communication technologies for remote maintenance in marine industry," 5th GNSS Vulnerabilities and Solutions Conference, 2011. https://bib.irb.hr/prikazi-rad?lang=EN&rad=520610, Accessed: 7 January 2016.

[21] G. Bakalar, and V. Tomas, "Possibility of Using Flow Cytometry in the Treated Ballast Water Quality Detection," Pomorski zbornik, vol. 51(1), pp. 43-55, 2016.

[22] G. Bakalar, "Comparisons of interdisciplinary ballast water treatment systems and operational experiences from ships," SpringerPlus, vol. 5(1), pp. 240. doi: 10.1186/s40064-016-1916-z., 2016.

[23] European Commission, Reflection Paper on Resudues in foodstuffs of animal origin, 2003.

[24] M. Cvetković, M. Grego, and V. Turk, "The efficiency of a new hydrodynamic cavitation pilot system on Artemia salina cysts and natural population of copepods and bacteria under controlled mesocosm conditions," Mar. Pollut. Bull., doi: 10. 1016/j.marpolbul.2016.01.030, 2016.

[25] G. Bakalar, "A thesis on a remotely controlled BWTS," The Ballast Water Times II, Journal of Institute NIOZ, project NSBWO (North Sea Ballast Water Opportunity)

[26] J. Aliotta, A. Rogerson, C.B. Campbell, and M. Yonge, Ballast Water Treatment by Electro-Ionization," 1st International Ballast Water Treatment R&D Symposium, Proceedings, pp 61-69, 2001.

[27] D. Mountfort, T. Dodgshun, and M. Taylor, "Ballast water treatment by Heat-New Zealand Laboratory $ Shipboard trials," 1st International Ballast Water Treatment R&D Symposium, Proceedings, pp 45-50, 2001. http://globallast.imo.org/wp-content/uploads/2014/11/Mono5.pdf. Accessed 15 March 2016.

[28] Case Western Reserve University, Environmental Health and Safety, Ultra Violet Radiation Safety., 2015. https://www.case.edu/ehs/Training/UV/UVsafety.pdf., Accessed: 17 January 2016.

AdriaHUB ICT platforma

Tatjana Škorjanc, Roberto Žigulić, Nikola Anđelić
Sveučilište u Rijeci Tehnički fakultet
tatjana.skorjanc@riteh.hr, zigulic@riteh.hr, nandjelic@riteh.hr

Sažetak – AdriaHUB platforma je transnacionalna ICT platforma za zapošljavanje razvijena u sklopu projekta "Premostiti tehničke razlike i društvene prepreke u smjeru pretvaranja jadranskog područja u središte održivog tehnološkog razvoja", skraćenog naziva AdriaHUB. AdriaHUB platforma je razvijena za jadransko područje, Hrvatsku, Srbiju, Crnu Goru i Bosnu i Hercegovinu pod vodstvom partnera projekta iz Italije, s ciljem poboljšanja integracije između sveučilišta i industrije i pronalaženja rješenja za najprikladniji plasman visokokvalificirane radne snage na tržištu rada.
Doprinos rada sastoji se u opisu napora uloženih u klasifikaciju prikupljenih informacija iz zemalja sudionica projekta, definiranje zajedničkih standarda i referentnih vrijednosti, kao i u pokretanje zajedničke platforme za zapošljavanje diplomanata na jadranskom području.

I. UVOD

Projekt "Premostiti tehničke razlike i društvene prepreke u smjeru pretvaranja jadranskog područja u središte održivog tehnološkog razvoja", skraćeno AdriaHUB, je projekt financiran iz sredstava IPA programa Jadranska prekogranična suradnja [1]. Instrument pretpristupne pomoći, skraćeno IPA, europski je program namijenjen pospješenju aktivnosti i napora koje država kandidatkinja čini u pretpristupnom razdoblju s ciljem članstva u Europskoj uniji (EU). U IPA programu Jadranska prekogranična suradnja sudjelovale su države članice EU - Italija, Grčka i Slovenija, Hrvatska, koja je u vrijeme početka programa bila kandidatkinja za ulazak u EU te države potencijalne kandidatkinje za ulazak u EU - Bosna i Hercegovina, Srbija, Albanija i Crna Gora. Samo određene regije u državama sudionicama - sve regije uz Jadran i neke druge, prikazano u [2], bile su prihvatljive za sudjelovanje u programu, pa tako i u AdriaHUB projektu. Naziv jadransko područje, koji se koristi u radu, odnosi se na te regije.

Vizija Europe kao društva zasnovanog na znanju, zasniva se na dostupnosti i primjeni znanja u svim segmentima života. Sveučilišta, kao ključni davatelji znanja i industrija kao njegov najveći potrošač su dva stupa ove vizije. Akademske institucije imaju dvostruku ulogu: istraživanja za stvaranje novih znanja i obrazovanja kao prenošenja postojećeg znanja.

Putem transfera tehnologije, sveučilišta na jadranskom području mogu djelovati kao nositelji istraživanja i tehnološkog razvoja, pomažući pritom poduzećima da prevladaju prepreke koje dolaze uz inovacije i konkurentnost. Međutim, niz država na jadranskom području uglavnom karakterizira velik broj malih poduzeća koja su rijetko u stanju iskoristiti prednosti koje im pružaju njihovi akademski istraživački partneri i najčešće nemaju dovoljne ekonomske i tehničke resurse za realizaciju zajedničkih projekata sa sveučilištima. Stoga uspostavljanje mehanizama za suradnju između sveučilišta i industrije može biti izvedeno korištenjem pretežno sveučilišnih resursa, uglednajući se na uspješne primjere iz prakse i pod vodstvom iskusnih vanjskih partnera. Dok suradnja u području istraživanja i razvoja progresivno raste, suradnja u području obrazovanja je tek na početku. Razlog tome se može naći i u činjenici da industrija nema jasan i potpun uvid u akademske institucije i procedure.

Nakon prethodnih iskustava akademsko poslovne suradnje koja je uspješno provedena u Italiji, Sveučilište u Bolonji iz Italije pokrenulo je projekt AdriaHUB s namjerom da razvije inovativnu platformu zasnovanu na informacijsko-komunikacijskim tehnologijama (engl. *ICT*), koja će biti u mogućnosti podržati širu suradnju i snažnu integraciju sveučilišta na jadranskom području i industrije. Odlučeno je da se u početku platformu pokrene i testira na drvnoj i drvoprerađivačkoj industriji u jadranskom području, s mogućnošću da se naknadno korištenje proširi na cjelokupnu industriju.

Sveučilište u Bolonji je, kao vodeći partner, ponudilo suradnju na projektu zainteresiranim partnerima u regijama prihvatljivim za sudjelovanje u IPA programu Jadranska prekogranična suradnja, sa razvijenom drvnom i drvoprerađivačkom industrijom u području.

Projektni konzorcij za realizaciju AdriaHUB projekta čine 23 partnera i suradnika (deset sveučilišta, pet javnih institucija i osam udruženja poduzeća) iz Hrvatske, Italije, Srbije, Crne Gore i Bosne i Hercegovine. Projekt je započet u listopadu 2012. godine. Trajanje projekta je u dva navrata produljeno, pa je predviđeni završetak u srpanju 2016. godine.

Partner projekta iz Hrvatske je Tehnički fakultet - Sveučilište u Rijeci, dok su pridruženi partneri Građevinski fakultet - Sveučilište u Zagrebu i Hrvatski zavod za zapošljavanje, Područni ured Dubrovnik. Popis svih partnera i pridruženih partnera i više o AdriaHUB projektu može se naći u [3].

U sklopu AdriaHUB projekta predloženi su mehanizmi podrške šire suradnje sveučilišta i industrije kroz:

- stvaranje međunarodne mreže visokoškolskih ustanova i privrednih subjekata za podršku sveučilištima sa ciljem usklađivanja obrazovanja i istraživanja s potrebama privrede,

Ovaj rad je napisan kao dio aktivnosti unutar projekta "AdriaHUB – Premostiti tehničke razlike i društvene prepreke u smjeru pretvaranja jadranskog područja u središte održivog tehnološkog razvoja". Ovaj projekt sufinanciran je od Europske unije te od IPA programa Jadranska prekogranična suradnja.

- pokretanje platforme za zapošljavanje visokokvalificiranih kadrova i povezivanje s bazom podataka konzorcija AlmaLaurea u Italiji,
- stvaranje baze podataka znanstveno-istraživačkih usluga koje nude sveučilišta u jadranskom području,
- poticanje primjene znanja za inovacije primjenjive u privredi, posebno u drvnoj industriji.

U ovom radu posebno će biti opisan doprinos Tehničkog fakultet u Rijeci, koji sudjeluje u projektu pod vodstvom prof. dr. sc. Roberta Žigulića, posebno u dijelu koji se odnosi na razvoj i pokretanje ICT platforme za zapošljavanje. Osoblje projekta, ukupno osam osoba, uključujući autore rada, radi na Zavodu za tehničku mehaniku, u Računalnom centru i Službi studentske evidencije. Na projektu su zaposlena dva mlada istraživača. U projektu sudjeluju profesori sa Zavoda za računarstvo i studenti studija računarstva.

II. ADRIAHUB ICT PLATFORMA

A. AlmaLaurea

Razvoj AdriaHUB ICT platforme za zapošljavanje. krenuo je od pozitivnog iskustva talijanske ICT platforme AlmaLaurea [4], u vlasništvu istoimenog međusveučilišnog neprofitnog konzorcija, partnera projekta, osnovanoga 1994. godine. AlmaLaurea uključuje 73 talijanska sveučilišta i približno 91% talijanskih diplomanata (podaci iz listopada 2015. godine).

Konzorcij potpomažu sveučilišta koja u njemu sudjeluju, talijansko Ministarstvo obrazovanja, sveučilišta i istraživanja (MIUR) i sva poduzeća i ustanove koje koriste bazu podataka i usluge koje nudi AlmaLaurea.

Osnovne aktivnosti AlmaLauree su :

- prikupljanje profila diplomanata – jednom godišnje, neposredno prije nego što diplomiraju, studenti popunjavaju detaljan upitnik sa podacima o školovanju,
- prikupljanje informacija o zaposlenju diplomiranih studenata - diplomirani studenti početno unesene podatke periodički ažuriraju,
- online baza životopisa diplomiranih studenata – alat koji olakšava usklađivanje ponude i potražnje radne snage na tržištu rada.

U AlmaLaurea platformi se odvijaju slijedeći procesi :

- studenti se registriraju na platformu, ispunjavaju upitnike, ispunjavaju životopise,
- osoblje sveučilišta šalje administrativne podatke u bazu podataka
- poduzeća se registriraju, akreditiraju, pretražuju bazu podataka diplomanata, nude zaposlenja, analiziraju prijave na zaposlenja
- osoblje AlmaLauree objedinjuje korisničke i administrativne podatke i pruža pomoć korisnicima

AlmaLaurea je referentna točka za sve koji se na raznim razinama susreću s tematikom studiranja i zapošljavanja diplomiranih studenata u Italiji, vrijedan izvor podataka pri analizi kvalitete sveučilišta i studija, zapošljivosti studenata s pojedinih studija i regija, visini očekivane plaće pojedinih profila stručnjaka, duljini čekanja na posao i slično.

B. Razvoj AdriaHUB ICT platforme

Pod vodstvom partnera projekta AlmaLauree i na iskustvima njihove ICT platforme, u projektu AdriaHUB je razvijena i pokrenuta AdriaHUB web orijentirana platforma za zapošljavanje za jadransko područje [5].

S obzirom na heterogenost jadranskog područja i različitosti obrazovnih sistema zemalja sudionica u projektu, pristupilo se prikupljanju relevantnih podataka, njihovoj klasifikaciji i izradi zajedničkih kataloga (engl. *decode tables*). Uloga kataloga je osigurati unaprijed definiran i ograničen skup izbora pri popunjavanju podataka od strane studenta kao i pri slanju podataka od strane osoblja visokoškolske ustanove.

Predložena struktura kataloga je na temelju prikupljenih primjedbi partnera projekta dodatno oblikovana, usuglašena i standardizirana. Definirani su AdriaHUB kodovi za podatke. Svaka ustanova partner je popunila kataloge s relevantnim podacima na službenom jeziku svoje države i na engleskom jeziku. Definirana su dva tipa kataloga: geografski i obrazovni.

Geografski katalozi su definirani na temelju prikupljenih podataka o naseljima i upravno-administrativnim jedinicama i sadrže jedinstvene nazive, lokalne i AdriaHUB kodove.

Tehnički fakultet u Rijeci je definirao i popunio hrvatski katalog sa hrvatskim i engleskim nazivima županija, odgovarajućim nacionalnim statističkim kodom, ASCII međunarodnim kodom županija, glavnim gradovima županija, imenima svih hrvatskih naselja, gradova i općina, poštanskim brojevima i kodovima.

Sve visokoškolske ustanove partneri na projektu su, sukladno specifičnostima obrazovnog sistema u svojim državama, uključene u Bolonjski proces, reformu u visokom obrazovanju europskih zemalja temeljenu na europskom sustavu prijenosa bodova, skraćeno ECTS (engl. *European Credit Transfer System*). Primjenom ovog sustava nastoji se osigurati usporedivost programa, pokretljivost studenata unutar i među sveučilištima. AdriaHUB projekt, posebno dio koji se odnosi na definiranje obrazovnih kataloga, u skladu je s tim nastojanjima.

Obrazovni katalozi sadrže jednoznačno definirane podatke, opise i kodove vezano uz sveučilišta, fakultete koji učestvuju u projektu, akademske i stručne nazive, njihovo stupnjevanje i popis srednjih škola u državama učesnicama. Klasificirane su akademske discipline i profili stupnjeva školovanja. Da bi se usporedili različiti sistemi ocjenjivanja i naziva uvedene su referentne vrijednosti.

Tehnički fakultet u Rijeci je izradio hrvatski katalog koji sadrži podatke o hrvatskim sveučilištima, fakultetima Sveučilišta u Rijeci, nazive studija na Tehničkom fakulte-

tu u Rijeci, odgovarajuće ETCS bodove, smjerove, ukupno trajanja studija u satima nastave i godinama, minimalnu i maksimalnu numeričku ocjenu i sve odgovarajuće kodove.

AdriaHUB web platforma izrađena je u skladu s važećim zakonima u državama iz kojih su partneri. Svaka akademska ustanova je pripremila tekst izjave kojom studenti pri registraciji daju suglasnost o korištenju svojih osobnih podataka u skladu s zakonom.

Tehnički fakultet u Rijeci je pripremio i koristio izjavu kojom studenti daju suglasnost da se njihovi osobni podaci koriste u skladu s Zakonom o zaštiti osobnih podataka u Hrvatskoj (»Narodne novine«, br. 106/12, pročišćeni tekst) za izradu baze podataka u okviru AdriaHUB projekta.

Osim izrade kataloga, velik je trud uložen od strane svih akademskih ustanova učesnica, pa i od Tehničkog fakulteta u Rijeci, da bi se definirali podaci koji se traže u životopisu studenta, kao i administrativni podaci studenata koje šalje ustanova, uzimajući u obzir sve različitosti obrazovnih sistema u zemljama i u ustanovama.

Talijanski partner AlmaLaurea je izradio modul koji je ujedinio prikupljene podatke iz akademskih partnera, kao i samu web aplikaciju za AdriaHUB platformu.

U svim akademskim ustanovama partnerima je napravljeno prevođenje tekstova na web stranicama na službeni jezik države. Tehnički fakultet u Rijeci je napravio prijevod na hrvatski jezik. Ustanove iz Srbije, Bosne i Hercegovine i Crne gore su, osim latiničnog pisma, tekstove upisale i na ćiriličnom pismu.

U prvom koraku je testiran i pusten u rad studentski dio platforme. Na Tehničkom fakultetu, kao i na drugim visokoškolskim ustanovama, AlmaLaurea je napravila online obuku AdriaHUB osoblja na fakultetu za korištenje sustava. Provedeni su testovi sa podacima stvarnih studenata koji su se prijavili na testnu platformu nakon informiranja i upućenog im poziva. Rezultati izvršenog testiranja su dokumentirani, te su napravljene potrebne korekcije. Nakon uspješno provedenog testiranja, studentski dio AdriaHub platforme je pušten u rad.

U drugom koraku cijeli postupak je ponovljen za dio platforme namijenjen poduzećima.

C. Opis AdriaHUB web platforme

AdriaHUB web platforma (slika 1.) se u korisničkom smislu sastoji od dijela namijenjenog studentima, dijela namijenjenog poduzećima, te dijela za AdriaHUB osoblje.

U dijelu namijenjenom studentima AdriaHUB web platforma nudi diplomantima visokih učilišta uslugu izrade životopisa koji su vidljivi i dostupni poduzećima radi mogućeg zapošljavanja.

Svakom diplomantu nude se sljedeće besplatne usluge:

- dodjeljivanje korisničkog računa i lozinke koji vrijede bez vremenskog ograničenja;
- dostupnost i rasprostranjenost životopisa na internetu;
- trajna mogućnost ažuriranja životopisa;
- pristup slobodnim radnim mjestima objavljenim na internetu od strane poduzeća;
- mogućnost direktnog odgovora na poslovne ponude na web-stranici slanjem životopisa poduzećima;
- pristup online informacijskim uslugama (bilteni, dokumentacija, integracija podataka i drugo);
- pristup statističkim podacima o uspjehu studiranja diplomanata.

Izrada životopisa je omogućena diplomantima visokih učilišta koja su sudjelovala u projektu. U sklopu životopisa uz osobne informacije (ime, prezime, mjesto rođenja i drugo), navedeno je obrazovanje osobe (sveučilište, fakultet, razina visokog obrazovanja, ocjena studiranja,



Slika 1.
AdriaHUB web platforma

datum diplome, naslov, tema i sažetak završnog rada, ključne riječi, voditelj završnog rada, i drugo). U životopisu se uz dosadašnje radno iskustvo navodi i vrsta željenog posla, uključen je opis jezičnih vještina i tehničkih sposobnosti kao i pregled svih dosadašnjih relevantnih aktivnosti, publikacija i zasluga. Moguće je i priložiti link na web stranicu s dodatnim podacima po volji diplomanta.

Diplomanti, koji su završili preddiplomski ili diplomski studij, pristupaju web stranici u sklopu koje unose svoje osobne podatke, ispunjavaju upitnik, kreiraju životopis i daju „Izjavu o suglasnosti s prikupljanjem i obradom osobnih podataka".

Nakon registracije na AdriaHUB web platformu podaci o diplomantima (uspjeh pri studiranju, životopis, itd.) postaju dio baze podataka.

U dijelu namijenjenom poduzećima AdriaHUB web platforma nudi alat za traženje radne snage. Nakon registracije poduzeća i dobivenog odobrenja od strane AdriaHUB osoblja, voditelji odjela ljudskih potencijala u poduzećima mogu preko interneta pregledavati životopise diplomanata koji su se uključili u projekt. Životopisi se mogu odabrati i spremiti koristeći napredno pretraživanje i definirajući ciljne karakteristike profila budućeg zaposlenika. Slika 2. prikazuje rezultat pretraživanja sa unesenim kriterijima: državljanstvo - Crna Gora; godine starosti - manje od ili jednako 30; stupanj obrazovanja - druga razina visokog obrazovanja.

Osim pretraživanja životopisa, poduzeća mogu i nuditi poslove. U ponudi zaposlenja definiraju se podaci o poduzeću, opis posla i detalji o traženom profilu posla. Unesu se podaci o osobi za kontakt kojoj sustav automatski šalje prijave diplomanata. Da bi ponuda zaposlenja bila vidljiva studentima, mora biti odobrena od strane AdriaHUB osoblja.

Registrirana poduzeća mogu pretraživati životopise i dobivati prijave na ponuđene poslove od studenata koji su se registrirali na sustav iz Hrvatske, Srbije, Crne Gore i Bosne i Hercegovine. Isto tako registrirani studenti poslovne ponude mogu dobiti od registriranih poduzeća iz ranije navedenih država.

U početnoj fazi rada platforme, registrirana su samo poduzeća iz drvne i drvoprerađivačke industrije, međutim ubrzo nakon toga, mogućnost registracije je, bez ograničenja, omogućena svim zainteresiranim poduzećima.

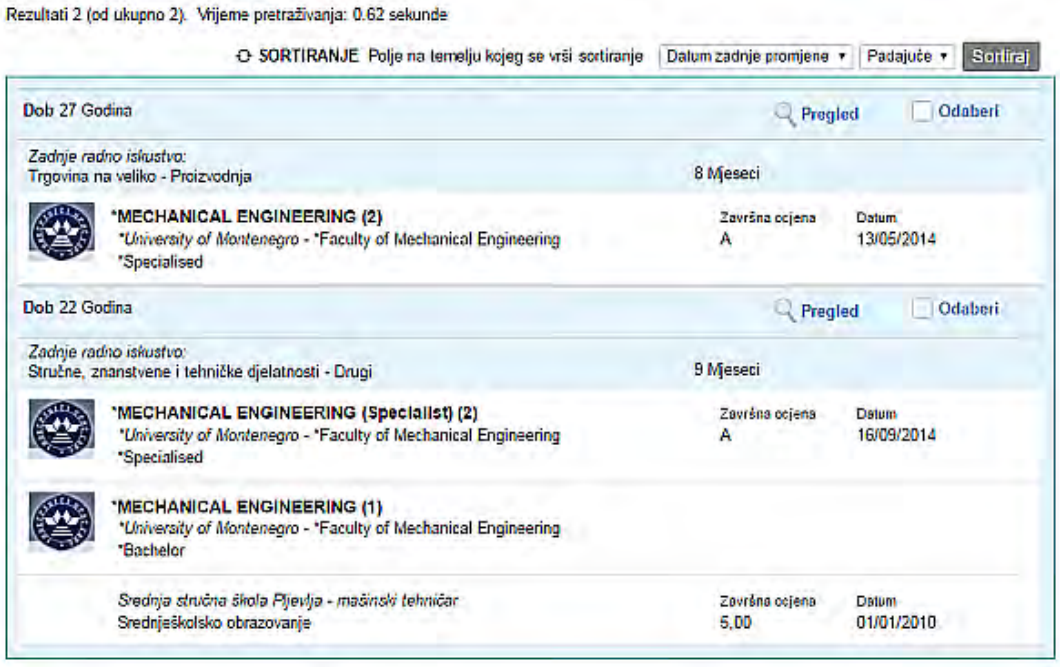
D. Razvoj ETL modula

Svaka akademska ustanova je s obzirom na svoje specifičnosti napravila vlastiti ETL modul (*data extract, transform, load*), aplikaciju za izdvajanja podataka diplomanata iz postojeće baze podataka, njihovo certificiranje i prijenos certificiranih podataka u AdriaHUB web platformu.

Za razliku od drugih portala za zapošljavanje, na kojima su podaci osoba koje traže zaposlenje neprovjereni, na AdriaHUB platformi certificiranje daje dodatnu vrijednost podacima, jer poduzeća mogu biti sigurna u točnost podataka vezanih uz studiranje.

Na Tehničkom fakultetu u Rijeci, u suradnji sa studentima studija računarstva i profesorima sa Zavoda za računarstvo, izrađena je aplikacija „Alat za automatsku certifikaciju studenata" za certificiranje studenata koji su završili studij na Tehničkom fakultetu u Rijeci i koji su se prijavili za sudjelovanje na projektu AdriaHUB.

Certificira se prije svega podatke vezane uz završeni studij na Tehničkom fakultetu u Rijeci, a osim toga, u aplikaciji se provjerava i podatke kao što su ime, prezime, datum i mjesto rođenja, JMBAG (jedinstveni matični broj akademskog građana) i drugi.



Slika 2.
AdriaHUB web platforma – rezultat pretraživanja životopisa

Na Tehničkom fakultetu u Rijeci, kao i u većini hrvatskih visokoškolskih ustanova, podaci o studentima , njihovim ocjenama i tijeku studiranja unose se u aplikaciju ISVU (Informacijski Sustav Visokih Učilišta). ISVU je sustav koji omogućava svim visokim učilištima Republici Hrvatskoj informatizaciju većine administrativnih poslova vezanih uz proces studiranja i mogućnost analize toga procesa s ciljem donošenja odluka utemeljenih na tim podacima.

Pravilnikom o studijima Sveučilište u Rijeci je propisalo da se studenti, osim brojčanom ocjenom: 5, 4, 3, 2 i 1, ocjenjuju i slovnom ocjenom iz ETCS sustava ocjena: A, B, C, D, E, F i FX te ETCS skalom ocjenjivanja u postocima od 0 do 100% ocjenskih bodova. ECTS slovne ocjene i ocjenski bodovi se ne mogu evidentirati u ISVU sustavu, pa se na Tehničkom fakultetu u Rijeci, osim ISVU sustava, koristi i vlastita aplikacija ISVUpatch preko koje se evidentiraju slovne ocjene, ocjenski bodovi, završni ili diplomski rad i suplement diplome, odnosno kompletan završetak studija.

Za certificiranje se koriste većinom podaci u ISVU patch-u, ali se neki podaci, koji u njemu ne postoje, kao naprimjer JMBG (Jedinstveni matični broj građana) moraju preuzeti iz ISVU-a, što dodatno otežava certificiranje.

„Alat za automatsku certifikaciju studenata“ je napravljen kao web aplikacija pomoću standardnih tehnologija PHP, HTML, AJAX, CSS, JavaScript i Jquery.

Redoslijed postupaka koji se provode nad podacima je slijedeći:

- Podaci se koje su diplomanti nakon registracije unijeli u AdriaHUB web platformi dodaju se u bazu podataka aplikacije za certificiranje, nazvanu RitehBaza, putem uvoza .cvs tekstualnih datoteka ili putem sučelja za dodavanje podataka. Tako dodani podaci mogu se uređivati i brisati
- Pripremljeni podaci uspoređuju se s postojećom ISVUpatch bazom studenata na Tehničkom fakultetu u Rijeci
- Provjeravaju se ime, prezime, JMBAG, datum i mjesto rođenja studenta. Na slici 3. je prikazan primjer provjere zapisa. Podaci u postojećoj ISVUpatch bazi vezani uz adrese i studije povezuju se s AdriaHUB katalozima.
- Podatak je certificiran ako zadovolji sve provjere. Certificirane podatke može se pojedinačno ili grupno pripremiti za slanje u AdriaHUB portal.. Podaci se izvoze u .xml formatu, prema unaprijed definiranoj .xsd shemi.
- Necertificirani podaci mogu se uređivati i brisati.
- Certificirani i izvezeni podaci se arhiviraju.

E. Pokretanje AdriaHUB web portala i službe za korisnike na ustanovama

Na svim viskoškolskim ustanovama, članicama projekta, napravljen je AdriaHUB web portal ustanove sa ciljem informiranja kolega na fakultetu, studenata i diplomanata te poduzeća.

Tehnički fakultet u Rijeci na svom AdriaHUB web portalu [6] redovito objavljuje informacije o AdriaHub projektu, posebno vijesti o projektnim aktivnostima na ustanovi kao i najave događanja. Portal je izvor informacija za studente i za poduzeća. Objavljena su česta pitanja i odgovori i pokrenut je „helpdesk“ - služba za korisnike.

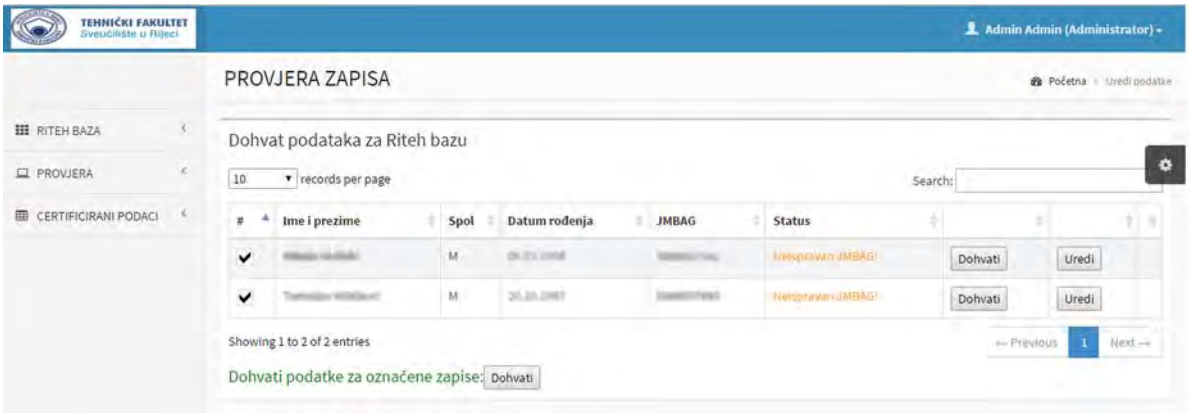
F. Baze podataka znanstvenoistraživačkih usluga sveučilišta

Stvorene su nove interne baza podataka znanstvenoistraživačkih usluga koje nude sveučilišta u jadranskom podruju. U kontaktu s privredom ispitane su mogućnosti za primjenu tog znanja za inovacije posebno u drvnoj industriji.

Tehnički fakultet u Rijeci je kontaktirao i upoznao s projektom devet poduzeća iz drvne i drvoprerađivačke industrije iz Primorsko-goranske županije i poslao im upitnik koji je korišten za preciznije definiranje tehnoloških zahtjeva i potreba za tehničko-stručnim profilima kompanija iz drvnoprerađivačkog sektora.

Prikupljeni podaci su analizirani te je zainteresiranim kompanijama poslan "Poziv za dostavu ideja". Na poziv su se javila poduzeća Lokve d.o.o., Lokve, Ferenčić d.o.o., Viškovo i Pilana Potok Čabar d.o.o., Čabar.

Partneri iz industrije pozvani su na radionicu "Dynamics and Control” održanu u studenom 2014. godine na



Slika 3.
Aplikacija „Alat za automatsku certifikaciju studenata“

Tehničkom fakultetu u Rijeci, gdje su predstavili svoja poduzeća. Održano je sedam predavanja tematski vezanih za drvo kao materijal te drvnu i drvoprerađivačku industriju. Na radionici su predstavljena četiri diplomska rada i održane tri laboratorijske vježbe, sve vezano uz drvoprerađivačku industriju. Održana je rasprava o stanju drvoprerađivačke industrije Primorsko-goranske županije.

G. Baza pilot projekata

Na temelju iskazanih potreba industrije, svaki partner na projektu mogao je prijaviti više pilot projekata i prijaviti se na predložene pilot projekte drugih partnera. Pilot projekti su ograničeni na drvnu i drvnoprerađivačku industriju U sklopu AdriaHUB projekta stvorena je baza od 14 pilot projekata na kojima su sa svojim profesorima, radili i mladi istraživači.

Tehničkom fakultetu Sveučilišta u Rijeci, prihvaćeni su pilot projekti koje je vodio u prema iskazanoj potrebi i u suradnji s poduzećima Pilana Potok Čabar d.o.o. Čabar i Ferenčić d.o.o. Viškovo i u suradnji s SCM grupom, partnerom iz Italije:

- "Istraživanje utjecaja materijala, geometrije, brzine vrtnje i topline koja se generira rezanjem drva na vibracije lista kružne pile. Istraživanje veze između strukturnih vibracija i buke. "
- "Istraživanje dinamike potporne strukture CNC stroja koja utječe na lošu površinsku obradu dr va. "

Tehnički fakultet u Rijeci je sudjelovao i u pilot projektu predloženom od strane partnera projekta Fakulteta za mašinstvo i građevinarstvo u Kraljevu, Univerziteta u Kragujevcu:

- "Istraživanje i primjena regulacijskog algoritma za pneumatsku paralelnu robotsku platformu za strojnu obradu drva."

Provedena istraživanja i ispitivanja na projektima rezultirala su izvještajima kojima se analiziralo postojeće stanje, te su dane preporuke za poboljšanja. Sudionici pilot projekata, profesori i mladi istraživači, su objavili slijedeće radove radove na 8. Međunarodnom kongresu Hrvatskog društva za mehaniku održanom 2015. godine u Opatiji: "*Variation of natural frequencies by circular saw blade rotation*", "*Effect of design parameters in circular saw blades*", "*A nonlinear vibrational model of the woodworking bandsaw*".

U publikaciji [7], zajedno s radovima drugih projektnih partnera, objavljeni su radovi: "Stanje drvne i drvoprerađivačke industrije Primorsko-goranske županije", "Buka praznog hoda kružne pile", "Vibracije lista kružne pile", "Vibracije konstrukcije stroja za obradu drveta i njihov utjecaj na kvalitetu obrade" i "Frekvencijski sadržaj vibracija tračnih pila pri pojavi fenomena *washboardinga*".

Otvoreno je ukupno oko 15 novih radnih mjesta na ustanovama partnerima u jadranskom području, vezano uz projekt. Od toga su na Tehničkom fakultetu Sveučilišta u Rijeci zaposlena dva mlada istraživača na poslovima istraživanja s ciljem primjene znanja na rješavanju problema i potreba u drvnoj industriji. Oni su, zajedno s mladim istraživačima iz drugih ustanova, partnera projekta, sudjelovali na osposobljavanjima za daljnja istraživanja i stručnim osposobljavanjima. U okviru rada na pilot pro-jektima obučeni su za znanstveni rad, sudjelovali su na međunarodnim kongresima i bili u prilici objaviti svoje radove u poznatim znanstvenim časopisima.

III. ZAKLJUČAK I BUDUĆI RAD

AdriaHUB ICT platforma je transnacionalna web orijentirana platforma za zapošljavanje koja uspostavlja veze između sveučilišta i industrije u jadranskom području. Na taj način se s jedne strane omogućava poslodavcima lakše pronalaženje viskokvalificirane radne snage, a s druge strane se diplomantima olakšava pristup tržištu rada.

Poduzeća iz Hrvatske, Srbije, Crne Gore i Bosne i Hercegovine, koja se registriraju u sustav mogu pretraživati životopise studenata, nuditi poslove i dobivati prijave na ponuđene poslove od studenata koji su se registrirali na sustav iz ranije navedenih država. Posebna vrijednost web platforme je što su podaci vezani uz studiranje certificirani od strane visokoškolske ustanove.

Da bi AdriaHUB platforma postala samoodrživa nakon završetka AdriaHUB projekta, bilo bi potrebno uključiti u platformu veći broj akademskih ustanova i njihovih studenata, što bi dovelo do registracije većeg broja poduzeća i obratno. Mogućnost za budući rad je i uključivanje svih akademskih ustanova i njihovih diplomanata u državama sudionicama projekta, pa tako i Hrvatskoj, u AdriaHUB platformu. Ona bi tada, kao što je to danas u Italiji platforma AlmaLaurea, postala vrijedan izvor podataka pri odabiru studija, analizi kvalitete sveučilišta i studija, zapošljivosti studenata s pojedinih studija i regija, visini očekivane plaće pojedinih profila stručnjaka, duljini čekanja na posao, i raznih drugih statističkih podataka za srednjoškolce, studente, sveučilišta, zavode za zapošljavanje, privredne komore, ministarstva obrazovanja i gospodarstva, i sve druge koji se susreću s tematikom studiranja i zapošljavanja diplomiranih studenata.

LITERATURA

[1] IPA Adriatic Cross-border Cooperation Programme Version as of 20/11/2013, http://www.adriaticipacbc.org/download/PROGRAMME_DOCUMENTS/IPA_OP_amended_20112013.pdf

[2] M. Savioa, M. Stefanovic, C. Fragassa, "Merging Technical Competences and Human Resources with the Aim at Contributing to Transform the Adriatic Area in Stable Hub for a Sustainable Technological Development", International Journal of Quality Research 2016; 10(1):1-12. http://ijqr.net/journal/v10-n1/0.pdf

[3] Službene stranice AdriaHub projekta www.adria-hub.eu/

[4] AlmaLaurea www.almalaurea.it/

[5] AdriaHUB platforma international.almalaurea.it/adriahub/?lang=hr

[6] AdriaHUB projekt – Sveučilište u Rijeci Tehnički fakultet, www.riteh.uniri.hr/adria_hub/

[7] M. Živković, N. N. Nedić, C. Fragassa, I. Vicović, "Drvo i tehnološki sistemi obrade na teritoriji zapadnog Balkana," Fakultet inženjerskih nauka Univerziteta u Kragujevcu, Kragujevac, 2015.

Mjerenje kvaliteta usluge mobilnog placanja m:Pay

S.Salapura, PhD*, Vesna Žlof, dipl.ing.*
* Mtel a.d. Banja Luka, BiH
sasa.salaura@mtel.ba, vesna.zlof@mtel.ba

Sažetak - Kako je sve više kompanija prisutno on-line a sve više korisnika je mobilno, istraživanje odrednica i mjerenje kvaliteta mobilnih servisa postaju od ključnog značaja. Istraživanja u ovoj oblasti još uvijek su u začetku, postojeća istraživanja su sporadična i posmatraju uslugu doživljenu od strane krajnjih korisnika ili iz perspektive poslovnih korisnika, a rezultat su uglavnom ograničena saznanja.

Ovaj rad predstavlja pokušaj da se popuni ovaj nedostatak u stručnoj literaturi. Okosnica ovog rada je m-pay: novokreirani servis za mobilno plaćanje računa telekomunikacione kompanije Mtel a.d. Banja Luka, jedne od tri telekomunikacine kompanije u BH. Mtel je prva kompanija u regiji koja je svojim korisnicima ponudila ovaj savremen način plaćanja računa za telekomunikacione usluge.

U ovom radu biće odgovoeno na pitanje da li mogu korisnici ovog servisa ocjeniti kvalitet servisa putem trinaest ključnih atributa kvaliteta mobilnog servisa: dizajn mobilne aplikacije, jednostavnost korištenja, tehnička pouzdanost, korisnost, intuitivnost, prezentacija, tačnost, raspoloživost, pouzdanost servisa, fleksibilnost, sigurnost i vjerodstojnost, ranjivost i tajnost.

Biće primjenjen prilagođeni alat za ocjenu kvaliteta servisa mobilnog plaćanja(WebSQ skala) te prikazani rezultati mjerenja.

Pokazaćemo da li se WebSQ alat takođe može koristiti za dijagnostiku i tako pomogne davaocima usluga pri identifikaciji slabosti u njihovim servisima, omogućivši uvođenje korektivnih mjera.

Ključne riječi: WebSQ upitnik, mobilno plaćanje, kvalitet elektronskih usluga QoS

Abstract - Since the number of companies present online is growing, as well as the number of mobile phones users, the study of indicators and measuring the quality of mobile services become essentially important. Researches in this area have just started, the existing studies are sporadic and only observe the service as seen by the end-users or from the perspective of business users, which results in generally limited knowledge.

This paper is an attempt to fill the gap in the expert literature. The backbone of this paper is m:Pay: new service for mobile payment of bills issued by the telecommunication company Mtel a.d. Banja Luka, one out of three telecommunication companies in BH. Mtel is the first company in the region which offered to its users this modern way of paying bills for telecommunication services.

In this paper we will provide answer to the question whether the users of this service may evaluate the service quality on the basis of thirteen key features of the mobile service quality: mobile application design, simplicity of use, technical reliability, usability, intuitiveness, presentation, accuracy, availability, service reliability, flexibility, security and authenticity, vulnerability and privacy.

We will apply customized tool for the assessment of the quality of mobile payment service (WebSQ scale) and present the results of the measurement.

We will also show if the WebSQ tool can be used for diagnostics and thus help service providers to identify the weaknesses of their services and allow them to introduce the correction measures.

Key words: WebSQ questionnaire, mobile payment, quality of electronic services Qos

I. UVOD

Savremene kompanije nastoje da koliko god mogu poboljšaju odnose sa potrošačima, korisnicima ili klijentima. Odlični odnosi su za kompanije od izuzetnog značaja iz razloga što samo one kompanije koje kontinuirano unapređuju odnose sa svojim korisnicima i koje su u stanju da kod njih izazovu zadovoljstvo još boljom uslugom, sasvim novom ili poboljšanom postojećom, te se na taj način u svijesti korisnika pozicioniraju kao kompanije koje vode računa o svojim potrošačima, postaju konkurentije na tržištu.

Kao posljedica ovih nastojanja se može identifikovati i briga zaposlenima, kojima je olakšavano izvršavanje postavljenih zadataka, pa je jasno da će se efikasno obavljen zadatak uz zadovoljstvo zaposlenih odraziti na zadovoljstvo korisnika ili potrošača. Tražeći odgovor kako se postaviti na ove izazove, kompanije koje su i savremene i odgovorne su razvile više modela, a jedan od njih je i upotreba mobilnih aplikacija.

Upravo zato se u ovom radu razmatra značaj mobilnih aplikacija u postizanju konkurentnosti institucija različitih djelatnosti, kako na zadovoljstvo svojih korisnika, tako i na povećanju efikasnosti zaposlenih uz ostvarivanje korporativnih ciljeva kompanije.

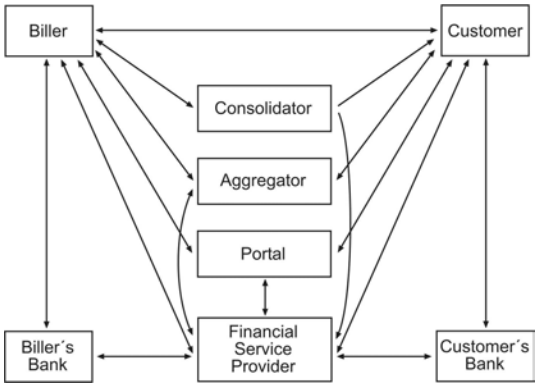
U radu se želi pokazati da postojeći model vrednovanja poslovnih procesa primjenom WEB servisa može da se primjeni i na mobilne aplikacije razvijene za potrebe telekomunikacione industrije. Rezultati dobijeni primjenom prilagođenog alata mogu pomoći ostvarivanju obostranih ciljeva (kompanije i korisnika), tj. postizanja koristi i za telekomunikacione kompanije i za krajnje korisnike.

II. USLUGA EBPP (ELECTRONIC BILL PRESENTMENT AND PAYMENT)

Usluga EBPP se javila kao odgovor na potrebe kompanija koje su smatrale da je tradicionalno upravljanje računima (dnevnim, sedmičnim ili mjesečnim) koje se sastoji od štampanja računa te slanja istih poštom, skupo i prevaziđeno te njihovih zahtjeva da se korisnik usmjeri da račune preuzima elektronski. Istovremeno je korisniku omogućeno da nakon uvida kroz isti kanal plati dospjeli račun. Neke kompanije su imale vlastite sisteme za elektronsku dostavu i plaćanje računa, neke su poslovanje prepustile novim kompanijama koje su namjenski formirane zbog ovog načina poslovanja, a treće kompanije su se opet odlučile za kombinaciju: sami će dostavljati račune a naplatu će prepustiti trećoj strani.

Ukratko, EBPP obuhvata kreiranje i dostavu računa uz istovremeno obezbjeđivanje najmanje jedne metode plaćanja.

Poslovni model jednog tipa EBPP usluge prikazan je na slici:



Slika 1. Poslovni model EBPP usluge; Izvor [1].

III. M:PAY MOBILNI SERVIS ZA PLAĆANJE TELEKOMUNIKACIONIH RAČUNA

Prateći trendove u razvoju i ponudi novih servisa u okruženju i šire, kompanija Mtel ad Banja Luka je u poslovnoj saradnji sa partnerskom poslovnom bankom razvila novi servis pod nazivom mpay koji nudi uslugu mobilnog plaćanja računa za telekomunikacione usluge kao i dopunjavanje prepaid računa upotrebom mobilne aplikacije na Android i iOS mobilnim platformama na jednostavan i siguran.

Krajnji korisnici usluge su Mtel korisnici koji imaju otvoren transakciojni račun u poslovnoj banci.

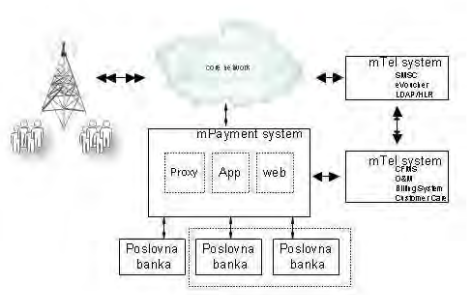
Osnovni skup funkcionalnosti mobilne aplikacije obuhvata:

- Aktiviranje usluge putem mobilne aplikacije
- Prijem i plaćanje vlastitih m:tel korisničkih računa
- Prijem i plaćanje m:tel korisničkih računa trećih lica (po referenci)
- Dopunjavanje prepaid računa
- Pregled izvršenih transakcija
- Podešavanja – promjena PIN koda

Usluga mpay je dostupna sa bilo kojeg mjesta na kojem je moguće pristupiti internetu mobilnim telefonom.

Da bi navedeni skup funkcionalnosti bio omogućen krajnjem korisniku, bilo je potrebno angažovati i u projekat uključiti veliki broj subjekata. Neki od njih su postojeći samostalni servisi, dok su ostali namjenski kreirani za potrebe servisa.

Pojednostavljena arhitektura sistema prikazana je na slici:



Slika 2. Pojednostavljena arhitektura m:payment sistema i veza sa okolinom, Izvor [2]

Centralni element strukture je mPay sistem koji je uvezan sa internim sistemima kompanije (LDAP, SMSC, mail server, eVoucher, Central Fault Management System, Billing, O&M clients), te poslovnim bankama [2].

IV. KVALITET ELEKTRONSKIH USLUGA

Od prvih dana prodaje robe i dobara putem interneta, internet se izdvojio i ustoličio kao veoma važan kanal za prodaju. Istovremeno su se pojavile kompanije koje proizvode i usluge nude isključivo putem interneta, dok

skoro sve kompanije kreiraju vlastite WEB stranice koje koriste kao izvor informacija o proizvodu te kao alternativni način prodaje proizvoda i ponude usluga, te kao način da privuku klijente sa ograničenim vremenom za obavljanje pojedinih usluga (online bankarstvo).

U početku se mislilo da su niske cijene i prisustvo na internetu dovoljni za uspjeh. Međutim, uskoro se pokazalo da je kvalitet usluga mnogo važniji faktor za uspjeh [3]. Uvidjelo se da kada potrošači/kupci nisu mogli izvršiti transakcije kupovine i plaćanja, kada proizvodi nisu dostavljeni na vrijeme ili nisu uopšte dostavljani, kad na elektronsku poštu nije pravovremeno odgovarano, a kad se željenim informacijama nije moglo pristupiti, da je održivost internet kanala za prodaju usluga ugrožen [4].

Mnogobrojni dokazi iz poslovanja kao i iz akademske literature pokazuju nedostatak adekvatnog kvaliteta usluge koja se pruža putem Interneta [5]. Ukoliko kompanije žele da potencijalni korisnici prihvate njihove internetske prezentacije, one moraju preusmjeriti fokus svog elektronskog poslovanja sa e-trgovine na e-usluge, znači sa same transakcije na sve ono što se dešava prije, u toku i nakon transakcija [4].

Proširivanje koncepta kvaliteta usluga na elektronski kontekst opisano je u radu [6] gdje je utvrđeno da pet dimenzija kvaliteta usluge [7] može biti svedeno na tri dimenzije kvaliteta elektronske usluge (a) opipljivost (b) kombinovana dimenzija reagovanja, pouzdanosti i garancije, i (c) saosjećanje. Ustanovljeno je da je opipljivost najbitnija dimenzija u povećanju lojalnosti kupaca, dok je kombinovana dimenzija najkritičnija u povećanju povjerenja kupaca. Međutim, kriteriji navedeni u skali su promijenjeni da bi bili prilagođeni elektronskom kontekstu (npr. opipljivost je predstavljena kao kriterij o izgledu WEB stranice) tako da skale nisu dirketno uporedive. Iz ovog razloga proučavanje kvaliteta elektronske usluge zahtijeva razvoj nove skale koja nije jednostavno prilagođavanje tradicionalnih skala na novi kontekst.

U radu [4] je ustanovljeno da postoje četiri dimenzije kvaliteta elektronskog servisa (WebSQ).

To su:

- kvalitet Web lokacije (WQ),
- kvalitet informacije (QI),
- kvalitet elektronskog servisa (SQ) te
- sigurnost servisa (SeQ).

Generalno, kvalitet web lokacije je vezan sa njenom upotrebljivošću i predstavlja odgovor na pitanje koliko je laka za korišćenje te koliko efikasno će korisnici ispuniti svoje zadatke [8].

Mjera kvaliteta informacija se uglavnom fokusira na rezultat proizveden od strane sistema, te na vrijednost, korisnost i važnost sistema na način kako je doživljen od strane korisnika [9].

Cijeli niz provedenih istraživanja identifikovala su više kriterija koje potrošači koriste pri ocjenjivanju i vrednovanju kvalite, a posebno su istaknuti atributi koji se odnose na:

- raspoloživost elektronskih servisa i dostupnost osoblja zaduženog za podršku,
- pouzdanost servisa (sigurne, pravovremene, pouzdane i ispravne transakcije), te
- fleksibilnost (fleksibilnost u mogućnosti pristupa servisu i kontaktiranju davaoca usluga) [4].

Jedan od identifikovanih kriterija koje korisnici koriste pri ocjenjivanju Web usluga je privatnost i sigurnost [11]. Atributi servisa dobijeni od strane ispitanika [4] odnose se na:

- sigurnost i vjerodostojnost,
- ranjivost
- tajnost

Na osnovu prikupljenih podataka razvijen je model vrednovanja poslovnih procesa primjenom web servisa [4]. Prikupljeni podaci su iskorišteni da se odrede determinante i osnovne dimenzije elektronsih bankarskih servisa, kao i relevantni atributi kvaliteta. Rezultati tog istraživanja su iskorišteni za razvoj modela vrednovanja bankarskih Web servisa, te kako je razvijen i unaprijedjen alat za mjerenje tih servisa. Svrha ovog alata je efikasnije vrednovanje poslovnih procesa banke primjenom Web servisa.

Cilj ovog rada je da se pokaže da li se postojeći model može iskorisiti da korisnici ocjene kvalitet mobilnog servisa putem trinaest ključnih atributa kvaliteta mobilnog servisa: dizajn mobilne aplikacije, jednostavnost korištenja, tehnička pouzdanost, korisnost, intuitivnost, prezentacija, tačnost, raspoloživost, pouzdanost servisa, fleksibilnost, sigurnost i vjerodstojnost, ranjivost i tajnost.

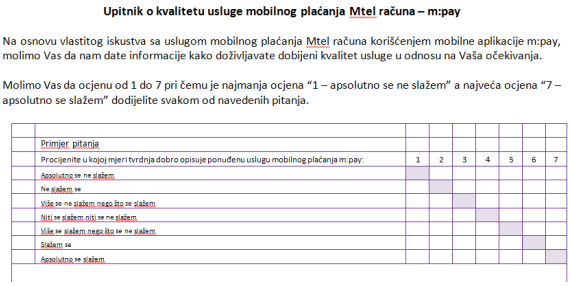
Takođe će se ponuditi odgovor na pitanje da li se razvijeni alat za ocjenu kvaliteta elektronskog bankarskog servisa (WebSQ) može prilagoditi i iskorisiti za ocjenu kvaliteta servisa mobilnog plaćanja m:pay.

V. OCJENA KVALITETA SERVISA MOBILNOG PLAĆANJA M:PAY

U cilju ocjenjivanja kvaliteta servisa mobilnog plaćanja m:pay, postojeći alat za ocjenu kvaliteta elektronskog bankarskog servisa (WebSQ) u formi upitnika je prilagođen za ovu namjenu.

Upitnik je ponuđen manjem broju korisnika postojećeg servisa i to iz praktičnih razloga: datum komercijalnog početka rada servisa i krajnji rok za predaju rada nalaze se u istom mjesecu. Međutim, kako se u ovom radu ne radi o ispitivanju kvaliteta servisa mobilnog plaćanja već o pokazivanju da li se postojeći model može, uz manja prilagođenja, iskoristiti za ispitivanje kvaliteta servisa mobilnog plaćanja, te demonstrirati mogućnosti upitnika i način tumačenja dobijenih podataka, smatramo da se manji broj ispitanika ne utiče bitno na postavljeno pitanje.

Korištena je 7-stepena Likertova skala gdje se na svako postavljeno pitanje moglo odgovoriti izborom ocjene od jedan do sedam. Najslabija ocjena je jedan („apsolutno se ne slažem“) dok je najjača ocjena sedam („apsolutno se slažem“). Neutralan stav („niti se slažem, niti se ne slažem“) odgovara ocjeni četiri. Likertova skala korišćena u prilagođenom upitniku prikazana je na slici:



Upitnik o kvalitetu usluge mobilnog plaćanja Mtel računa – m:pay

Na osnovu vlastitog iskustva sa uslugom mobilnog plaćanja Mtel računa korišćenjem mobilne aplikacije m:pay, molimo Vas da nam date informacije kako doživljavate dobijeni kvalitet usluge u odnosu na Vaša očekivanja.

Molimo Vas da ocjenu od 1 do 7 pri čemu je najmanja ocjena "1 – apsolutno se ne slažem" a najveća ocjena "7 – apsolutno se slažem" dodijelite svakom od navedenih pitanja.

Primjer pitanja						
Procijenite u kojoj mjeri tvrdnja dobro opisuje ponuđenu uslugu mobilnog plaćanja m:pay:	1	2	3	4	5	6	7
Apsolutno se ne slažem							
Ne slažem se							
Više se ne slažem nego što se slažem							
Niti se slažem niti se ne slažem							
Više se slažem nego što se ne slažem							
Slažem se							
Apsolutno se slažem							

Slika 3 Likertova skala korištena u upitniku

U nastavku su ilustracije koje prikazuju slika uzoraka dobijenih odgovora za pojedina pitanja.



Korisnički servis reaguje na hitne zahtjeve u razumnom roku.

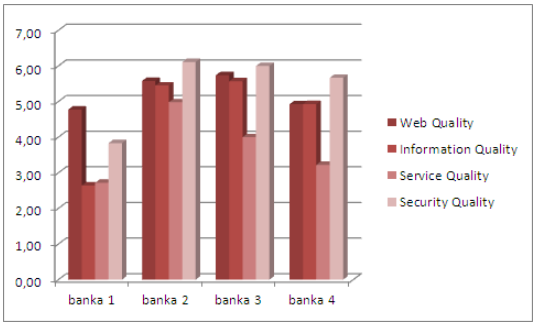
Ocjena	Broj ispitanika	
	Frekvencija	%
Ocjena 1	0	0%
Ocjena 2	0	0%
Ocjena 3	2	33%
Ocjena 4	1	17%
Ocjena 5	1	17%
Ocjena 6	2	33%
Ocjena 7	0	0%
Ukupno	6	100%

Slika 4 Slika uzorka za pitanje: „Korisnički servis reaguje na hitne zahtjeve u razumnom roku“

Servisi koje neka kompanija nudi mogu na različite načine da budu pod uticajem uključenih partnera. Negdje će vanjski partner da svojim kvalitetom pospješi pozitivno mišljenje o ponuđenom servisu, ali nisu rijetke situacije da se o servisu misli loše isključivo zbog lošeg segmenta koji je pod direktnim uticajem partnera.

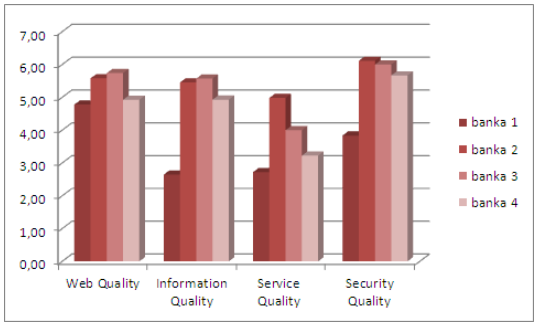
Zato primjenom ovog alata možemo vidjeti da li korisnici koji su nezadovoljni jednim od atributa ponuđenog servisa mogu da se grupišu na osnovu vanjskog partnera.

Moguća situacija pri ocjenjivanju kvaliteta servisa u zavisnosti sa kojom komercijalnom bankom se obavlja plaćanje prikazana je na narednim slikama.



Slika 5 Srednja ocjena banke

Na osnovu ovakvih rezultata istraživanja možemo zaključiti da su korisnici naše mobilne usluge najzadovoljniji ukoliko svoje poslovanje završavaju preko poslovne banke broj 2, a nešto slabiju ocjenu je dobila poslovna banka broj 3. Istovremeno vidimo koje segmente kompanija koja pruža mobilnu uslugu treba unaprijediti u poslovnoj banci 1 da bi izbjegla odlazak korisnika za koji nije direktno odgovorna. Sada, kada ima ovu informaciju, kompanija je u mogućnosti da unaprijedi ponuđeni servis.



Slika 6 Uporedni pregled dimenzija po bankama

Istovremeno, na osnovu rezultata istraživanja i prikaza kao na slici 6, možemo vidjeti da je kompanija veoma dobro procjenila važnost prve dimenzije (WQ). U konkrentom slučaju, ova dimenzija i ne zavisi previše od poslovnih banaka, tj. samo kompanija utiče na kvalitet ove dimenzije. S druge strane, možemo vidjeti da korisnici previše zadovoljno sa dimenzijama IS i SQ. Za ove dimenzije znamo da nisu pod kontrolom samo kompanije. Naprotiv, na njih skoro velikim dijelom utiču poslovne banke sa kojima kompanija sarađuje. U ovakvoj situaciji, preporuka za kompaniju bi bila da pokuša sa poslovnom bankom 1 da unaprijedi dimenzije IQ i SQ [4].

VI. ZAKLJUČAK

U ovom radu je na primjeru novokreiranog servisa za mobilno plaćanje m:pay prikazana primjena postojećih alata za ocjenu kvaliteta elektronskog poslovanja.

Pokušali smo dati odgovor na pitanje da li korisnici ovog servisa mogu ocjeniti kvalitet servisa putem trinaest ključnih atributa kvaliteta mobilnog servisa: dizajn mobilne aplikacije, jednostavnost korištenja, tehnička pouzdanost, korisnost, intuitivnost, prezentacija, tačnost, raspoloživost, pouzdanost servisa, fleksibilnost, sigurnost i vjerodstojnost, ranjivost i tajnost.

Primjenjen je prilagođeni alat za ocjenu kvaliteta servisa mobilnog plaćanja te su prikazani rezultati mjerenja.

Pokazano je da se isti alat takođe može koristiti za dijagnostiku čime pomaže kompaniji koja pruža m:pay usluga da identifikuje slabosti vlastitih servisa omogućivši uvođenje korektivnih mjera.

Ovo istraživanje je bilo ograničeno na utvrđivanje odrednica i mjerenja WebSQ i nije uzimalo u obzir relativnu važnost svake od utvrđenih dimenzija u odnosu na njihov doprinos ukupnoj procjeni kvalitete servisa.

Kada bi davaoci usluga bili u mogućnosti da znaju koji su aspekti elektronski pruženih usluga od najveće važnosti korisnicima, bili bi u mogućnosti da bolje iskoriste (ograničene) resurse i da usmjere razvoj na ta kritična područja.

Takođe, ovo istraživanje ne prati kako se ocjenjivani kriteriji WebSQ mijenjaju tokom vremena tako da se treba nadati da će dugotrajne studije davaocu usluge dati uvid u promjenjive potrebe usluge.

Ovo je prvi put da se ovaj alat koristi za ocjenu usluge elektronskog poslovanja u drugim, tj. nefinansijskim djelatnostima i očekivano ne postoje dokazi o razlikama u kriterijima koji se koriste za procjenu WebSQ za razne djelatnosti.

Prethodno je u [4] rečeno da bi u smislu dobijanja potpunije slika o WebSQ-u u finansijskoj djelatnosti, trebalo istražiti i ulogu treće strane davaoca usluge (telekomunikacione kompanije, davaoci internet usluga, kartične kuće i procesori kartičnih transakcija) te njihov uticaj na doživljeni kvalitet WebSQ od strane korisnika. Takođe je navedeno da nivo servisa isporučenog od treće strane (posebno davaoci internet usluga) može imati uticaj (pozitivan ili negativan) na doživljeni kvalitet usluge od strane korisnika, tako da bi se naredna istraživanja mogla odvijati i u ovom pravcu.

Provođenjem konkretnog istraživanja sa precizno prilagođenim alatom koji bi obuhvatio aspekte usluge mobilnog plaćanja računa može se pokazati da je uticaj treće strane na kvalitet elektronske usluge dvosmjeran i obostran: već dugo znamo da je kvalitet elektronskih servisa koje nude finansijske institucije u mnogome podložan uticajima telekomunikacionih kompanija koje su prisutne kao partner za davanje internet usluga, a ovaj rad nagovještava ono što buduća istraživanja treba da potvrde a to je da i finansijske institucije mogu da utiču na kvalitet servisa koji je ponudila telekomunikaciona kompanija.

LITERATURA

[1] K.Terplan, "Electronic Bill Presentment and Payment", Copyright © 2003 CRC Press, LLC

[2] V.Žlof, S.Salapura, "Mobile payment service m:Pay", Mipro 2016, Opatija, Croatia, 2016

[3] Ahmad S.,"Service Failures and Customer Defection: A Closer Look at Online Shopping Experiences", ManagingService Quality, 12 (1), 19-29., 2002

[4] Salapura S. „Model vrednovanja poslovnih procesa primjenom web servisa“, disertacija, Beograd, 2015

[5] Lennon R., Harris J., "Customer Service on the Web: A Cross-Industry Investigation", Journal of Targeting, Measurement and Analysis for Marketing, 10 (4), 325-38., 2002;

[6] Gefen D., "Customer Loyalty in E-Commerce", Journal of the Association for Information Systems, 3, 27-51., 2002

[7] Parasuraman, Valarie A. Zeithaml, Leonard L. Berry, "A Conceptual Model of Service Quality and Its Implications for Future Research", The Journal of Marketing, Vol. 49, No. 4 (Autumn, 1985), pp. 41-50

[8] Zeithaml, V.A., Parasuraman, A. and Malhotra, A. “e-service quality: definition, dimensions and conceptual model” Marketing Science Institute Working Paper Series, Cambridge, MA, (2000)

[9] Grover, V., Teng, J.T.C., and Fiedler, K.D. “Investigating the Role of Information Technology in Building Buyer-Supplier Relationships” Journal of the Association for Information Systems (3) pp.217-245, 2002

[10] Zeithaml, V.A. “Service excellence in electronic channels”, Managing Service Quality 12 (3) pp.135-138, 2002

[11] Hoffman, D.L., Novak, T.P., and Peralta, M.A. “Building Consumer Trust Online”, Communications of the ACM 42 (4) pp.80-85., 1999

Uspostava sustava upravljanja identitetima u Carinskoj upravi

Miljenko Hajnić, struč.spec.ing.comp.*, dr.sc. Drago Cmuk **
* Ministarstvo financija, Carinska uprava, Zagreb, Hrvatska
** KONČAR Electronics and Informatics Inc., Zagreb, Hrvatska
mhajnic@carina.hr
dcmuk@koncar-inem.hr

Ovaj rad obrađuje uspostavu središnjeg sustava upravljanja digitalnim identitetima i ovlaštenjima (IdM) u informacijskom sustavu Carinske uprave, velike nacionalne organizacije s preko 3000 zaposlenika. Osnovna ideja rada je prikazati kako su u zatečenom sustavu administracije korisnika i njihovih ovlaštenja detektirani problemi i ograničenja, koje su aktivnosti poduzete na planiranju i dizajniranju novog sustava, koje tehnologije i alati su korišteni za izradu informatičkog rješenja s centraliziranom autentikacijom, referentnim imeničkim sustavom identiteta korisnika, aplikacija, poslovnih rola, radnih mjesta i organizacijskih jedinica, a sve u cilju uštede na troškovima administracije, povećanju učinkovitost i ubrzanju protoka roba i usluga. Rezultat rada je analiza i opis zatečenog stanja, financijski izračun učinaka administracije korisnika te razrada i unaprjeđenje ključnih procesa kao visoko razrađena podloga za tehničku provedbu projekta. Izgrađena je arhitektura novog sustava i opisan tijek provedbe i testiranja informatičkog rješenja središnjeg IdM sustava koji osigurava povrat investicije već u prvoj godini korištenja.

I. UVOD

Zatečeni sustav upravljanja identitetima (engl. Identity Management, skraćeno IdM) Carinske uprave (CU) oslanjao se na proces unosa i odobravanja novog zaposlenika radi kreiranja domenskog korisničkog računa za pristup informacijskom sustavu. U CU se svakodnevno koristi veliki broj poslovnih aplikacija. Većina poslovnih aplikacija ima vlastiti sustav autentikacije i autorizacije, a korisnici zbog velikog broja aplikacija često zaboravljaju pristupne lozinke. Centralizirani sustav autentikacije i autorizacije nije postojao, a broj poslovnih aplikacija konstantno se povećava. Administriranje korisnika predstavljalo je sve veće opterećenje, a krajnji učinak odražava se na državni proračun RH i proračun EU, jer je EU prvenstveno carinska unija.

Rješenje problema i ublažavanje posljedica pronađeno je u razvoju središnjeg sustava upravljanja identitetima i podizanju organizacijske svijesti o potrebi prilagođavanja modernim načinima poslovanja. CU mora jednostavnije i brže upravljati ovlaštenjima i korisnicima Informacijskog sustava Carinske uprave (ISCU) kako bi mogla uspješno pratiti trendove kretanja i razmjena roba i usluga, naplatiti obvezna davanja u proračun te primjereno pratiti dinamiku poslovanja Carinske uprave.

Rad je koncipiran u nekoliko poglavlja koja opisuju početno stanje sustava, analizu, opseg poslova, arhitekturu, planirane i odrađene aktivnosti kroz faze uspostave IdM sustava, zaključak te popis korištene literature.

Radi osiguranja sigurnosti i povjerljivosti određeni dijelovi sustava nisu opisani u detalje kao i način njihovog rješavanja. Određeni nazivi, šifre, adrese i podaci korišteni u ovom radu su izmijenjeni radi zaštite ISCU na način da izmjenom ne utječu na ideju i rezultat rada.

II. POČETNO STANJE SUSTAVA

A. *Organizacijska struktura*

Radi razumijevanja zatečenog stanja sustava i procesa koji se odvijaju u CU, potrebno je razumjeti organizacijsku strukturu. CU broji preko 3000 zaposlenika, a djelatnost obavlja na teritoriju cijele RH. Svaka ustrojstvena jedinica CU, njih preko 200, je informatizirana i nalazi se unutar iste domene. Informatička infrastruktura prati organizacijsku strukturu. Oko 70% zaposlenika CU odrađuje poslove u uredovno vrijeme od ponedjeljka do petka, od 8 do 16 sati, a ostatak zaposlenika radi u smjenama i poslove odrađuje 365 dana u godini, od 0 do 24h.

B. *Informatizacija procesa i tehnologije*

Informatička podrška poslovnim procesima u CU pruža se putem nekoliko različitih aplikativnih rješenja razvijenih i izrađenih u suradnji s nekoliko različitih dobavljača. Zaposlenici u svakodnevnom radu koriste od 7 do 10 različitih aplikacija. Dio aplikacija poslužuje se iz domene CU dok se ostatak aplikacija poslužuje putem vanjskih partnera. Aplikativna rješenja izrađena su na različitim tehnologijama, a svako aplikativno rješenje ima samostalnu bazu korisnika s pripadajućim podacima za autentikaciju i autorizaciju. Imenički sustavi i baze korisnika nisu međusobno povezani niti postoji sinkronizacija između baza korisnika na aplikacijama. Svaki zaposlenik CU u prosjeku koristi od 7 do 10 različitih korisničkih imena i lozinki.

C. *Politike*

Sigurnosne politike CU postavljene su tako da zaposlenici moraju periodično mijenjati lozinke. Takva politika ima opravdanje jer osigurava višu razinu sigurnosti, ali se negativno odražava na zaposlenike koji u kraćim vremenskim intervalima moraju pamtiti puno različitih lozinki, ali i na administratore korisničkih računa. Rezultat provođenja sigurnosne politike je veliki broj korisničkih zahtjeva za resetiranje lozinke.

D. *Ovlaštenja i administracija*

Poslovni zadaci obavljaju se na razini i nadležnosti organizacijske jedinice u kojoj zaposlenik radi. Hijerarhijska struktura poštuje se prilikom dodjele ovlaštenja pristupa resursima. Zaposlenik ne može pristupiti resursima druge organizacijske jedinice bez obzira što se nalaze u istoj razini, a niti resursima viših razina te se zaposlenici premještaju u druge organizacijske jedinice kako bi se poslovni zadaci mogli obaviti na vrijeme.

Posljedica takvih migracija korisnika je dnevna obrada između 20 i 50 zahtjeva za ažuriranje korisničkih računa. Propisani kanali za dostavu zahtjeva za ažuriranje korisničkih računa su e-mail, fax i redovna pošta. Aktivnosti administratora uključuju kreiranje računa, resetiranje lozinki, promjene ovlaštenja, promjene mjesta rada i deaktiviranje korisničkog računa. Procedura administracije može potrajati i po nekoliko sati, ovisno o količini zahtjeva i broju administratora.

III. ANALIZA

Tablica 1. prikazuje godišnji bruto utrošak na plaće IT administratora [1] koji upravljaju identitetima i ovlaštenjima, a odražava se kroz najmanje jednog administratora na svakoj domeni, što je na godišnjoj razini trošak od 36 prosječnih osobnih dohodaka [2].

TABLICA 1. GODIŠNJI BRUTO UTROŠAK NA PLAĆE ADMINISTRATORA

Broj administratora korisnika	Broj mjeseci	Prosječni bruto dohodak administratora	Trošak plaća administracije korisnika
3	12	12.219,81 kn	439.913,16 kn

Trošak upravljanja odražava se još i kroz trošak „hladnog pogona“ zaposlenika dok čekaju ažuriranje korisničkog računa. Za izračun troška hladnog pogona koristit će se podaci o robnoj razmjeni RH s inozemstvom za razdoblje od siječnja do prosinca 2014.godine koje iznosi Državni zavod za statistiku [3] (DZS) te podataka o ukupnom prihodu CU koji je u 2014.godini iznosio 20.925.654.994 kuna [4]. Oko 2100 zaposlenika (70% od ukupno zaposlenih) izravno sudjeluje na poslovnim postupcima koji imaju financijsku posljedicu za CU. Tablica prikazuje broj zaposlenika koji sudjeluje u postupcima s financijskom posljedicom te vrijednosti robne razmjene RH s inozemstvom.

Broj službenika	Robna razmjena – uvoz u RH (kn)	Robna razmjena – izvoz iz RH (kn)	Ukupan prihod (kn)	Broj radnih dana u godini
2100	130,67 mlrd.kn	79,09 mlrd.kn	20,92 mlrd.kn	365

U prosjeku se dnevno odradi 35 zahtjeva za ažuriranjem korisničkih računa, što uključuje promjenu mjesta rada i ovlaštenja, resetiranja lozinki, kreiranja i deaktiviranja korisničkih računa. 35 zaposlenika čini 1,67% zaposlenika koji izravno sudjeluju u postupcima s naplatom.

Za potrebe inicijalne analize korišteno je prosječno vrijeme obrade zahtjeva od 1h, a to je vrijeme kada zaposlenik ne može odrađivati poslovne zadatke. Tablica 2. prikazuje izračun prosjeka naplate prihoda po jednom zaposleniku u sat vremena (Navedena kalkulacija ne govori kako bi izostali prihod nužno bio ostvaren, već pokazuje ekstrapoliranu procjenu gubitaka koji se mogu očekivati u sličnim uvjetima. Utjecaj drugih poslova koji se mogu raditi za vrijeme nedostupnosti sustava kao ni kumulativne posljedice periodičkog pauziranja procesa nisu uzete u obzir jer nadilaze okvire ovog rada.).

TABLICA 2. IZNOS NAPLATE PRIHODA PO ZAPOSLENIKU NA SAT

20.925.654.994 kn / 365 dana / 2100 zaposlenika / 24 sata = 1.137,51 kn

Temeljem gornjeg izračuna vidljivo je kako svaki zaposlenik, koji izravno sudjeluje u prihodu CU, u prosjeku svaki sat ostvari prihod od 1.137,51 kn. Na godišnjoj razini trošak hladnog pogona za 35 zaposlenika, koji dnevno čekaju 1 sat na ažuriranje korisničkog računa, može se prikazati kao prihod koji se mogao ostvariti, a iznosi gotovo 15 milijuna kuna (Tablica 3.).

TABLICA 3. MOGUĆI OSTVARENI PRIHOD U GODINI DANA

35 zaposlenika * 1.137,51 kn * 365 dana + 439.913,16 kn = 14.971.475,66 kn

Financijski pokazatelji provedene analize ukazuju kako postoji nedvojbeni motiv za uspostavu središnjeg sustava upravljanja identitetima.

Analizom je također utvrđeno nekoliko područja za unaprjeđenje i nekoliko područja koja trenutno uopće nisu dokumentirana, pokrivena procedurama ili politikama.

IV. OPSEG POSLOVA

Opseg poslova na uspostavi IdM sustava koncentriran je kroz tri kategorije [5]:

- Standardizacija – dimenzioniranje, usklađivanje i prilagođavanje sustava i podataka standardima
- Certifikacija – utvrđivanje potrebnih kvalifikacija za sudjelovanje u procesu uspostave IdM sustava
- Upravljanje - utvrđivanje postojanja politika, procedura i dokumentacije procesa, njihova usklađenost s vizijom, poslovnom politikom CU i zakonskom regulativom te upravljanje odnosima s vanjskim partnerima.

A. *Upravljanje rizicima*

Za prikaz rizika u procesu uspostave IdM sustava korišten je model kvadranta, odnosno koordinatnog sustava u kojem su rizici raspoređeni temeljem utvrđenih vrijednosti utjecaja i vjerojatnosti pojave po osi x i y. Os x prikazuje vrijednosti vjerojatnosti pojave, a os y

prikazuje vrijednosti utjecaja rizika na uspostavu sustava [6]. Kvadrant čini temelj za donošenje odluka kojim rizicima treba upravljati i na koji način. Položaj rizika u kvadrantu određen je korištenjem vrijednosti utjecaja i vjerojatnosti pojave rizika (Tablica 4. i Tablica 5.).

Tablica 4. prikazuje utjecaj pojave rizika, a iskazuje se vrijednostima na skali od 1 do 100.

TABLICA 4. VRIJEDNOSTI UTJECAJA POJAVE RIZIKA

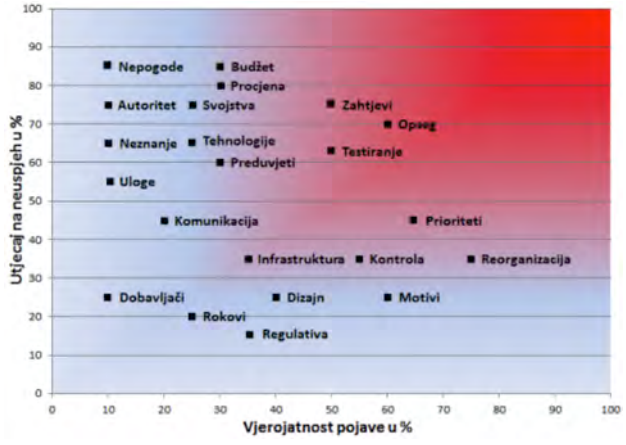
Utjecaj	Opasnosti	Vrijednost
Gotovo zanemariv	osnovne rutine su dovoljne za kontrolu rizika	1-20
Mali	neće ugroziti neke od aktivnosti programa	21-40
Umjeren	zahtijevat će prilagodbe programa	41-60
Velik	neće utjecati na neke od ciljeva i učinaka programa	61-80
Sigurna prijetnja	ako se pojave program će prestati postojati	81-100

Tablica 5. prikazuje vjerojatnost pojave rizika, a koja se iskazuje vrijednostima na skali od 1 do 100.

TABLICA 5. VRIJEDNOSTI VJEROJATNOSTI POJAVE RIZIKA

Vjerojatnost	Očekivanja	Vrijednosti
Gotovo nikad	samo u iznimnim situacijama	1-20
Malo vjerojatno	javlja se, ali nije uobičajeno	21-40
Moguće	može se javiti ponekad	41-60
Vjerojatno	javlja se u većini situacija	61-80
Gotovo sigurno	u pravilu se očekuje	81-100

Slika 1. prikazuje identificirane rizike i njihov položaj u kvadrantu.



SLIKA 1. UTJECAJ I VJEROJATNOST POJAVE RIZIKA

Za svaki identificirani rizik izrađen je plan aktivnosti i protumjera na uklanjanju ili ublažavanju posljedica nakon moguće pojave rizika.

B. *Određivanje uloga i odgovornosti*

Nakon odrađene inicijalne analize postojećeg stanja i utvrđivanja postoji li opravdani razlog za pokretanje daljnjih radnji, utvrđene su uloge i odgovornosti dionika budućeg sustava. Osnovne uloge u uspostavi IdM sustava su [5]: Publika - svi dionici sustava, Voditelj tima, Nadzornik, Direktor, Savjetnik, Komunikator i Odobravatelj. Timovi i članstvo u ulogama sastavljeno je od zaposlenika CU i članova vanjskih partnera. Ovakav pristup potreban je ukoliko se određene aktivnosti moraju provesti i na infrastrukturi vanjskog partnera koji za CU ima ulogu davatelja usluge.

V. DIZAJN IDM SUSTAVA

U fazi dizajniranja sustava odrađene su „aktivnosti vezane uz svaku komponentu arhitekture: Upravljački okvir, Tehnička referentna arhitektura, Arhitektura podataka, Procesna arhitektura, Okvir interoperabilnosti i Politike.“ [5]

A. *Upravljački okvir*

Utvrđeno je kako su struktura i djelokrug CU podložni promjenama. Strukturne promjene uglavnom su potaknute proširenjem djelokruga rada CU ili racionalizacijom financijskih troškova poslovanja. Promjena upravljačkog okvira nije česta pojava, a vjerojatnost promjene je jednom u 5 godina. U fazi dizajna poželjno je uključiti mogućnost jednostavne prilagodbe sustava organizacijskim promjenama kao jednu od funkcionalnosti sustava kako bi se u slučaju promjene strukture ili djelokruga rada sustav prilagodio bez većih intervencija.

B. *Tehnička referentna arhitektura*

„Tehnička referentna arhitektura ima za zadatak prikazati shemu postojećeg informacijskog sustava, koje su komponente u funkciji, međusobne veze i sve tehničke pojedinosti komponenti sustava, što će arhitektima biti temelj za izradu najboljeg rješenja i integraciju IdM sustava“ [5] u ISCU. Sve poslovne aplikacije izrađene su od strane vanjskih izvođača na temelju specifikacija CU. Za sve poslovne aplikacije postoji razvojna dokumentacija, zapisnik o preuzimanju, tehnička dokumentacija, izvorni kôd i korisničke upute za korisnike. Korisničke upute za administraciju korisnika ne postoje u CU jer se administracija korisnika i dodjela ovlaštenja vršila na strani vanjskog partnera.

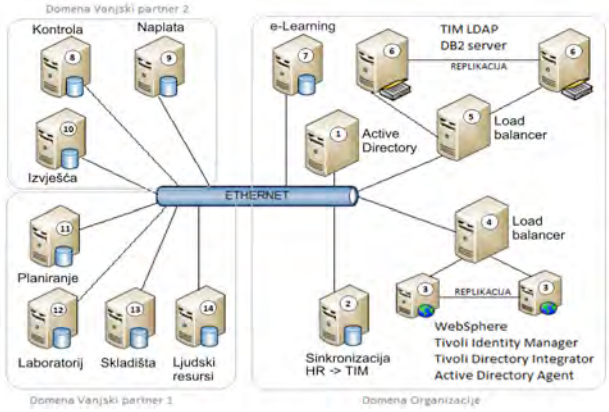
Osnovne tehničke komponente informacijskog sustava čine: Imenički sustav Active Directory, Sinkronizacijska baza podataka, WebSphere Application Server, WAS Network load balancer, Tivoli Identity Manager, DB2 load balancer, TIM LDAP, DB poslužitelj i aplikacije koje se poslužuju iz različitih domena.

Slika 2. prikazuje tehničku arhitekturu informacijskog sustava Carinske uprave, osnovne komponente i njihovu povezanost unutar i izvan domena partnera.

C. *Arhitektura podataka*

„Glavna namjena izrade arhitekture podataka je prikaz trenutnog stanja i budućeg stanja podataka o identitetima

koje će koristiti IdM sustav te njihova međusobna povezanost.“ [5]



SLIKA 2. TEHNIČKA ARHITEKTURA SUSTAVA

Svaka poslovna aplikacija CU koristi vlastiti izvor podataka o korisnicima aplikacije i njihovim ovlaštenjima. Struktura podataka u tim izvorima međusobno se razlikuje jer je svaki izvor podataka o korisnicima dizajniran posebno za potrebe te aplikacije. Podaci prikladni za konsolidaciju podataka o korisnicima su ime i prezime korisnika, OIB i korisničko ime. Većina baza korisnika sadrži i OIB podatak, ali nemaju svi korisnici upisan OIB. Ne postoji zajednički jedinstveni identifikator identiteta već svaki imenički sustav i baza podataka o korisnicima koristi vlastiti jedinstveni identifikator. Pravilo kreiranja korisničkog imena postoji samo na domeni CU i primjenjuje se u AD-u.

Obzirom na veliki broj aplikacija koje posjeduju vlastite baze korisnika, na određenom broju aplikacija uključenim u proces uspostave IdM sustava izvršeno je usklađivanje podataka na način da je svaki identitet korisnika u bazama aplikacija povezan s pripadajućim identitetom korisnika u AD-u kroz federaciju imenika. Usklađivanjem podataka utvrđeno je da postoji određeni broj korisnika čiji jedinstveni identifikatori, odnosno korisnička imena ne odgovaraju domenskom korisničkom imenu u AD-u. Problem neusklađenosti riješen je zapisnički, konstatiranjem da su korisnici do određenog datuma i vremena koristili jedno korisničko ime, a nakon tog datuma nastavljaju koristiti drugo korisničko ime, koje odgovara njihovom domenskom korisničkom imenu.

Drugo rješenje navedenog primjera može se ostvariti kontroliranom zamjenom podatka o korisničkom imenu u evidenciji.

Usklađivanje podataka identiteta mora se odraditi radi održanja kontinuiteta zapisa i evidencija koje je jedan identitet ostvario. „Preporuka je da se jednom entitetu u CU dodjeljuje samo jedan digitalni identitet te da se atributi identiteta unose u sustav samo jednom.“ [7]

Cilj usklađivanja podataka identiteta je postizanje stanja u kojem jedan korisnik ima jedinstveno korisničko ime na svim poslovnim aplikacijama domene CU i domenama vanjskih partnera, a radi omogućavanja centralizirane autentikacije. Prvi razlog uvođenja centralizirane autentikacije je činjenica da korisnici moraju pamtiti nekoliko različitih lozinki za prijavu na aplikacije, a drugi razlog je dodatna sigurnosna mjera CU da se lozinka korisničkog računa pohranjuje na samo jednom mjestu, unutar CU, nikako kod vanjskog partnera.

Svakoj aplikaciji koja se uključuje u IdM potrebno je kreirati identitet. CU kreirala je identitete aplikacijama koje su obuhvaćene novim sustavom upravljanja identitetima.

Svaka poslovna aplikacija posjeduje vlastiti set poslovnih rola koje su na raspolaganju zaposlenicima. Poslovne role zapisane su u bazama podataka samih aplikacija. Poslovne role svih aplikacija potrebno je popisati i svakoj poslovnoj roli kreirati identitet sa svim potrebnim atributima. CU kreirala je identitete poslovnih rola koje su obuhvaćene novim sustavom upravljanja identitetima.

Šifrarnik organizacijskih jedinica koji se koristi u poslovnim aplikacijama ugrađen je u svaku aplikaciju koja ga koristi. Obzirom da se identičan šifrarnik može nalaziti na više različitih mjesta, ažuriranje šifrarnika predstavlja ponavljajući posao na svim aplikacijama koje ga koriste, uz veću mogućnost greške kod ažuriranja. Za uključivanje aplikacije koja koristi šifrarnik u IdM sustav potrebno je uspostaviti središnji referentni izvor podataka za organizacijske jedinice kreiranjem identiteta svakoj organizacijskoj jedinici.

Šifranik radnih mjesta (RM) zapisan je kroz zasebnu bazu podataka koju poziva TIM kada se želi odraditi određeni upit po korisniku. Šifrarnik RM u TIM-u, pored šifre i naziva radnog mjesta posjeduje i atribut hijerarhije koji se koristi za određivanje nadređenosti zaposlenika u CU i za potrebe dodjele ovlaštenja. Nadređena osoba ima pravo dodijeliti ovlaštenja traženom korisniku.

Kreiranjem pet prethodno opisanih grupa identiteta (korisnici, aplikacije, poslovne role, organizacijske jedinice i radna mjesta), centraliziranom pohranom tih identiteta u TIM i prihvaćanjem TIM-a kao referentnog izvora podataka od strane vanjskih partnera, ostvaruje se preduvjet kontroliranog upravljanja identitetima. Sve promjene podataka se evidentiraju, a mogu ih izvršiti isključivo ovlašteni korisnici.

Za uspješnu provedbu uvođenja IdM sustava potrebno je usuglasiti i odrediti pravila koja govore tko, što, gdje i kada smije koristiti ili ažurirati atribute ovih identiteta. Pravila se mogu uspješno kreirati putem matrica.

Tablica 6. pojednostavnjeno prikazuje uzajamne odnose između identiteta aplikacija i identiteta poslovne role, odnosno koje sve aplikacije može koristiti pojedina poslovna rola.

Uzajamne odnose između svih pet grupa identiteta (korisnici, aplikacije, poslovne role, organizacijske jedinice i radna mjesta) potrebno je definirati putem matrica. Ovakav pristup upravljanja ovlaštenjima omogućava da se novom zaposleniku ili novom korisniku

sustava dodijele sva potrebna ovlaštenja prije nego što je stigao na svoje novo radno mjesto.

TABLICA 6. ODNOS IDENTITETA APLIKACIJA I POSLOVNE ROLE

Aplikacije Poslovne role	e-Learning	Kontrola	Naplata	Izvješća	Planiranje	Laboratorij	Skladišta
Carinik I	X	X	X	X	X	X	X
Informatika	X	X	X		X	X	
Financije	X		X	X	X		X
Kadrovi	X			X	X		

D. *Procesna arhitektura*

Procesna arhitektura utvrđena je stvaranja predodžbe arhitektima sustava kako sustav funkcionira [5]. Popisani su svi poslovni procesi u CU koji koriste identitete i ovlaštenja, a čine ih procesi kreiranja identiteta zaposlenika, deaktiviranja identiteta, evidencija i kreiranje identiteta korisnika koji nisu zaposlenici CU, dodjela i promjena ovlaštenja, promjena korisničkih podataka, resetiranje lozinke i upravljanje identitetima aplikacija, poslovnih rola i šifrarnika.

Za analizu procesa u CU korišten je dijagram uzroka i posljedica, poznatiji kao Ishikawa dijagram koji omogućava identifikaciju, kategorizaciju i grafički prikaz uzroka problema [9]. Kao glavni uzrok problema fokusirano je loše upravljanje identitetima i ovlaštenjima.

Za modeliranje procesa korišteni su alati Bizagi Modeler i Bizagi Studio. Modeli procesa omogućili su identifikaciju 6 kategorija iz kojih dopiru uzroci lošeg upravljanja, a to su: Obrasci, Komunikacija, Korisnici, IT Administratori, Procesi i Informacijski sustav CU.

E. *Okvir interoperabilnosti*

Neophodno je ostvariti jednaka pravila pristupa za sva tijela s kojima CU ostvaruje informatičku razmjenu podataka, a u tu svrhu izrađen je Okvir interoperabilnosti. Dokument koji treba podržati da odvojeni sustavi čija su sučelja potpuno poznato, međusobno djeluju i funkcioniraju bez ikakvih ograničenja u pristupu i implementaciji“ [10].

Nakon usvajanja dokumenta Okvir interoperabilnosti od strane uprave, isti se treba priložiti uz svaki natječaj za izradu programskog rješenja ili objaviti na službenim web stranicama.

F. *Politike*

Politike predstavljaju interna pravila CU koja se provode radi provedbe poslovnih procesa u zadanim okvirima. CU usvojila je politike informacijske sigurnosti, kontrole pristupa, mrežne sigurnosti, upravljanja prijenosnim uređajima i radom na daljinu, upravljanja zapisima o radu sustava, upravljanja zaporkama te procedure i radne upute koje iz politika proizlaze.

Politike se često pozivaju na norme. „Politike su srce arhitekture IdAM sustava, definiraju prihvatljive aktivnosti i procese koji će se koristiti, a predstavljaju temelj za izvršavanje poslovnih procesa“ [5].

VI. KONFIGURIRANJE SUSTAVA

CU koristi IBM-ova rješenja za poslovne potrebe već duži niz godina, uključujući TIM i IBM Maximo Asset Management. Vanjski partneri također koriste IBM-ove tehnologije za posluživanje poslovnih aplikacija iz svojih okolina. Zbog zajedničke orijentiranosti prema IBM-ovim tehnologijama kao najprikladnije rješenje pokazali su se postojeći TIM i Maximo sustavi. TIM je nadograđen na verziju koja pruža sve potrebne funkcionalnosti. Odluka CU o korištenju TIM i Maximo alata za potrebe uvođenje IdAM sustava temelji se na slijedećim činjenicama: postojeći alati CU imaju sve potrebne funkcionalnosti, gotova rješenja drugih proizvođača nisu razmatrana jer bi nabava i uvođenje novih alata predstavljao dodatni trošak CU, IBM rješenja za upravljanje identitetima tehnološki su kompatibilna s velikim brojem poslovnih aplikacija, što će integraciju učiniti znatno lakšom i bržom u odnosu na potencijalna rješenja drugačijih tehnologija, IBM rješenja za upravljanje identitetima predstavljaju vrh svjetske ponude gotovih rješenja [11][12]. Iako na tržištu postoje besplatna ili jeftinija gotova rješenja, za postojeće alate koje CU koristi potrebna je i osigurana tehnička podrška za specifične konfiguracije i nadogradnje, što kod jeftinijih rješenja nedostaje ili se u konačnici dodatno naplaćuje. Količina dodatnih naplata može učiniti rješenje skupljim od postojećih alata.

Korištenjem postojećih alata kao dodatni trošak predstavlja samo izrada adaptera za svaku pojedinu aplikaciju koja nije izrađena na IBM tehnologiji, a koja će se integrirati s IdAM sustavom. Integracijski adapter je programsko rješenje koje transformira podatke jednog sustava prema formi, sintaksi i potrebama drugog sustava, a koji može biti izgrađen na drugačijoj platformi i koristiti drugačije tehnologije. Svrha adaptera je da se iz matičnog IdM sustava podaci vezani uz identitet propagiraju na drugi sustav, u njemu razumljivom obliku, a da se ti podaci pohrane na potpuno jednak način, kako bi to odradio IT administrator drugog sustava kroz vlastito sučelje.

Preduvjet za uspješno prebacivanje podataka na drugi sustav ili izmjenu podataka na drugom sustavu je potrebno ovlaštenje i uspostavljen odnos povjerenja između dva sustava. Ovlaštenje je u obliku korisničkog računa i vjerodajnica, a odnos povjerenja se ostvaruje putem SSL/TLS protokola (engl. Secure Sockets Layer, skraćeno SSL i Transport Layer Security, skraćeno TLS).

Prilagođavanje specijaliziranih alata, kao što su IBM Tivoli Identity Manager, IBM Tivoli Directory Integrator, IBM Maximo te dogradnje softvera na potrebne verzije odrađene su od strane vanjskog partnera CU na temelju održanih radionica u kojima je CU predstavila poslovne zahtjeve, funkcionalnu i nefunkcionalnu specifikaciju IdM sustava.

Maximo sustav lako se integrira sa TIM-om jer su oba alata razvijena od strane istog proizvođača i jednakim tehnologijama. Za ovu integraciju potrebno je povezati alate pomoću gotovog adaptera na način da Maximo preuzima i prikazuje podatke iz TIM-a u grafički i korisnički prihvatljivijem sučelju. Maximo pruža jednostavnije sučelje intuitivnog dizajna koje krajnji korisnik može koristiti bez dodatne obuke. Pomoću Maximo alata poslovne role mogu se dodjeljivati i uklanjati korisnicima ISCU. Za dodjeljivanje i uklanjanje poslovnih rola izrađena je web forma koja sadrži osnovne informacije o korisniku kojem se ovlaštenja dodjeljuju ili uklanjaju. TIM je prilagođen i konfiguriran prema postojećim povezanim sustavima i aplikacijama uključenim u IdM sustav.

VII. PROVEDBA SUSTAVA

„Za osiguranje kvalitetnog sustava, kvalitetom je potrebno upravljati. PMBOK® Guide definira tri osnovna elementa upravljanja kvalitetom: planiranje, osiguranje i kontrolu kvalitete. Planiranje kvalitete uključuje identifikaciju standarda kvalitete koji su relevantni za projekt i utvrđivanje kako ih ispuniti“ [8]. Carinska uprava izradila je dokument, Plan upravljanja kvalitetom, kako bi upravljačka struktura potvrdila da isporučeni rezultati sustava ispunjavaju svrhu. Plan upravljanja kvalitetom je formalni okvir koji definira procese i njihovu primjenu za osiguranje uspjeha provedbe, u skladu s odgovarajućim metodologijama i standardima.

Izrađen je i plan aktivnosti za osiguranje kvalitete radi praćenja kvalitete i kontinuiranog unaprjeđenja sustava. Plan aktivnosti uključuje elemente korisničkog zahtjeva i specifikacija IdM sustava. Temelju specifikacija utvrđuje se što će se i kako mjeriti, kako će se prikupljati podaci i s čime se prikupljeni podaci moraju usporediti. Za svaku aktivnost potrebno je rezervirati resurse, odrediti odgovornost, a sve aktivnosti grupirati u plan za osiguranje kvalitete. Za konačnu ocjenu projekta planirana je i eksterna revizija u kojoj će se kroz proces prikupljanja podataka i stručne procjene dokaza ocijeniti kvaliteta IdM sustava. Time će upravljačka struktura dobiti konkretne preporuke za daljnje poboljšanje sustava [9].

VIII. TESTIRANJE IDAM SUSTAVA

Testiranja sustava još nisu odrađena u potpunosti, a odradit će se i automatizirano i ručno, zbog prirode testova. Automatiziranim testiranjem izvršit će se pomoću unaprijed programiranih testnih scenarija, snimanjem radnji i rezultata pomoću alata za automatizaciju testiranja, a za to će se koristiti specijalizirani alati vanjskog partnera.

Statičkim testiranjem analizirana je dokumentacija, arhitektura, uporaba standarda, dijagrami toka, te programski kôd. Statičko testiranje provodi se periodički kroz razgovor s dobavljačem softvera, a provjera programskog kôda od strane stručne i kvalificirane osobe. Dinamičkim testiranjem provjerit će seprogramski kôd koji se izvršava na stroju. Dinamički testovi mogu biti funkcionalni i nefunkcionalni. Funkcionalni testovi provest će se na temelju funkcionalne specifikacije, a nefunkcionalni testovi na temelju nefunkcionalne specifikacije, kako bi se utvrdilo ispunjava li informatičko rješenje svrhu.

IX. ZAKLJUČAK

Ovaj rad prikazao je zatečeno stanje upravljanja identitetima u Carinskoj upravi, poteškoće i probleme uočene prilikom administriranja korisnika i njihovih ovlaštenja na sustavu. Uspostava IdM sustava ostvarena je arhitekturom koja definira pohranu identiteta korisnika, poslovnih aplikacija, poslovnih rola, radnih mjesta i ustrojstvenih jedinica u središnji IdM sustav sa referentnim sustavom šifara za sve poslovne aplikacije. Time su osim evidentnog unaprjeđenja poslovnog sustava ostvarene i znatne financijske uštede.

Rad donosi i tehnički okvir za odabir programskog rješenja za upravljanje identitetima Tivoli Identity Manager te integraciju s IBM Maximo programskim rješenjem. Navedenim je rješenjem osim optimiranja sustava i posljedičnih ušteda ostvareno je i korisnički usmjereno grafičko sučelje koje znatno olakšava korištenje sustava korisnicima čije zanimanje nije iz područja informatike ili računarstva. Implementacija sustava je trenutno u tijeku i očekuje se tijekom 2016. godine.

LITERATURA

[1] Administrator baze podataka, http://www.mojaplaca.hr/plata/informacijske-tehnologije/administrator-baze-podataka.

[2] Kalkulator za izračun plaće , http://www.moj-posao.net/Kalkulator-Placa/?salary=6+985&location=263&children=&supporting=&disabilityCoeff=0&type=n2g.

[3] Državni zavod za statistiku, http://www.dzs.hr/App/PXWeb/PXWebHrv/Menu.aspx?px_tableid=Robna+razmjena+s+inozemstvom%5cod+2013.+do+2014.+godine.fld%5cCountry_2_HR.px&px_language=hr&px_type=PX&px_db=Robna+razmjena+s+inozemstvom&rxid=fc9d580f-2229-4982-a72c-cdd3e96307d3.

[4] Carinska uprava: Izvješće o radu Carinske uprave za 2014.godinu, http://www.carina.hr/Dokumenti/Download.aspx?args=X6++atFcCwZYk2tHNPHiT0k2PzA++CB9.

[5] P. WINDLEY: Digital Identity (2005)

[6] Paul Hopkin, Institute of Risk Management: Fundamentals of Risk Management: Understanding, Evaluating and Implementing Effective Risk Management, Kogan Page Publishers (2014)

[7] G.WILLIAMSON, D.YIP, I.SHARONI, K.SPAULDING: Identitiy management, A Primer (2009)

[8] Project Management Institute: A Guide to the Projest Management Body of Knowledge – 3rd edition, Newtown Square (2004)

[9] K.H.ROSE: Project Quality Management, Why, What, and How, J. Ross Pub. (2005)

[10] http://interoperability-definition.info/, pristup 23.09.2015.

[11] IBM Security Intelligence: Identity Governance: IBM Security Is a Leader, Again, in the New 2015 Gartner IGA Magic Quadrant, https://securityintelligence.com/identity-governance-ibm-security-is-a-leader-again-in-the-new-2015-gartner-iga-magic-quadrant/.

[12] Linkedin: Identity & Access Management Market Overview, https://www.linkedin.com/pulse/20140604024600-12402425-identity-access-management-market-overview.

CE

International Conference on

COMPUTERS IN EDUCATION

Steering Committee

Chairs: Marina Čičin-Šain, University of Rijeka, Croatia
Ivanka Sluganović, I. Tehnička škola Tesla, Zagreb, Croatia
Jaak Henno, Tallinn University of Technology, Estonia
Hannu Jaakkola, Tampere University of Technology, Finland
Marko Čupić, University of Zagreb, Croatia
Snježana Babić, Univesity of Rijeka, Croatia

Members: Michael E. Auer, Carinthia Tech Institute, Villach, Austria
Ivan Gerlič, University of Ljubljana, Slovenia
Marija Marinović, University of Rijeka, Croatia
Željka Požgaj, University of Zagreb, Croatia
Jadranka Šunde, Defence Science and Technology Organization, Australia
Slavomir Vukmirović, University of Rijeka, Croatia

INVITED PAPER

New Informatics curriculum - Croatian tradition with world trends

Lidija Kralj
OŠ Veliki Bukovec, Udruga "Suradnici u učenju", Croatia
lidija.kralj@ucitelji.hr

Abstract - Croatian curriculum reform started in February 2015 as one of measures in Strategy for education, science and technology. During first phase of curriculum reform 55 documents were created - for all levels of pre-tertiary education, all subjects, cross curricular topics and frameworks for assessment, special education and gifted education. One of subject curriculum document is for subject Informatics whose authors are: Predrag Brođanac, Nikolina Bubica, Lidija Kralj, Zlatka Markučič, Marina Mirković, Maristela Rubić and Darka Sudarević with support: Antonela Czwyk Marić, Viktorija Hržica and Branka Vuk.

Aims of subject Informatics are to enable students to become digitally literate, to develop computational thinking, digital wisdom, critical thinking and use technology effectively, responsibly and respectfully. In order to reach that aims curriculum of Informatics is organized in four domains: Information and digital technologies, Computational thinking and programming, Digital literacy and communication and E-society.

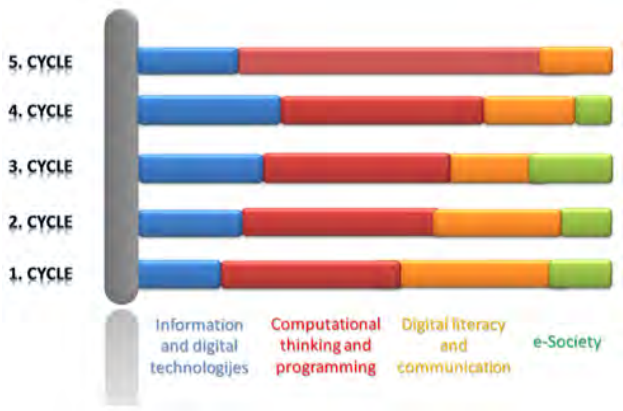
I. INTRODUCTION

Croatian curriculum reform started in February 2015 as one of measures in Strategy for education, science and technology. During first phase of curriculum reform 55 documents were created - for all levels of pre-tertiary education, all subjects, cross curricular topics and frameworks for assessment, special education and gifted education. One of subject curriculum document is for subject Informatics whose authors are: **Predrag Brođanac, Nikolina Bubica, Lidija Kralj, Zlatka Markučič, Marina Mirković, Maristela Rubić and Darka Sudarević with support: Antonela Czwyk Marić, Viktorija Hržica and Branka Vuk** [1].

Currently Informatics is elective subject in grades 5 to 8 in compulsory schools in Croatia and obligatory subject for one year in Gymnasium. Informatics' curriculum for obligatory schools is 10 years old and for secondary schools date from 1994, so new curriculum was a necessity. With new curriculum Informatics becomes elective subject in all grades of compulsory school and obligatory subject for two years in Gymnasium with 70 lessons per school year in all levels.

Generic competencies that the subject Informatics develops and promotes are:
- Creativity and innovation by creating digital content and algorithms
- Critical thinking and evaluation of technologies and information
- Problem-solving and decision-making with the help of ICT
- Information and digital literacy
- Personal and social responsibility by considering ethical issues
- Responsible and effective communication and collaboration in the digital environment
- Active citizenship, with safely, respectfully and responsibly technology use.
- Educational and professional development with use of ICT.

Aims of subject Informatics are to enable students to become digitally literate, to develop computational thinking, digital wisdom, critical thinking and use technology effectively, responsibly and respectfully. In order to reach that aims curriculum of Informatics is organized in four domains: Information and digital technologies, Computational thinking and programming, Digital literacy and communication and E-society.



Picture 1. Domains spread through all educational cycles in pre-tertiary education in Croatia

II. INFORMATION AND DIGITAL TECHNOLOGIES

The greatest power of computers lies in their ability of fast and safe storage and processing of a huge amount of information. Searching for information, retrieval and critical evaluation of information found in various sources and databases represent the main concept of today's use of digital technologies. Therefore, it is of utmost importance to understand the patterns of digital representation of various types of information such as numbers, text, sound, images or videos. Development of various strategies needed to recognise, define and explain patterns and relations within available information makes it possible to compose a new information structure. Visualisation and simulation methods used to demonstrate how a computer works in a simplified manner contribute to the development of abstract thinking. In order to develop abilities to choose and use a suitable digital technology and programmes for processing and presenting information, it is necessary to understand the fundamental concept of how computers and other devices work, to understand the patterns of data storage as well as basic characteristics and ways of transferring digital information.

III. COMPUTATIONAL THINKING AND PROGRAMMING

Development of computational thinking fosters various approaches to solving a problem, applicable on a computer. Such approach turns students not only into users of various tools but also into their designers. Logical thinking, modelling, abstracting and problem solving are skills which are developed in the process. Computational thinking is a universal skill which fosters precision and systematisation and it can be applicable in various fields of everyday life. The concept of abstraction, a fundamental concept of computational thinking, fosters the use of metacognitive skills and enables us to work on complex problems by breaking them into several simple problems. A good ICT education, based on computational thinking and creativity enables understanding and alteration of the world around us. The process of solving a problem by designing a computer programme involves standard procedures of programme development, but it also involves innovation, entrepreneurship and initiative in design and development of new models and products by applying digital technologies. The process of programming develops various skills such as self-confidence, persistence and precision in error correction, ability of communication and team work, all directed towards a certain goal.

IV. DIGITAL LITERACY AND COMMUNICATION

Digital literacy and communication involves understanding of hardware and software solutions and development of various skills necessary for online cooperation and communication. The main prerequisite for choosing the right technology and programmes and their efficient and innovative application in various fields of life is to understand their possibilities. It is necessary to develop digital literacy from an early age and during the years of schooling in order to prepare students for life and work in digital community. Various programmes for communication and cooperation enable students to develop communicational and social skills, exchanging opinions and experiences while developing the sense of tolerance, respect of differences and appreciation of other people's opinion. By continuous development and competent, critical and creative use of information and communication technology students take up an active role in creating their own positive digital footprints. By working individually or in teams, they choose appropriate digital content and programmes, they create and develop their own digital content. Positive attitude and openness towards new technological advances will provide a simpler adjustment to the future technologies.

V. E-SOCIETY

It is the right of every individual to access the digital society. It is also a vast source of possibilities to use various digital services. Digital technologies simplifies the access to education, entertainment, culture and many other services and fosters active participation in democracy. Besides, it changes methods, time and place of work and enables us to be more creative and more efficient in performing our tasks. The e-Society members should be using the Internet responsibly, safely and efficiently, whether it is reading the news or using public services such as e-Class register, e-Citizen system or e-Health. Each e-citizen should understand what personal data are and how to protect them. They should also know how to protect themselves from frauds, threats and cyberbullying and how to react to inappropriate behaviour, how to respect other people's privacy or where to look for help if they stumble across an inappropriate content or contacts. High quality, modern and innovative education based on the use of ICT will enable students to become well educated citizens of e-society who take good care of their health, digital reputation, safety and environment.

VI. LEARNING OUTCOMES

Learning outcomes are stated for every grade and for all domains. There are 8 to 14 clearly stated learning outcome for each grade and every learning outcome is described in four levels of accomplishment (satisfactory, good, very good and outstanding). Short recommendations are available for all learning outcomes so teachers could easier plan how they will accomplish stated outcome. In every domain learning outcomes are organized for spiral model of learning so in every year knowledge is building on old knowledge, level of students independent work is raising and content is getting wider or deeper.

In new curriculum learning outcomes are not stated in chronological order, they not require same amount of time and some of them can be realized in different combinations. For example outcomes from domain Digital literacy and communication easily mix with outcomes from all other domains, and some outcomes from e-Society will be realized through whole school year with all other outcomes.

Table shows learning outcomes for domain Computational thinking and programming from 1st grade (students aged 6/7) till 4th grade in Gymnasium (students aged 18/19). Similar spiral structure is recognizable in other three domains.

Table 1. Learning outcomes for Computational thinking and programming

Primary school - grades 1 - 8
Students will be able to B. 1. 1 solve a simple logical task B. 1. 2 follow and demonstrate steps in the process of solving a simple task.
B. 2. 1 analyse a sequence of instructions which perform a simple task and correct a wrong order if necessary B. 2. 2 design a sequence of instructions which uses repetition.
B. 3. 1 design a program by using a visual environment and sequence, selection and repetition and evaluate suggested solutions with teacher's assistance B. 3. 2 organise data in a useful manner.
B. 4. 1 design a program by using a visual environment and sequence, selection and repetition and input values B. 4. 2 solve more complex logical tasks with or without a computer
B. 5. 1 use a programme language to design programmes using input and output values and repetition B. 5. 2 create an algorithm for solving simple tasks, check the validity of the algorithm, detect and correct the mistakes.
B. 6. 1 design, follow and adjust programmes which contain selection and conditional repetition structures; anticipate the behaviour of simple algorithms which can be displayed in form of a diagram, words of a spoken language or a programming language B. 6. 2 explore and solve a more complex problems by decomposing them into smaller subproblems.
B. 7. 1 develop algorithms for solving various problems by using a programming language while using appropriate structures and data types B. 7. 2 apply search algorithm (sequential) when solving a problem B. 7. 3 design and create modular programmes which contain subprogrammes in the programming language B. 7. 4 use a simulation while solving a problem, not necessarily a computer problem.
B. 8. 1 identify a real life problem, create a programme to solve the particular problem, document the functioning of the programme and present it to other students B. 8. 2 identify and describe sorting algorithm, apply a sorting algorithm to solve a particular problem in a programming language B. 8. 3 identify and describe a possibility of applying recursive procedures to solve particular problems, explore further possibilities for applying recursion.
Gymnasium - grades 1 - 4
Students will be able to B. 1. 1 analyse the problem, define input and output variables and detect steps to the solution of the problem B. 1. 2 apply simple data types and explain their choice, apply various types of expressions, operations, relations and standard functions for modelling a simple problem in a particular programming language B. 1. 3 develop an algorithm and create a programme in a particular programming language for solving a problem by using the selection and conditional repetition structures.
B. 2. 1 analyse fundamental algorithms with simple data types and fundamental programme structures and apply them in the process of solving new problems B. 2. 2 detect smaller units in a particular problem, solve them and integrate them the final solution of the problem B. 2. 3 differentiate between various complex data types in a particular programming language and use functions and methods in a problem solving procedure. B. 2. 4 analyse data sorting as an important concept for solving various problems B. 2. 5 create an algorithm in collaboration with other students, implement the algorithm in a particular programming language, test it, document it and present its possibilities and limitations to other students.
B. 3. 1 visualize and present a problem graphically B. 3. 2 solve a problem by using complex types of data defined by a particular programming language B. 3. 3 solve a problem by using recursive function B. 3. 4 compare various sorting and searching algorithm B. 3. 5 evaluate algorithms according to their time complexity B. 3. 6 design an object model with related complex data structures, implement it in a particular programming language B. 3. 7 analyse traditional cryptographic algorithms and describe a fundamental idea of modern cryptographic systems B. 3. 8 define a real life problem and create a software solution going through all phases of programming, present and evaluate the software solution.
B. 4. 1 solve a problem by using abstract data structures B. 4. 2 design an application with a graphic user interface for a real life problem B. 4. 3 use modelling and simulation to present and understand natural phenomena B. 4. 4 use various programming paradigms to solve a real life problem B. 4. 5 define a real life problem and create a software solution going through all phases of programming, present and evaluate the software solution.

Table 2. Learning outcomes for Digital literacy and communication

Primary school - grades 1 - 8
C. 1. 1 use recommended programmes and digital educational resources with teacher's assistance C. 1. 2 follow simple steps to create a simple digital content with teacher's assistance.
C. 2. 1 choose a device or a programme for accomplishing simple tasks according to teacher's recommendations C. 2. 2 create digital content by combining various forms of content with teacher's assistance C. 2. 3 collaborate and communicate with familiar people in safe digital environment with teacher's assistance.
C. 3. 1 independently choose a device and a programme among other recommended programmes and devices and identify their usage C. 3. 2 design simple digital content according to instructions C. 3. 3 use safe digital environment for communication in collaborative activities C. 3. 4 identify roles and activities which require collaborative online environment.
C. 4. 1 choose an appropriate programme to solve a particular task, recommend a programme to other students and explore possibilities of similar programmes C. 4. 2 design a plan to create a digital content, design the content and evaluate the results C. 4. 3 plan and carry out simple ideas in collaborative online environment.
C. 5. 1 adapt the user interface of the operating system to his own needs, individually discover and display additional possibilities of the operating system C. 5. 2 use the opportunities the system offers for storing and organising data C. 5. 3 design a plan for creating a digital content, create a digital content, store it in e-portfolio and evaluate it. C. 5. 4 use multimedia programmes to accomplish more complex ideas in communication or collaboration environment.
C. 6. 1 Create, publish and present digital content with an online or offline programme while respecting the terms of use and privacy settings C. 6. 2 use an online storage service and appropriate programmes which support learning, research and collaboration C. 6. 3 collaborate with other students in the process of creating an online digital content.
C. 7. 1 get to know and use various platforms and programmes which he can find and install according to their needs C. 7. 2 prepare, design and publish his own website in accordance with good practices in the fields of intellectual property, think critically on good and bad features of a certain online content
C. 8. 1 retrieve, describe and compare various services for publishing online content, describe the procedure of publishing an online content C. 8. 2 individually retrieve information and software, choose appropriate sources of information and create and publish or share digital content C. 8. 3 design, develop, publish and present his work by using ICT and collaborative activities.
Gymnasium - grades 1 - 4
C. 1. 1 retrieve data and information, choose appropriate sources of information as well as edit, create and publish or share his own digital content C. 1. 2 explore the Internet services and opportunities for learning, doing business and future development C. 1. 3 collaborate and work on a project in online environment.
C. 2. 1 use a collaborative online environment to work on a joint project and explore the influence which incorporating computer systems in various devices might have on everyday life C. 2. 2 analyse programmes with regard to the licence and prerequisites for its installation C. 2. 3 compare strategies of data collection according to relevance and reliability of the source, choose a programme to effectively analyse and present data and results.
C. 3. 1 plan, create, present and evaluate a multimedia project.

VII. CONCLUSION

New Croatian curriculum for Informatics is built on good tradition that Croatia has in computer science, especially in programming where Croatian students achieve amazing results on international level. New curriculum also includes modern trends toward digital literacy for all, development of critical thinking and evaluation and also put stress on e-safety, respectfully and responsibly technology use - which was completely missing in old curriculum.

We are expecting to see new curriculum in experimental implementation in school year 2016/17 and continue to work on curriculum improvements so we do not have to wait another 22 years to get new curriculum.

REFERENCES

[1] P. Brođanac, N. Bubica, L. Kralj, Z. Markučič, M. Mirković, M. Rubić, D. Sudarević, A. Czwyk Marić, V. Hržica, B. Vuk; Nacionalni kurikulum nastavnoga predmeta Informatika, CKR, 2016. http://www.kurikulum.hr/wp-content/uploads/2016/03/Informatika.pdf [24. 4. 2016.]

PAPERS

Creativity, Communication and Collaboration: Grading with Open Badges

I. Salopek Čubrić* and G. Čubrić**
* University of Zagreb, Faculty of Textile Technology/Department of Textile Design and Management, Zagreb, Croatia
** University of Zagreb, Faculty of Textile Technology/Department of Clothing Engineering, Zagreb, Croatia
ivana.salopek@ttf.hr, goran.cubric@ttf.hr

Abstract - An important and relatively new technology for teaching and assessing soft skills is the use of *digital badges*, and more specifically - *open badges (OBs)*. Such badge is a validated indicator of accomplishment, skill, quality or interest that can be earned in various learning environments. The paper describes the use of open badges to grade soft skills in higher education. The process starts with survey that is conducted in order to define the skills that are most important for the future employment of students. It continues with the development of pedagogical approach and badging environment. The paper also gives the critical overview of the results of the survey conducted with students that were involved into the badging process. Based on the reported experiences, it is to be concluded that that the concept and technology of OBs, coupled with appropriate pedagogical approaches, could be a viable solution for teaching, assessing, grading and recognizing soft skills.

I. INTRODUCTION

Soft skills have been defined as a dynamic combination of cognitive and meta-cognitive skills, interpersonal, intellectual and practical skills [1]. They help people to adapt and behave positively, so that they can deal effectively with the challenges of their professional and everyday life. Currently, one of the most cited research study focused on soft skills has found that soft skills, especially motivation and ability to interact with others, are increasingly valued for employment [2]. Regarding the relevance of soft skills for employability, many research projects have been focused towards the investigation of this topic from different perspectives. As a result of investigations, the development of soft skills has already been established as a necessity among employers in most western societies for the purpose of sustained growth of socio-economic benefits [3]. There is a number of existing projects dealing with soft skills, either among students, among participants of VET or employed/unemployed individuals. The overall goal of those projects is to place an accent on the importance of soft skills and to develop different initiatives and models to assure the recognition of skills in focus. The list of selected projects, including the title and aim, is given in the Table 1, [4-10].

There is a number of ICT tools that teachers use in class to help their students to develop soft skills. Teachers, educational institutions and other relevant organizations use different tools and services for teaching and assessing learners' soft skills. Most of these tools are general-purpose teaching tools and services, or even ICT tools developed with wider objectives in mind.

Table 1. List of projects dealing with soft skills

Acronym	Project title	Aim related to soft skills
E-QUA	Erasmus Quality Hosting Framework	To improve the recognition of the soft skills acquired in non-formal and informal learning environments
DAISS	Job Matching Diagnostics for Assessing Soft Skills and Work Role Preferences	To adapt and transfer soft skills diagnostic and job-matching tool, by testing in a range of environments
ModES	MODernising higher Education through Soft skills accreditation	To increase mobility of students and young professionals trough accreditation of soft skills
MASS	Measuring and Assessing Soft Skills	To develop personal and social skills of young people
QA	„Querschnittsqualifikationen für Auszubildende“ Soft Skills in der Ausbildung	To determine important soft skills in teaching
NESSIE	NEtwork for Soft Skills Innovation for Employment	To share innovative methodologies in measuring and assessing soft skills for people in the workplace and those seeking to enter the labour market
CREDNET	Network for the Accreditation of the Managerial Skills	To develop a certification methodology and criteria to validate informal learning especially for the credibility and recognition of the competencies established

An important and relatively new technology that facilitates recognition and credentialing of different skills and learning achievements, and can further increase learning motivation, is that of digital badges, and more specifically Open Badges (OBs), [11]. Badges are an online record of achievements, tracking the recipient's communities of interaction that issued the badge, and the work completed to get it [12]. In other words, the badge is an image file, it contains information about who earned it, what they had to do, when the badge was issued and who issued it. This badge can then easily be added to the personal portfolio, and can be added in clusters that are shared with various groups. In addition, the badge can include a link to the evidence of meeting the criteria for obtaining a badge [13].

Digital badges have emerged on the internet to recognize and communicate achievement in a variety of learning contexts, but particularly within informal learning contexts. They are currently implemented as "micro-credentials" to convey skills acquisition and academic achievement with more transparency [14].

According to Anderson and Saubl [15], the use of digital badges is more powerful than traditional grades or scores, because they:
- represent the effectiveness of skills (as opposed to the oral tests),
- are based on the standards or criteria (not to normative assessment),
- encourage students to take control of their learning process as well as publicly available process,
- enable learning as a social process where assessments of professors and colleagues may be part of the final award of badges and finally,
- allow connections with other metadata (e.g. the national/international standards to deal with this and how).

There is a whole range of organizations that support badges like Badge Alliance (global organization standards for badges), the Mozilla Foundation, Collective shift, the MacArthur Foundation, World the Badge and so on. Currently, a number of institutions and organizations award badges, such as IBM, Microsoft, NASA, Design of America, Michigan State University, NYC Department of Education, Smithsonian American Art Museum, University of Illinois and others [16]. Examples of open badges awarded by the company Microsoft are given in the Table 2, [17].

This paper presents the results accomplished within the EU LLP research project "Grading soft skills - GRASS", at one partner institution – University of Zagreb, Faculty of Textile Technology. The whole GRASS project focuses on representing soft skills of learners in a quantitative, measurable way, so these skills can become the subject of formal validation and recognition. The overall objective of project is to create mechanisms that enable to continuously support, monitor, assess, and acknowledge development of learners' soft skills with the help of state-of-the-art ICT tools in resolving the stated challenges, thus leading the OBs concept to its full potential [18].

Table 2. Examples of Open Badges developed by Microsoft company

Badge	Title	Description
Member Designations		
	Microsoft Innovative Educator (MIE)	Educators who are passionate and committed to empowering learning through technology in their classrooms.
	Microsoft Innovative Educator Expert	Educators who are thought leaders in innovative learning with technology and want to amplify their reach and voice to all educators.
	Microsoft Innovative Educator Fellow	Educators who have demonstrated and built strong MIE Expert or Trainer achievements and can dedicate more time to work with Microsoft and educators to transform classroom learning.
School Designations & Badges		
	Microsoft Partners in Learning School	School that has educators participating in the professional learning community, Microsoft Educator Network.
	Microsoft Innovative School	A school committed to exploring, planning and developing a culture of innovation by participating in the Microsoft Innovative Schools Program online.
	Microsoft Mentor School	Demonstrated leadership in developing and implementing innovative practices and community leadership in Microsoft Innovative Schools Program.
Professional Learning Badges		
	Partners in Learning Local Forum Participant	Educators share and get their learning activity evaluated online or at in-person events.
	Partners in Learning Regional Forum Participant	By invitation, gives educators in a region the opportunity to network and share best practices.
	Microsoft Global Forum Participant	By invitation, gives educators from regional events the opportunity to network and share best practices.

II. DEVELOPMENT OF APPLICATION CASE AND GRADING SOFT SKILLS

A. Definition of soft skills in focus

At the beginning, it was important to define which soft skills are most important for the profile of the students that study at the University of Zagreb, Faculty of Textile Technology and are participants in this study. As main instruments for the definition are used questionnaire and focus group. The aim was to get the feedback from students regarding the soft skills that are considered as the most important for the future employment. In the survey participated 50 students. For the study is developed a questionnaire consisted of 14 questions [19, 20]. As the basis for the preparation of questionnaire are used the outcomes of previous projects that extracted the lists of

most important soft skills [6, 7]. In the questionnaire is included set of 30 personal, social and methodological soft skills and students are asked to grade their perception of importance of each listed skill for the success in employment and work. For the questionnaire is used bipolar 5 - level Likert scale with following responses: 1 - strongly disagree to 5 - strongly agree. In the first section of questionnaire, students were also asked to provide personal information regarding their socio-economic characteristics. The percentage of respondent's grades is given in the Table 3, [19, 20].

Regarding the definition of skills which all students consider most important for employment and work (i.e. skills with highest grades), on the top of the list is creativity (average grade 5,0).

Table 3. The percentage of respondents grades

Type of soft skill	Percentage of respondents grades (%)				Average value
	2	3	4	5	
Communication			40,91	59,09	4,59
Literal/writing ability		40,91	59,09		3,59
Verbal ability		13,64	63,64	22,73	4,09
Negotiation ability			100,00		4,00
Costumer/User orientation		81,82	18,18		3,18
Collaboration/Teamwork			100,00		4,00
Leadership			100,00		4,00
Conflict solving		77,27	18,18	4,55	3,27
Social/Contact Network		77,27	22,73		3,22
Creativity				100,00	5,00
Proactivity		81,82	18,18		3,18
Management skills			86,36	13,64	4,13
Planning and goal defining		13,64	86,36		3,86
Analysis skills			100,00		4,00
Efficient time organization		9,09	90,91		3,90
Priority setting			100,00		4,00
Decision making			100,00		4,00
Adaptability to changes			100,00		4,00
Results orientation ability			100,00		4,00
Continuous improvement	9,09	22,73	63,64	4,55	3,63
Learning skills	4,55	72,73	22,73		3,18
Commitment	4,55	68,18	27,27		3,22
Criticism		9,09	90,91		3,90
Responsibility			95,45	4,55	4,04
Professional ethics			95,45	4,55	4,04
Empathy		77,27	22,73		3,22
Tolerance to stress	18,18	54,55	27,27		3,09
Self-awareness		63,64	36,36		3,36
Life-balance		100,00			3,00
Socio-cultural adaptability		95,45	4,55		3,05

Creativity is followed by communication (average grade 4,6) and additional 10 skills that got average grade 4,0. Among the skills that are considered as least important, the life-balance, socio-cultural adaptability and tolerance to stress are pointed out. The questionnaire is followed by a focus group discussion. The group is formed from the teachers at the same institution. On the basis of conducted survey and discussion within the focus group at the Faculty, it is decided to focus at badging of the following soft skills: creativity, communication and collaboration.

B. Setting up an application case

After the definition of important soft skills, an application case for badging is developed. The definition of choosen skills, as well as needed learning outcomes defined according to guidelines proposed by the ModEs project, are given in the Table 4, [6].

Table 4. The definition and learning outcomes of the skills in the focus

CREATIVITY	
Defini-tion	The ability to contribute with new ideas to develop improvements in the products or services of the organization as well as in the activities performed in the job, with the aim of responding to the needs of evolution of the organization.
Learning Outcomes	Students are able to: 1. apply new methods for completing required tasks, eventually finding a "better way". 2. employ strategies to deal with the constantly changing professional landscape. 3. improve their ability to respond practically and creatively to problems and opportunities. 4. use frameworks and strategies for enabling a supportive environment for creativity and innovation, e. g. exchange ideas in web-forums, facilitate team processes in an atmosphere of mutual respect and support.
COMMUNICATION	
Defini-tion	The ability to transmit ideas, information and opinions clearly and convincingly both verbally and in writing, while listening and being receptive to the proposals of others.
Learning Outcomes	Students are able to: 1. communicate transparently as a result of increased congruence or confidence. 2. practice active listening. 3. present clearly and confidently to an audience. 4. monitor and manage communication patterns in a group. 5. communicate expressively and effectively in face-to-face and online settings. 6. communicate expressively and effectively in written communication
COLLABORATION (Teamwork)	
Defini-tion	The ability to build relationships of participation and cooperation with other people. It involves sharing resources and knowledge, harmonizing interests and contributing actively to reach the objectives of the organization.
Learning Outcomes	Students are able to: 1. evaluate their own performance in communication, conflict resolution, and decision-making. 2. use a wide range of team processes (brainstorming, structured organization, meetings, ...) effectively. 3. conduct a self-evaluation of themselves as team members and identify specific areas for improvement. 4. interact and work effectively with others.

For each of the defined soft skills (i.e. creativity, collaboration and communication) is clearly specified an importance statement, pedagogical approach and specific metric. Detailed schemes of work are designed to aid in the lesson planning process by outlining the entire overarching plan of a course of study.

C. Design of badging environment

For grading soft skills with Open Badges, the following badges are designed: Creative individual, Communicator, Collaborator (all three to be given after one successfully completed task), as well as Exceptionally creative individual, Great communicator and Perfect collaborator (all three to be given to students with 3 badges earned for the same skill). All badges are shown in the Table 5.

In the first phase, i.e. pilot phase, the BadgeOS is used to develop, publish and issue badges. BadgeOS is a free plugin to WordPress that alows a user to create and issue open badges. Although the experience of work in BadgeOS was positive, the intention was to incorporate the use of badges into the platforms that are widely used for e-learning at the University of Zagreb, i.e. Moodle.

D. Grading soft skills

In the grading process participated 17 MSc students attending the course Methods of Textile Making I.

Table 5. Developed Open Badges (OBs)

Skill in focus	OB Design	OB Name	Issuing when:
Creativity		Creative individual	the highest level of creativity is achieved
Communi-cation		Communicator	the highest level of communication is achieved
Collabora-tion		Collaborator	the highest level of collaboration is achieved
Creativity		Exceptionally creative individual	three “Creative Individual” OBs are earned
Communi-cation		Great communicator	three “Communicator” OBs are earned
Collabora-tion		Perfect collaborator	three “Collaborator” OBs are earned

Within the course, the professor taught the regular content, but gave additional effort to motivate students to develop their soft skills in the context of the course. Students are requested to work on different tasks either in small groups (3-4 members) or individually, and to present the outcomes in the classroom. The task is to be done in 3-4 classes (i.e. 135-180 minutes), or within a week (if the task is to be done at home). Before the start of the task, professor showed students similar tasks completed by previously enrolled students, with no further guidance. Students showed completed task in the front of all students and uploaded it to the course webpage (if applicable).

The skill level is discussed and assessed both by professor and students. The level is assessed by filling in a questionnaire with 5-point scale. The guidelines to be followed during next task are given by professor, who also issues open badges to the most successful students.

III. EVALUATION OF STUDENTS' REACTIONS

In order to receive the reactions of 17 students towards the use of Open Badges, another questionnaire is developed. It includes 10 questions (7 to be answered using a 5-point Likert scale, and 3 open-ended). The survey among students is conducted during the semester (Phase I) and at the end of semester (Phase II). The results are given in the Tables 6 and 7. The results of survey, obtained during the Phase I, indicated that students consider all three observed soft skills (creativity, communication and collaboration) very important for a textile and fashion designer (the average grades on Likert scale are: 5,00 for creativity; 4,71 for communication; 4,29 for collaboration).

Table 6. The results of students' attitude towards importance of soft skills

Statistic indicator	Importance of a single skill for textile and fashion designer		
	Creativity	Communication	Collaboration
Phase I			
Mean	5,00	4,71	4,29
Median	5,00	5,00	5,00
Mode	5,00	5,00	5,00
Minimum	5	3	3
Maximum	5	5	5
Variance	0,00	0,37	0,84
Coefficient of variation	0,00	12,97	21,32
Standard Error	0,00	0,16	0,24
Phase II			
Mean	5,00	4,93	4,71
Median	5,00	5,00	5,00
Mode	5,00	5,00	5,00
Minimum	5	4	3
Maximum	5	5	5
Variance	0,00	0,07	0,37
Coefficient of variation	0,00	0,27	12,97
Standard Error	0,00	0,07	0,16

Table 7. The results of students' attitude towards OBs

Statistic indicator	Understanding of OB concept	Preference towards gaining OBs	Importance of OBs for career
Phase I			
Mean	4,00	4,71	3,92
Median	4,5	5,0	4,0
Mode	5,0	5,0	3,0
Minimum	1	3	3
Maximum	5	5	5
Variance	1,53	0,37	0,84
Coefficient of variation	31,01	12,97	23,33
Standard Error	0,33	0,16	0,25
Phase II			
Mean	4,57	5,00	3,79
Median	5,00	5,00	4,00
Mode	5,00	5,00	4,00
Minimum	3	5	3
Maximum	5	5	5
Variance	0,42	0,00	0,34
Coefficient of variation	14,14	0,00	15,29
Standard Error	0,17	0,00	0,15

Besides those skills, they also proposed to include persistence and the ability to find a solution.

The overview of results showed that students like the idea of badging and consider it very motivating. The results also indicated that some students do not completely understand the whole concept of badging, and are not so sure that open badges will be important for their future career (average grade for the importance is 3,93).

The results of survey conducted in the Phase II, i.e. after the end of badging process are even more positive. Students gave even higher grades for the importance of communication and collaboration for a textile and fashion designer (average grades are 4,93 and 4,71). According to the results, it seems that, after this phase, the whole concept of badging is more understandable and student's motivation towards gaining OBs is even higher. Still, there is one open question - the importance of OBs for their future carrier.

Students also reported that the assessment made them think more deeply about soft skills and that they realized how soft skills are ignored in the whole educational system. A number of students started to think about enrollment into courses that will improve their soft skills.

IV. CONLUSION

This paper gives an overview of the badging concept, starting from the definition of skills that will be in the focus of badging, continuing with the development of pedagogical approach and badging system, and finally concluding with evaluation of the whole process.

Based on the experience in badging, authors believe that Open Badges should be considered as one of the best (if not the best) means of recognizing and credentialing soft skills that are presently available. The major advantage of OBs lies in the traceability and transparency of learning evidence associated with a badge as a digital credential. Furthermore, OBs could be beneficial to teaching and assessing soft skills as a good motivational mechanism, as well as a means of supporting alternative forms of assessment.

With all this in mind, authors of this paper believe that the concept and technology of Open Badges, when coupled with appropriate pedagogical approaches, could be an useful solution for teaching, assessing, grading and recognizing soft skills. This assumption is to be further verified in the GRASS project and a number of activities to be started in the following period.

AKNOWLEDGEMENT

This paper has been funded with support from the European Commission within LLP project Grading soft skills: GRASS, Project No. 543029-LLP-1-RS-KA3-KA3MP.

This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

[1] D. Haselberger, P. Oberhuemer, E. Perez, M. Cinque and F. Capasso, ″Mediating soft skills at higher education institution. Guidelines for the design of learning situations supporting soft skills achievement″, EU LLP project: Modernising higher education through soft skills (MODES), 2014.

[2] P. Moss and C. Tilly, ″Soft skills and race: An investigation of black men's employment problems″, Work and Occupations, vol. 23 (3), pp. 252-276, 1996.

[3] J. Andrews and H. Higson, ″Graduate employability, Soft skills versus hard skills business knowledge: A european study″, Higher education in Europe, vol. 33(4), pp 411-422, 2008.

[4] E-QUA, ″E-QUA Handbook″, June 2015. [Online]. Available: http://www.equa-project.eu/ [Accessed Jan 22, 2016].

[5] DAISS, ″Project report″, Sept 2013. [Online]. Available: http://daiss-project.eu/ [Accessed Jan 22, 2016].

[6] ModES, ″ModES Handbook″ [Online]. Available: http://www.modesproject.eu/en/the-modes-handbook.aspx [Accessed Jan 22, 2016].

[7] K. Kechagias, ″Teachnig and Assessing Soft Skills″, Sept 2011. [Online]. Available: http://www.mass-project.org/attachments/396_MASS%20wp4%20final%20report%20part-1.pdf [Accessed Jan 22, 2016].

[8] Forschungsinstitut Betriebliche Bildung, ″Soft Skills in der Ausbildung″, Nov 2015. [Online]. Available: http://www.f-bb.de/fileadmin/Veranstaltungen/QA-Praesentation_Soft-Skills-in-der-Ausbildung_HP.pdf [Accessed Jan 22, 2016].

[9] ADAM, ″NEtwork for Soft Skills Innovation for Employment″ 2012. [Online]. Available: http://www.adam-europe.eu/prj/9722/ project_9722_en.pdf [Accessed Jan 22, 2016].

[10] CREDNet, ″Results overview″ 2012. [Online]. Available: http://www.crednetmanager.eu/index.php/en/results [Accessed Jan 22, 2016].

[11] J. Jovanović, and V. Devedžić, ″Open badges: Novel means to motivate, scaffold and recognize Learning″, Technology, Knowledge and Learning, vol. 20(1), pp. 115-122, 2015.

[12] Open Badges, ″Badges/onboarding-issuer″, 2016. [Online]. Available: https://wiki.mozilla.org/Badges/Onboarding-issuer#A._Mozilla_Open_Badge_Infrastructure_.28OBI.29 [Accessed Jan 22, 2016].

[13] I. Glover, ″Open Badges: A visual method of recognising achievement and increasing learner motivation″, Student Engagement and Experience Journal, vol. 2(1), 2013.

[14] Acclaim: Open Badges for higher education, Pearson learning, (2013) http://www.pearsoned.com/wp-content/uploads /2013/12/ Open-Badges-for-Higher-Education.pdf

[15] D.M. Anderson, and S. Staub, "Postgraduate digital badges in higher education: Transforming advanced programs using autheentic online instruction and assessment to meet the demands of a global marketplace″, Procedia – Social and Behavioral Sciences, vol. 195, pp. 18-23, 2015.

[16] Mozilla, ″Participating issuers″ 2016. [Online]. Available: http://openbadges.org/participating-issuers/ [Accessed Jan 22, 2016].

[17] Microsoft Educator Community, ″Badges and Achievments″ 2016. [Online]. Available: http://www.educatornetwork.com/pd/ Earnbadges [Accessed Jan 22, 2016].

[18] GRASS, ″Grading Soft skills" 2016, [Online]. Available: https://sites.google.com/site/llpgrassproject/ [Accessed Jan 22, 2016].

[19] M. Stracenski Kalauz, I. Salopek Čubrić, and G. Hudec, ″Perception of soft skills in textile designers' education″, Global Journal on Humanities & Social Sciences, vol. 1 (1), pp. 33-37, 2015.

[20] I. Salopek Čubrić, M. Stracenski Kalauz, and G. Hudec, ″Soft skills' evaluation in textile engineering and design″, Proceedings of ICED 2014 - Educational Development in Changing World, pp. 1-9, 2014.

A study of students' attitudes and perceptions of digital scientific information landscape

Radovan Vrana
Department for information and communication sciences,
Faculty of humanities and social sciences, University of Zagreb, Zagreb, Croatia
rvrana@ffzg.hr

Abstract - This paper presents results of the research of students at the Faculty of humanities and social sciences in Zagreb, Croatia about their attitudes and perceptions of digital scientific information landscape. The results of the research show that use of digital scientific information resources including use of open access information resources has become a constant activity among students in their daily academic activities. The students who participated in the research demonstrated their familiarity with the key characteristics of digital information resources which suggests their increased competence for selection and use of digital scientific information resources in their education. The increase in use of scientific digital content on mobile devices has also become a steady trend which calls for adapting digital content used in student's education. Unfortunately, students are still mostly self-taught about use of digital scientific information resources which should be changed.

I. INTRODUCTION

Digital scientific information resources supported by ICT have got an exceptionally important role at universities as they create new innovative opportunities and challenges at the same time in learning and research. [1] Recent curricular and organizational changes at universities worldwide initiated by developments in ICT encouraged more intensive use of digital information resources in general and especially more intensive use of open access scientific information resources. [2] In Croatia, as in many other countries, use of open access scientific information resources has finally gained momentum as this type of information resources is now used more frequently than ever before [3],[4],[2],[5] especially in students' daily academic activities such as: creating supplement to lectures, writing term papers, preparing and carrying out students' project, writing their final theses, learning about subject areas etc. [2] To find out more details about students' attitudes and perceptions of locating and use of digital scientific information resources that create information landscape in their learning process at the Faculty of humanities and social sciences in Zagreb, Croatia, a research study was initiated and this paper presents the results of that study. By digital scientific information landscape, we mean the sum of all available scientific resources in digital format which create information landscape.

II. DIGITAL SCIENTIFIC INFORMATION RESOURCES AND THEIR USE BY STUDENTS

In general, to be able to participate actively in programs of study at universities, students have to be well acquainted with all types of digital information resources with a special focus on scientific information resources such as: e-books, e-articles, e-databases, e-journals, digital theses and dissertations, research data as well as with e-standards and patents, e-reports, e-news [8] etc. Use of digital scientific information resources is directly related to academic activities of students and their academic achievements. [8] The process of acquisition of knowledge about use of these information resources usually starts at the first year of the study. However, students are not always interested in introductory courses about use of library digital scientific information resources and services because their effort doesn't result in credits like many other courses do [7]. Special attention is given to use of open access digital scientific information resources as they "provide opportunities for students and scholars to conduct research from personal computers and workstations located in their homes, offices and laboratories" [6, 149]. As financial conditions at universities and in academic libraries around the world worsened over the years, open access scientific information resources became more attractive. Open access to scientific information resources such as scientific journal articles means they are freely available on the public internet, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself [9]. According to Ivwighreghweta and Onoriode [10], the primary advantage of open access information resources is that "the entire content is available to users everywhere regardless of affiliation with a subscribing library". Open access should be part of system of education in every country "(…) so students can become effective advocates for access to their own work, or for access to research that can aid them in becoming informed and critical researchers, consumers, and citizens" [11, 212].

III. RESEARCH

During their education students are encouraged to trust in digital information resources (with focus on scientific information resources) "with a presumption that, the wealth of information available in electronic formats can utterly

accomplish their scholastic needs as a better substitute to traditional print services" [8, 580]. As a result, students use digital information resources more frequently than they used to do. Several studies investigated students' views on digital information resources and attitudes towards digital environment in general. Study about use of digital resources done by Slater [12] contrasted use of e-books and printed books locally selected in collections of the researched university and collections available on the internet which are used more and more frequently. Another study [13] investigated acceptance and usage of open access scientific communication by postgraduate students in Tanzania, and found out that students were aware of open access; however, inadequate online scientific communication skills and slow internet connectivity were identified as key constraints for students to exploit open access. Stanton [14] investigated doctoral students in New Zealand and their perceptions of open access and found out that awareness of open access and institutional repositories was fairly low among investigated students, even in a research-focused community of doctoral students. Turan and Bayram [15] executed a study about the views and habituations of the university students in regards to their use of digital library and found out that students use the internet resources for school assignments but to digital library because it is still not among their priorities and they are not familiar with the modes of use of digital library. Previous studies in Croatia on use of digital information resources including open access to information resources as a support to academic achievement showed an increase in use of such resources [16],[17]. The study in this paper presents the newest study of students at the FHSS and their attitudes and perceptions of digital scientific information landscape. The purpose of this study is to find out what are attitudes and perceptions of students at the FHSS in relation to the digital scholarly information landscape in which they learn and do research. The objective is to collect data would provide an insight into these two selected aspects of use of information resources. The basic presumptions (hypotheses) are that digital scientific information resources have become truly integrated into the learning process of students at the FHSS and that digital scientific information resources have increased their influence on learning and research at the FHSS in comparison to the previous studies which included the students at the FHSS. A semi structured self-administered online questionnaire survey was undertaken using a convenience sample of undergraduate, graduate and postgraduate students at the FHSS. The online questionnaire was available from April 28th till May 12th 2015. An invitation to participate in the research was sent to the students' mailing list at the FHSS. Since the students are not in obligation of participating in the mailing list, additional call for participation in research was published on the main page of the Web site of the FHSS. The total response rate was 87 students,

IV. FINDINGS

Question 1. What is the year / type of your study? (N=87)

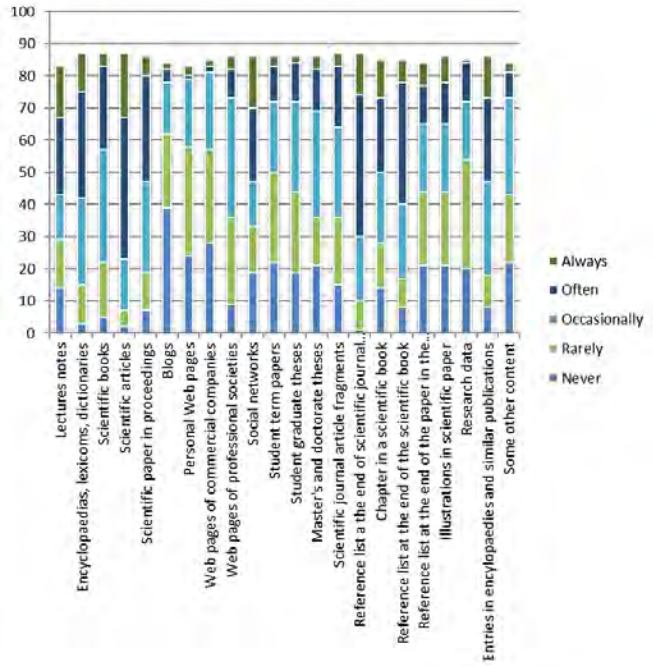
TABLE I. YEAR / TYPE OF STUDY

	%	N
First year – undergraduate study	17.2%	15
Second year – undergraduate study	10.3%	9
Third year – undergraduate study	11.5%	10
Fourth year – undergraduate study	0.0%	0
First year – graduate study	18.4%	16
Second year – graduate study	23.0%	20
Advanced university student	12.6%	11
Not a student of a reformed study program	2.3%	2
Postgraduate study	4.6%	4

Participants in this survey are students from different years of study at the FHSS but most participants in this research were students of senior years of study.

Question 2. What type of digital information resources do you use in your learning? (multiple answers)

Figure 1. Type of digital information resources used in the learning process.



In this question the respondents were given a choice of 23 different types of digital information resources. According to the results presented in Figure 1, the respondents assessed value of different information resources in a different way. Full text scientific journal articles were most frequently used information resources in this research, followed by textbooks, encyclopedias, conference papers, social networks, parts of scientific journals articles, chapter in books, Web pages of professional societies, entries in encyclopedias, lectures notes followed by other digital information resources. The choice of digital information resources is, in some cases, determined mainly by the study program.

Question 3. Where do you start your search for digital information resources you usually use in your learning process? (multiple answers) (N=87)

TABLE II. STARTING POINTS IN SEARCH FOR DIGITAL INFORMATION RESOURCES NECESSARY FOR LEARNING.

	%	N
Search engine	87.4%	76
The FHSS library	79.3%	69
Full text databases	64.4%	56
LMS Omega at the FHSS	63.2%	55
Google scholar	49.4%	43
Course Web pages	47.1%	41
FHSS digital repository	28.7%	25
Senior year students	26.4%	23
Library outside the FHSS	19.5%	17
Friend	9.2%	8
Personal Web pages of professors	4.6%	4
Bookshop	3.5%	3
Somewhere else	2.3%	2

Somewhere else: my own library (1), online encyclopedias (1). Three most popular starting points in search for digital information resources needed in the learning process are search engines such as Google, Bing or Yahoo!, the FHSS library and remote information resources to which the FHSS library is providing access i.e. full text databases. It is interesting to observe that the participants in this survey seek support of their senior colleagues less frequently than they use search engines and the FHSS library which leads to the conclusion that the students are relying more on themselves when searching content for their learning than they rely on their senior colleagues with more learning experience.

Question 4. Which of the following characteristics of digital information resources do you find important (multiple answers)? (N=87)

TABLE III. CHARACTERISTICS OF DIGITAL INFORMATION RESOURCES IMPORTANT TO STUDENTS.

	%	N
Search	97.7%	85
Storage of content on electronic media	88.5%	77
Unlimited access to the content by multiple persons	87.4%	76
Sending content to other person	62.1%	54
Print	60.9%	53
Copy / paste	58.6%	51
Highlighting of the content	31.0%	27
Some other characteristics	1.2%	1

Easy locating, storing and unlimited access to the digital information resources facilitate better learning and were, therefore, chosen by the students in this research most frequently because of their convenience in daily use. Other characteristics such as sending (sharing) content, printing, copy / paste (which is not available in all digital content) were also frequently chosen by the students in this research while highlighting of the content (a characteristic often related to printed material) were chosen less frequently.

Question 5. Mark answers which denote your reasons for use of digital information resources in learning. (N=87)

TABLE IV. REASONS FOR USE OF DIGITAL INFORMATION RESOURCES IN LEARNING

	%	N
Writing of the term paper	96.6%	84
Fast access to information	87.4%	76
Use (mostly) free of charge	77.0%	67
Writing of the homework	70.1%	61
Great quantity of available information	69.0%	60
Search for news in area of my study	62.1%	54
Searching for answers on specific questions	60.9%	53
Literature review of and scientific area	60.9%	53
Novelty of information	58.6%	51
Satisfaction of use (in general)	44.8%	39
Support to my research	42.5%	37
Content quality	42.5%	37
Student project	28.7%	25
Teaching colleagues how to use digital content in learning	5.8%	5

Answers collected in this question represent the students' view on current and future role on digital information resources in the learning process (fast access to information, free of charge, great quantity of available information, novelty of information, content quality), support to research (support to my research), current use of digital information resources in the learning process (writing of the term paper, writing of the homework, literature review of and scientific area, student project, teaching colleagues how to use digital information resources in learning and other reasons such as search for news in area of my study, searching for answers on specific questions, satisfaction of use (in general). All these reasons are valued differently in this research according to the priorities the students have in their daily activities.

Question 6. Which devices do you use for access to digital information resources used in learning? (N=87)

TABLE V. USE OF DEVICES FOR ACCESS TO DIGITAL INFORMATION RESOURCES USED IN LEARNING

	%	N
My own portable computer	85.1%	74
Computer in a public facility	83.9%	73
Smartphone	57.5%	50
My own desktop computer	32.2%	28
Family computer	21.8%	19
Tablet	16.1%	14
Feature phone	5.8%	5
Friend's computer	3.5%	3
By using some other device	3.5%	3
Television set with networking capability	1.2%	1
Computer in a public place	0.0%	0

Other devices: Amazon Kindle touch (1), Digital audio player for listening audio books (1), Kindle (1). Participants in this research use their own portable computers and smartphones (first and third most used device). Use of portable computers is followed by use of computers in a public facility (such as in library at the FHSS). Tablets as still not owned by many students in this research but they certainly carry a potential for more frequent use in the learning process. Use of book readers such as Amazon Kindle is still very modest.

Question 7. What are ways in which you learned how to use digital information resources in learning? (multiple answers) (N=87)

TABLE VI. WAYS OF LEARNING HOW TO USE DIGITAL INFORMATION RESOURCES

	%	N
By myself	92.0%	80
Introductory course for freshmen organized as a part of the study you enrolled other study you enrolled	31.0%	27
Course organized by Faculty's library	18.4%	16
In some other way	11.5%	10
Introductory course for freshmen organized for all students at the Faculty	4.6%	4
Course organized by students' club	2.3%	2
Online course	2.3%	2
Course taken outside the Faculty	1.2%	1

In some other way: I don't need a course – pointers given by professors during lectures are sufficient (1), professors gave us information during lectures (1), I asked professors and librarians (1), by word of mouth from other students (1), in courses at the department for information and communication sciences (1), lectures at the Faculty (1), friends helped me (1), course taken at another university (1), my colleagues helped me (1), in introductory courses and at the course about information literacy (1). The results show that the students participating in this research are relying mostly on themselves when they needed to learn how to use (access, organize, store and apply) digital information resources. It is encouraging to see that almost one third (31%) of the participants participated in an introductory course for freshmen organized as a part of the study they enrolled and 18.4% participated in a course organized by Faculty's library. In addition to self-learning and courses at the FHSS, the students find their own ways of getting acquainted with digital information resources such as taking pointers from professors, librarians etc.

Question 8. Are you familiar with the term open access to digital scientific information resources? (N=87)

TABLE VII. FAMILIARITY WITH TERM OPEN ACCESS

	%	N
Yes	87.4%	76
No	12.6%	11

Open access to scientific information resources is present in academic institutions worldwide for almost two decades, and it allows students and teaching staff to access (mostly) digital information resources without or with little obstacles. It also facilitates publishing of various types of scientific and professional papers written by students themselves. The results in this research hold a promise that open access will continue to be an important source of knowledge to all students at the FHSS.

Question 9. What does open access to digital scientific information resources represents to you? (N=86)

TABLE VIII. IMPORTANT OF OPEN ACCESS TO STUDENTS

	%	
(Mostly) free access to content	86.1%	74
Easier access to learning material learning	80.2%	69
Possibility of access to greater quantity of learning material	66.3%	57
Equality of access to content regardless of financial status	62.8%	54
Quick access to learning material	59.3%	51
Long term access to learning material	57.0%	49
Possibility of subject area overview	52.3%	45
Possibility of sharing information about such content with colleagues	36.1%	31
Versatility of content	36.1%	31
Possibility of making my own decision about what content to include in learning	34.9%	30
Possibility of easier organization of content in digital repository	20.9%	18

The results in this question supplement answers from the previous questions and indicate that the students have recognized and understood well known advantages of open access.

Question 10. In your opinion, should open access digital information resources constitute the majority in the reading lists of university courses? (N=86)

TABLE IX. OPEN ACCESS AS THE MAJORITY IN THE READING LISTS OF UNIVERSITY COURSES

	%	N
Yes	73.3%	63
No	11.6%	10
I don't know	15.1%	13

73.3% of the participants in this research think that open access information resources should constitute the majority in the reading lists at the FHSS. At the moment, one such digital information resource, the digital repository at the FHSS is still in the initial stage of its development but it is expected that it will provide more digital content to the students in near future.

Question 11. Please, estimate the current availability of open access information resources to you for the purpose of learning (N=86)

TABLE X. AVAILABILITY OF OPEN ACCESS INFORMATION RESOURCES FOR THE PURPOSE OF LEARNING

	%	N
Inaccessible	1.2%	1
Hardly accessible	12.8%	11
Accessible	66.3%	57
Easy accessible	12.8%	11
I can't make estimation	7.0%	6

Two thirds of students who participated in this research estimated that open access information resources are accessible while equal number of the students estimated that such information resources are equally easy and hardly accessible. The number of the students who had a positive experience (easy accessible and accessible open access information resources) with availability of open access information resources is high (total of 79.1%). While this is solely an estimation of availability of open access information resources based on the students' view of the matter, and not exact measurement, it is still valuable to know how the students see the availability of open access.

Question 12. Please, estimate your need for digital open access information resources for the purpose of learning (N=87)

TABLE XI. NEED FOR OPEN ACCESS INFORMATION RESOURCES FOR THE PURPOSE OF LEARNING

	%	N
Completely unnecessary	0.0%	0
Partially unnecessary	1.2%	1
Partially necessary	27.6%	24
Completely necessary	70.1%	61
I can't make estimation	1.2%	1

The majority of the students who participated in this research consider digital open access information resources to be completely necessary. This result could serve as a solid foundation for building of digital collections of open access scientific and learning material at the FHSS.

Question 13. Please, indicate your degree of agreement or disagreement about the following statements about digital information resources.

The total of 97.7% of the students think that open access information resources are partially or completely necessary indicating that there is very real need for a type of content that would help the students to complete their academic obligations with less problems. Due to the quantity of data, the results in this question are presented in two figures.

Figure 2. Statements about digital information resources (part 1)

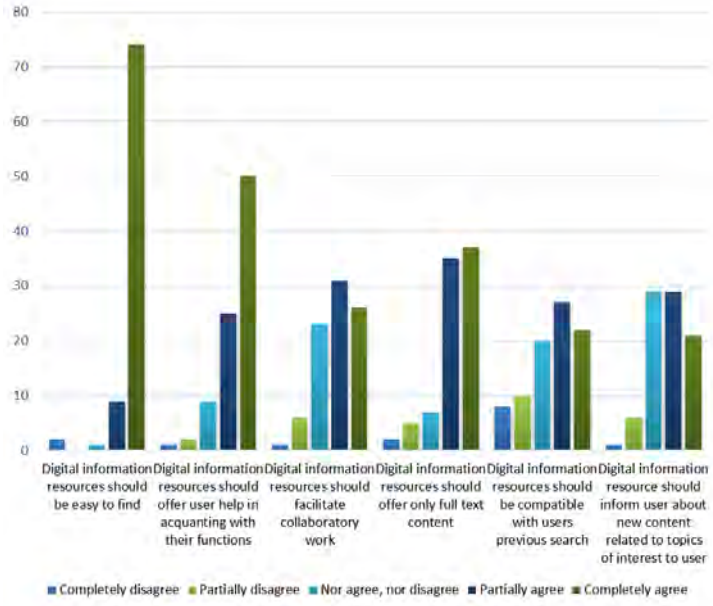
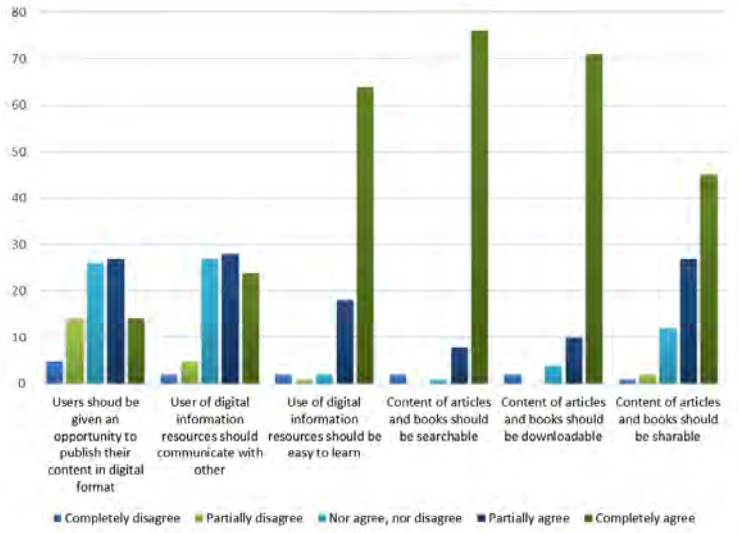


Figure 3. Statements about digital information resources (part 2)



V. CONCLUSION

Digital scientific information resources have become ubiquitous in daily activities of students at the FHSS. The research of attitudes and perceptions of digital scientific information landscape confirmed positive attitudes and familiarity of students at the FHSS with digital information resources. This research is the follow-up of previous similar research at the FHSS done by the same author. Each new research of the students' attitudes and perceptions about digital information resources show that their use has increased as well as students' competence in recognizing, selection and use of digital information resources necessary for their education. Furthermore, the results indicate a need for adapting digital content to fit the screens of new generations of mobile devices owned by students. The results confirmed both hypotheses: that digital scientific information resources have become truly integrated into the learning process of students at the FHSS and that digital scientific information resources have increased their influence on learning and research at the FHSS in comparison to the previous studies which included the students at the FHSS. Finally, the results also show the need for change in pedagogical approach of use of digital information resources as universities endorse more intensive of use of such resources.

REFERENCES

[1] K. Johnson, H. Trabelsi and T. Tin. Library support for online learners: e-resources, e-services and the human factors in Theory and practice of online learning. Athabasca: Athabasca University, 2004, pp. 349-366.

[2] R. Vrana. Open access initiative and its impact on use of digital information resources for research and teaching in higher education institutions in Croatia in MIPRO 2014 37th International Convention: Proceedings, Rijeka: MIPRO, 2014, pp. 751-756.

[3] I. Hebrang Grgić. Open Access to Scientific Information in Croatia: Increasing Research Impact of a Scientifically Peripheral Country. Saarbrücken: Lambert Academic Publishing, 2011.

[4] I. Hebrang Grgic and A. Barbaric. The future of open access in Croatia: a survey of academic and research libraries. Library Review, Vol. 60, pp. 155–160, 2011.

[5] R. Vrana. Access to digital information resources as a support to academic achievement in Central European Conference on Information and Intelligent Systems - CECIIS 2014. Varaždin: FOI, 2014, pp. 144-151.

[6] I. Fourie, I and H. Fourie. Targeting users in information provision – more than researchers, students and professionals. Library Hi Tech, Vol. 32, pp.164-172, 2014.

[7] N.K. Sheeja. Undergraduate students' perceptions of digital library: A case study. The International Information & Library Review, Vol. 42, pp. 149-153, 2010.

[8] D.K. Swain. Students' keenness on use of e-resources. The Electronic Library, Vol. 28, pp. 580-591, 2010.

[9] Budapest Open Access Initiative (BOAI). Retrieved December 27, 2015. from http://www.budapestopenaccessinitiative.org/read downloaded: December 27th 2013.

[10] O. Ivwighreghweta and O.K. Onoriode. Awareness and Use of Open Access Journals by LIS Students at the University of Ibadan, Nigeria. Library Philosophy & Practice, p. 27, 2012.

[11] S. Davis-Kahl. Engaging undergraduates in scholarly communication: Outreach, education, and advocacy. College & Research Libraries News, Vol. 73, pp. 212-222, 2012.

[12] R. Slater. E-books or print books, "big deals" or local selections—What gets more use?. Library Collections, Acquisitions, & Technical Services. Vol. 33, pp. 31–41, 2009.

[13] F.W. Dulle. Acceptance and Usage of Open Access Scholarly Communication by Postgraduate Students at the Sokoine University of Agriculture and the University of Dar es Salaam, Tanzania. African Journal of Library, Archives and Information Science. Vol. 21, pp. 17-27, 2011.

[14] K.V. Stanton. Open access theses in institutional repositories: an exploratory study of the perceptions of doctoral students. Information research. Vol. 16, 2012. Retrieved December 27, 2015. from http://eric.ed.gov/?id=EJ971947

[15] F. Turana and Ö. Bayram,. Information Access and Digital Library Use In University Students' Education: The case of Ankara University. Procedia - Social and Behavioral Sciences. Vol. 73, pp. 736–743, 2013.

[16] R. Vrana. Access to digital information resources as a support to academic achievement in Central European Conference on Information and Intelligent Systems - CECIIS 2014., Varaždin: FOI, pp. 144-151, 2014.

[17] R. Vrana. Modes of use and preservation of digital information resources as a support to the learning process at the Faculty of humanities and social sciences in Zagreb in Proceedings of the 36th International Convention MIPRO 2013, Rijeka: MIPRO, pp. 685-690, 2013.

Researcher Measured - Towards a Measurement-driven Academia

H. Jaakkola*, J. Henno**, J. Mäkelä***, K. Ahonen*,
* Tampere University of Technology, Pori Department, Pori, Finland
** Tallinn University of Technology, Tallinn, Estonia
*** University of Lapland, Rovaniemi, Finland
email: hannu.jaakkola@tut.fi; jaak.henno@ttu.ee; jukka.makela@ulapland.fi; kimmo.ahonen@tut.fi

Abstract - In many countries, higher education is financed mainly through state grants. Traditionally, universities have autonomous status and the right to carry out their self-chosen policy. The autonomy and independence of universities have formed the cornerstone of their existence. In recent years there is a rising interest in monitoring the effectiveness of academic institutions. On the one hand, this interest comes from the external desire to control the efficient use of financing and, on the other hand, from the internal desire of the university administration to direct their scant financing in a productive way. In this model, university funding is seen as an investment to implement both government and university strategies. The significance of free external financing has increased simultaneously, but even this funding is based on heavy competition between institutions. To summarize, a transition towards the entrepreneurship model has also begun in universities. Alongside academic freedom, a culture of the evaluation and measurement of quality has been created; in this process, academic values may be consigned to the background. The paper addresses the changes in universities caused by the measurement directed culture. Digitalization supports this process by providing automated tools for collecting data and using it for measurement. The handling of the topic is partially general, and partially draws on the experiences (context) of the authors. The transition towards openness and networked activities has begun. It remains to be seen what will happen to traditional academic values.

I. INTRODUCTION

The starting point of this paper relates to two questions: "How well does measuring suit academic organizations?" and "Does measurement support or control academic freedom?" Even here, at the beginning, it is easy to state that no unambiguous answer to the questions can be given. It is also questionable whether academic freedom still exists. Academic institutions are a part of society and as a result follow the same economic rules as other institutions.

Universities have three tasks prescribed by law: to provide the highest academic education, to conduct scientific research, and to provide services to the surrounding industry and society (service injunction). The University Act of Finland [3] provides the following cornerstones (translated, condensed) to the issues addressed in this paper:

- The task of universities is to promote *free research* and *education in the sciences and humanities*, and to give the highest teaching based on the research. When taking care of its tasks, the universities have to operate in *interaction* with society and have to promote the *social effectiveness* of the research results and of artistic operation. (§ 2)
- The universities have independence whereby the freedom of science, the arts, and higher education is guaranteed. The independence covers matters that belong to internal administration. (§ 3)
- The Ministry of Culture and Education (Minedu) and the University agree on the quantitative and qualitative objectives that are set for the operations of the university, and their follow-up and evaluation principles. (§ 48)

The law specifies academic *research* as a primary task of the university; *higher education* (second task), and interaction with the society / *social effectiveness* (third task) is based on its achievements.

The term "*scientific*" is often connected to the research performed by universities. The concepts "*scientific research,*" "*science,*" and "*research*" have several definitions – in this paper we will not concentrate on these. The division into two types of research – *basic* and *applied* - may be worth pointing out. *Basic (fundamental) research* is aimed at improving scientific theories for improved understanding or prediction of natural or other phenomena. *Applied research*, in turn, uses scientific theories to develop technology or techniques to intervene and alter natural or other phenomena. Basic research fuels the innovations of applied science. [8]

Traditionally, universities are centers of basic research, but to an increasing degree they also focus on work that has applied characteristics. This is – at least partially – a consequence of the external "free" research funding sourced from industry and academia–industry collaboration activities; it is an essential part of the universities' third task. Even though historically research has been the task of universities, it is good to state that high-quality independent research institutions and industry are responsible for research to an increasing extent.

A *researcher* is someone who conducts *research*. In turn, research work is the search for knowledge and a systematic investigation to establish facts. Again, without

going into details and profound definitions, some typical characteristics are worth noting. Research work produces new knowledge, it must be possible to validate, and it brings new information, subject to *public criticism*. The following attributes and features characterize the concept of science: aspiration for truth, connection to practical reality, aspiration for objectivity, demand for consistency, aim for cyclic repair of one's own results, attitude of neutrality, work based on controlled creativity, and typically creating generalizations through abstractions.

This paper focuses on publishing - *subjecting of research results to public criticism*. One of the driving forces of the researcher is to achieve high visibility and wide publicity for the work, and to increase its scientific impact. Because scientific publication is to an increasing extent based on cooperation in networks, it is natural that the members of the network utilize each other's studies as a part of their own work. This further leads to the growth of cross-referencing inside the network, in a positive manner. The measurement-related discussion of this topic is covered in the following sections of our paper; it is handled in detail in [4].

What is the link between researcher and university? *Academic freedom* is the basic value of universities in the Western world. Its interpretation varies in different countries as well as within separate disciplines and even within the universities themselves. Academic freedom means independent thinking and responsibility in the building of one's own view. The term "free research" of the University Act above refers to the autonomous role of the university in directing its research focus. It is not clear if the academic freedom also reaches the researcher level – probably just in a few cases. The universities implement their predefined strategies and conduct research in their predefined focus areas. The practical work is conducted by research teams and ultimately by individual researchers. "Measurable" results are produced by individual researchers and then collected upwards into the results of the team and finally into the results of the organization. Eventually the funding available plays the key role in directing the goals of the research work - money talks and researchers, if they want to conduct their work, must primarily fulfill the goals set by the funding elements, secondarily to use the results to meet their own preferences.

The *research problem* addressed in our paper is "What should the motivation factors be for a researcher to conduct his / her research work?". *The hypothesis* is that there exists an accelerating transfer towards a measurement- and control-oriented academic research system, in which external forces have an increasing influence in controlling and directing the work of the researchers. In this process, the universities' autonomy may be questionable and especially the motivation factors of an individual researcher will come under strong pressure from external dominance. The research questions related to the problem setting are the following:

1. What kind of evidence do we have related to the hypothesis? This will be handled by analyzing the changes which have taken place in the culture of measurement (Section 2);
2. What are the current practices in measuring scientific productivity? (Section 3);
3. What kind of infrastructure is needed for measuring? (Section 4).

We are approaching the topic in two separate papers (in this conference). This paper focuses on the role of measurement, the other paper [4] on the role of digitalization in the process. This paper is structured according to answering the research questions. The approach derives from experiences from our own context (Finland, Estonia) – but can easily be generalized and used as a benchmarking reference in other contexts. Section 5 concludes the paper and summarizes the issues addressed.

II. INTRODUCTION

The focus of this paper is on publishing as one of the important outcomes of academic research. The ideal environment for it is based on academic freedom, in which the motivation factors are subjective and based on the goals set by the individual researcher. The five-step evolution path (reporting the experiences of the authors in a career as a researcher spanning more than 40 years) from academic freedom to a "market-driven system" is illustrated in Figure 1. It covers five steps of progress in time (period of about 20 years); the illustration is based on a conceptual model (adopted from information technology and used for modeling the relevant concepts and dependencies between them).

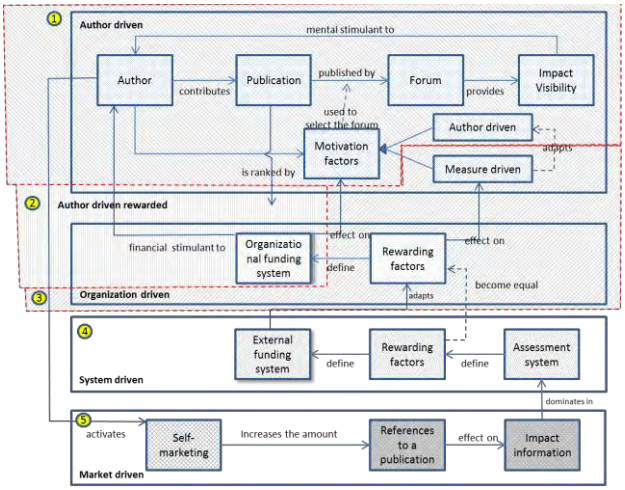


Figure 1. Five-step evolution path of academic publishing culture

In the the *author-driven system* (area ① in Figure 1) the author contributes the article to a publication, which is published by a forum providing visibility and impact to the paper. The author's motivation factors to select the forum are more or less *subjective*. The visibility and impact of the paper (in the selected forum) act as a mental stimulant to the author to continue. At this stage, publication was the researchers own "hobby" and the organization (university, research institute) did not have an interest in it.

Gradually universities started to highlight the importance of publications as one of the visibility factors of the organization; as a result, it started to reward

publication (extension ② in Figure 1; *author-driven rewarded* system); the measurements were simple, purely quantitative, and the reward was directed to the sub-organization of the author (laboratory, department, faculty, etc.). As a consequence, a *financial stimulant* was added to the motivation factors of the author – the more you publish, the more funding you receive.

Over time, the measurements became more complex and started to include some qualitative aspects; the transition – from the researcher's point of view – towards externally driven academic publication started. In the *organization-driven system* (extension ③ in Figure 1), the organization built a reward system, which became more and more complex (growing number of attributes with different reward point values – even qualitative attributes were included; the sum of the points reflects "the earnings"). At this stage, the author's motivation factors acquired a new component, *measurement-driven factors*. These tended to dominate over author-driven factors, which became gradually less and less significant). Academic publishing became a "business" for the author's sub-organizations. This step also includes a paradox: typically the budgeted reward sum was fixed inside the organization and divided by the total sum of reward points; the more points you collect, the less the value of one point.

In most cases universities and research institutes are dependent on external funding – in our case it is a question of public (government) funding, in addition to the complementary funding from industry and various other funding organizations (the latter funding instruments are not discussed here). Where previously this external funding division was based on very simple measurements (number of staff, number of students, etc.), in the transition to the next step in our evolution path, the *System-driven model* (extended by box ④ in Figure 1), the public funding model began to become more complicated. Nowadays, it also includes publication-related attributes that take into account both quantity- and quality-related aspects. In the situation where organizations are strongly dependent on external funding, their internal metrics gradually adopt the structure of the external metrics; the organization-defined metrics are finally replaced by the externally driven metrics that may conflict with the internal focus and strategies. This is the situation of today.

What is the response of the *intelligent researcher* in this case? The complex externally driven metrics system includes a lot of qualitative attributes – in publications these refer to the *scientific quality* of the papers. Because scientific quality is difficult (if not impossible) to measure, it is reflected by secondary metrics – *the impact factor of the paper* - the visibility of the paper indicated by the references to it by other authors in their papers. The impact of an individual paper can be increased by several means. The traditional approach is to publish papers in high quality forums, which are automatically followed by all researchers. The quality of a publication forum is indicated by the *impact factor of the forum* (journal, conference), i.e., how often papers published by this forum are referred to. The consequence of the *growth in impact factor* of a forum indicates a *decreasing acceptance percentage* of submitted articles. Another approach is to submit papers to such forums that are easy to reach; the fast growing amount of *open access publishing* supports this alternative. However, these forums suffer from quality problems, having very low (if any) impact factors and also questionable business models. The third approach is covered by the last step in our evolution path (extended by box ⑤ in Figure 1) – the transition towards a *market-driven system* has started. The intelligent researcher has noticed that it is possible to affect visibility. Because the scientific publication is to an increasing extent based on cooperation in networks, it is natural that the members of the network utilize each other's studies as a part of their own work. This further leads to the growth of cross-referencing inside the network, but also increases the total visibility of all the network members. This tendency, of course, should be distinguished from the contemporary culture of "likes." Social networks of scientists will have growing importance and consequences, not only on the visibility of an individual researcher but also indirectly in the changes of the impact factors of the publication forums – the more they are used, the higher the impact factor. The market-driven approach is discussed in detail in [4].

The same topic is handled by Sivula et al. [7] in their study related to the adaptation of researchers to the new funding models in academic institutions; their context is human sciences. They confirm the transition path introduced above. Before the 1990s, funding was mechanically based on the intake of students and the universities did not have the freedom to make a decision on the application of the funding independently (planned economy period). This model did not encourage universities to improve their productivity. In the 1990s, universities were given more freedom to plan the use of funding inside their university. In the middle of the 1990s, the role of publications was recognized and the Ministry of Education started to develop a calculation model to be used in the directing of financing to the universities. In the late 1990s, the model was based on the number of degrees taken. In the middle of the 2000s, the model was further developed to cover research-related indicators. The Ministry found it important to be able to compare the scientific productivity of the universities based on statistical data (KOTA database). Gradually, the importance of publishing grew as an indicator of quantity, and finally also the quality of the research outcome. The study as a whole includes results of a survey-based study among researchers – unfortunately it is beyond the scope of this paper to handle it in detail.

III. MEASUREMENT TOOLS AND ENVIRONMENTS

A. Elements of the measurement-oriented system

With reference to the discussion above, in relation to the transition towards a measurement-oriented culture and market-driven approach in publishing, we would like to point out the following aspects for further discussion:

- assessment in academic society on a general level;
- external measurement – the role of assessment between academic institutions;

- internal measurement – the role of assessment inside academic institutions.

The approach is experience-based and reflects the context of the authors – the Finnish academic system and the universities of the authors are used as a case; Estonian experiences are used to benchmark the reported aspects. The item on the list – *quality assessment in Finnish universities* is not handled in detail, because we want to concentrate on the academic publication process. *External measurement* covers the principles of the university funding system to point out the elements of external guidance in an "independent" university system. This describes the principles of financial flows from the Ministry of Education and Culture (Minedu) to the universities; there are 14 Universities and 26 Polytechnics (now called Universities of Applied Sciences) in Finland. The internal measurement section reports the current practice in one individual university (Tampere University of Technology - TUT). The aim is to concretize the contents of the steps introduced in Section 2.

The Finnish higher education system is based on public funding. Studies are free for students and external (complementary funding) is directed mainly to research activities. The share of complementary funding varies between universities; in TUT it is approximately 50%. Universities in Finland are certified by the Government; consequently, no general assessment for external certificates is needed. In higher education, the polytechnics and universities are responsible for the evaluation of their own operations and outcomes. In this they also receive support from the Higher Education Evaluation Council. The quality assessment of research is typically organized in the form of the Research Assessment Exercise (RAE), which is based on the work of independent assessment panels and focuses on improvement more than on assessment itself. Evaluation activity concerning education is concentrated into a single Education Evaluation Centre; in addition, the Academy of Finland organizes evaluations covering the quality of the science in different research sectors and implements forecasting to advance the development and impact of science.

B. *Elements of the measurement-oriented system*

The funding model applied by the Government is illustrated in Figure 2. The model as a whole includes a number of interesting items. To keep the focus of our paper we will concentrate on the publication part of the model only. As shown in Figure 2, research-based funding accounts for 34% of the total, including *13% based on publishing* (red circled area). The publications are ranked according to their impact value in five categories, having different weights (coefficients) for peer-reviewed publications: 3 for publications assessed on levels 2 and 3, 1.5 for papers on level 1, and 1 for papers on level 0; a reasonably low coefficient (0.1) is given for publications that are not assessed by the review panels. The journals, conferences, and publishers are assessed by the *Finnish Publication Forum* (JUFO) [1], which is a classification system of publication channels created by the Finnish scientific community to support the quality assessment of academic research (i.e. their impact). The categorization is based on open proposals by individual researchers and the work of evaluation panels representing various fields of science. The evaluation system is discussed in Section 4. In the current situation, over 31,000 publication forums are assessed by the panels.

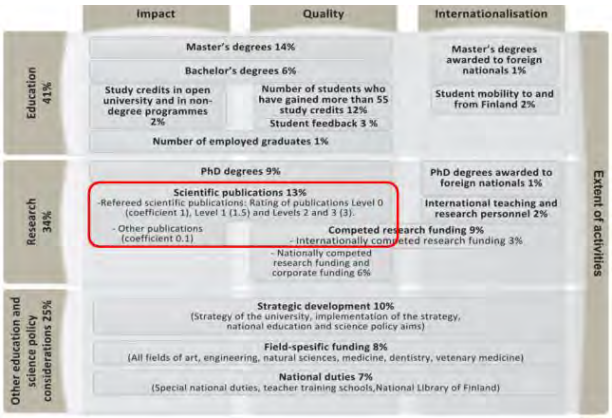


Figure 2. The university funding model [5]

In addition to the annual funding review, the long-term foresight process is used to guide educational policy – e.g., in taking into account regional needs (especially in vocational and practically oriented education). It is used primarily to analyze labor demand in different sectors and, as a result, also has direct consequences in government-level education strategy in the long term. In turn, these strategies affect the directing of the financing. Foresight looks far into the future. The policy outlined in the development plan for education and research influences the quality, quantity, and structure of education, notably target numbers, for the following years. The details of higher education provision are agreed in three-year performance agreements. Polytechnics agree with the Ministry of Education and Culture on the intake in each field of education, and universities on the target number of degrees and target intakes.

C. *Internal measurement*

The public funding flows from Government to universities according to the external model shown in Figure 2. Universities as independent bodies have the right to build an internal funding model, that may conflict / overlap / comply with the external one. However, in rational behavior, the models tend to unify in time, because it is beneficial to encourage internal processes to follow the rules of external ones (see □ in Figure 1). This means the partial loss of independence (to determine the importance of single activities) of universities although independence has been granted to the universities in the law.

The university's internal model (in the case of TUT) follows the structure of the external model, but is more complex. It is based on ranking scores representing different activities. Every sub-organization collects these scores based on their implemented activities, and the scores are summed ($SUM_{organization}$) on the university level ($GRAND_TOTAL=\sum_{i=1..n} (SUM_{organization(i)})$). Finally, the GRAND_TOTAL is used to determine the rank of a

[1] JUFO introduced: http://www.julkaisufoorumi.fi/en; Publication forum search: https://www.tsv.fi/julkaisufoorumi/haku.php?lang=en.

single score (SCORE_VALUE=FUNDING_AVAILABLE/GRAND_TOTAL; *in the TUT case slightly more than 2 EUR*. This calculation method is equal for all the sub-organizations, but does not provide predictable results in advance, because the value of the score depends on the GRAND_TOTAL. The paradox included in this procedure is that the more active the sub-organizations are, the less the score value (at least in theory it is a question of a *null-sum game*, in which the same amount of funding is available independently of the GRAND_TOTAL). It is also important to note that just a part of the funding targeted from government funding is directed to performance-based funding (in the TUT case, it is slightly less than 10% of the total funding directed to the sub-organizations). *TUT budget figures* followed this scheme in 2015: total budget 142 MEUR, Minedu 80 MEUR; 46 MEUR forwarded to the sub-organizations, performance-based funding 4.5 MEUR).

As noted above, only 5.6 % of Minedu funding is directed to the sub-organizations (laboratories, departments) as performance-based funding. The model covers 43 parameters in 11 categories. The categories follow the balanced score card (BSC) idea with elements from studies, degrees, publications, internationalization, supplementary funding, media coverage, and some other merits. The score values vary from 0.5 to 2,000, the highest score value belongs to doctoral dissertations (2,000 points using the 2 EUR estimation as discussed above means 4,000 EUR funding). In the case of *publications,* the score values are: refereed papers without JUFO ranking 40 (=80 EUR), JUFO level 1 papers 500 (1,000 EUR), and JUFO level 2 and 3 papers 750 (1,500 EUR). Compared to the external model, this is a slightly more author-friendly approach (especially in publications outside the JUFO rating).

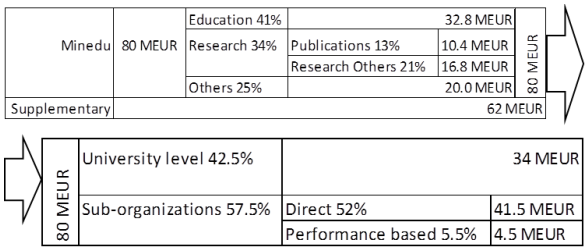


Figure 3. External and internal measurement summarized

Figure 3 summarizes the discussion related to the external and internal measurement – the TUT case is used to concretize the analysis; the upper part of the figure describes the incoming funding and lower part the division of funding inside the university. We accept that the model is greatly simplified, but it is sufficient to give a general idea on the matter. When the Ministry directs 13% of the whole sum of financing to the university on the basis of publications, the performance-based funding accounts for less than 6 % of the whole sum inside the university – including all of the more than forty components (of which publications are one). This means that the publications have higher importance to the university than to the individual sub-organization. The analysis also gives evidence to our hypothesis that the university performance-based bonus system has adopted elements from the external model, but the models are not uniform. It also validates our hypothesis that to an increasing extent, government-level (strategic) goals and political control have an effect on the goals of the universities – thus decreasing their independence.

Supplementary funding (competition-based funding from different sources for funding research projects etc.) is of high importance (50%) in the total budget of the university. In addition to direct consequences, it also has indirect leverage consequences which derive from the performance-based system. Even though this financing is used to fulfill the original goals, a part of it (in the form of overhead) remains to be used by the university beyond the control of the originator. This again includes one additional paradox. According to the statistics,[2] preparing one project application needs an average time of five working days (in simple cases; some demand much more work) and, in heavy competition, only 5-20% of applications are accepted for funding. Based on this calculation, 25-100 days of work are required for one accepted application. This is a lot of work, away from other activities, that may be partially rewarded through the performance-based funding bonuses (1,000 EUR project funding gives 1- 20 points in the calculations, depending on the level of difficulty related to the source of funding; 2-40 EUR additional performance-based funding; 100,000 EUR project funding gives 200-4,000 EUR, 1 Million gives 2,000-40,000 EUR.

IV. JUFO RATING SYSTEM

To manage the publication-based funding, a portal providing data about the impact factor classification is needed; it must also have an interface with the universities' credit reporting (ERP) system. The first classification was published in 2012. The forum covers various disciplines and the classification includes academic journals, book series, conferences, as well as book publishers. The classification rates the major foreign and domestic publication channels of all disciplines in three categories: 1 = basic level, 2 = leading level, 3 = highest level; 0=assessed, did not reach the status. The evaluation is performed by 23 discipline-specific Expert Panels composed of some 200 experts. The classification is suitable for discussions on the publication policy of research organizations and it is useful for several different evaluation, assessment, and rating purposes in the academic context. The forum has an independent position and assessments are made by independent experts. The scientific community has an opportunity to contribute and to influence the future development of the classification. The classification provides information on the impact of academic publication channels and the appreciation they enjoy within the scientific community. The objective is to encourage researchers to publish their outcomes in high-level forums. At the moment (Q1/2016), it covers more than 31,000 publication forums classified in nine main research areas. In the area of engineering and technology, it lists over 5,100 classified publication forums, over 9,000 in natural sciences (partially overlapping with the

[2] Opinion column in a major Finnish newspaper : Pekka Neittaanmäki (Professor), Valitus vie yliopistoilta uskottavuuden. Helsingin Sanomat, February 10th, 2016, page A5. In Finnish.

previous category), and in social sciences, over 8,500 publication forums – these cover the research topics of the authors of this paper.

The forum provides its users with a variety of services. The main purpose is to find information about the impact of different publication forum alternatives for one's research contribution. The search engine (forum search) serves this goal. Additional services cover information about open publishing classifications and publisher copyright policy according to Sherpa Romero model [6] and access to Directory of open access journals – DOAJ [2]. The fast progress of open access publishing has brought suspicious and questionable entrepreneurs to the market; the ethical rules of publishing have lagged behind business values. A well-known list related to this topic is the *Beall's list of predatory publishers* (list of questionable, scholarly open-access publishers) [1]. The American librarian Jeffrey Beall started to compile this list in the late 2000s; nowadays the lists cover both journals and publishers. JUFO has assessed *over 600* of these journals and flagged these with a "red warning sign" in the search results. This topic is handled in more detail in [4].

JUFO has proven to be an important tool for Finnish researchers. As discussed above, it can be used both in planning the publication policy of researcher groups and in analyzing the quality of their results. Without question, it is both useful and beneficial. Nevertheless, there are also challenges. The publication field is a dynamic one. The impact of publication forums is dynamic – impact may grow and it may also be lost. The assessment process is slow and looks backwards; however, slow progress is better than nothing.

Another challenge relates to the differences between sciences. The division of the forums assessed follows these principles: the combined publication volume of the series rated as Level 2 can be a maximum of 20 % of the aggregate publication volume (Levels 1–3) assigned to the Panel (each science field separately); the combined publication volume of the series rated as Level 3 can be a maximum of 25 percent of the aggregate Level 2 publication volume assigned to the Panel. The rest are on level 1 or 0. At the beginning, the distribution was monitored on the total level. It led to the situation in which "young sciences" had more difficulty to have their forums accepted on the higher levels of "traditional sciences". In this context, I refer to an unpublished study of Professor Markku Oivo (University of Oulu). In 2013, he compared the publication forum categories between software engineering (SE) and natural sciences (Physics, Biology) by benchmarking the publication history of these departments in 2012 in his university: level 3 – Physics 14, Biology 21, SE 2; level 2 - Physics 52, Biology 98, SE 7. He concluded "to achieve results typical for Physics and Biology, a single SE department in Finland would need to annually publish approx. 23% of all level 3 software engineering papers in the world and all SE departments would have to publish all the articles in these two forums to have the same impact as the benchmarked fields of science. Nowadays, the distributions have been harmonized between the sciences as well as possible. The reason is partially the discussion quoted above.

One more challenge, from the point of view of an individual researcher, relates to the JUFO dominance in the processes. Not all researchers are at the same stage of their research career. JUFO equalizes the researchers without taking into account their personal needs - sometimes participation in an acceptable level conference (peer reviewed, on level 0 or 1) would be more beneficial than publishing the same paper in a highly ranked journal. Young researchers are building their networks in the science society and it is important for them to meet and learn to know the researchers with a similar interest; visibility and impact in networks has high importance at the beginning.

The original purpose of JUFO was to guide researchers to publish in high-level forums. In 2015, it was embedded in the universities' funding model; this was not its original purpose. The model has connections with the neoliberalism[3], which has directed changes in Western societies since the 1990s. The judicial position of universities has changed from an independent institution to following an organizational model tightly controlled by society. Nowadays, an individual researcher, instead of an independent role, is a tool to implement the university's strategic goals. The decreased funding narrows the researchers' work opportunities in their native country and thus increases publication competition further.

V. CONCLUSIONS

The paper builds a "storyline" of the role of academic publishing, starting from traditional academic freedom, in which the researcher follows goals and purposes derived from his/her research topic, to the external, reward-driven system. The aim of rewarding has been to encourage researchers to publish their work in forums that have wide interest. This – at least partially – contradicts the goals rational to the researcher himself/herself. Academic publications are a fast growing sector with a variety of conferences, journals, and books available as alternatives; their number is incalculable. Publication has always been and will continue to be a sector of business. From the publisher point of view, it is guided by a business model, from the researcher point of view, by different research-oriented goals – these have to fit with each other. The researcher needs short publication delays and high visibility, however only (highly) ranked forums are valuable. What is a ranked forum? Visibility indicates impact and is achieved by good quality publication forums. Good quality, i.e., the impact of the forum is the consequence of a wide group of readers. The impact of a paper comes from citations – i.e., the paper is beneficial for other researchers in the field.

From the researcher point of view, the funding provided based on publications is important; funding depends on the impact level of the publication forum. Managing the impact levels of publication forums needs some technical support. The JUFO platform is one example of this. JUFO has analyzed the most commonly used publication forums and nowadays it is used both by researchers and by administrators. Its original goal was to support researchers to publish in high quality forums and

[3] https://en.wikipedia.org/wiki/Neoliberalism

to avoid publishing in the low ranked forums. The original purpose was not to be used as an instrument for funding – although it seems to work even for this purpose. Another question is the binding of publishing to financing. With reference to the numbers in Section 3.3, it is somewhat questionable whether JUFO really guides researchers to publish in the best forums (in our case organization, two JUFO 1 papers have more value than one JUFO 2-3 paper). The acceptance percentage, however, is significantly higher in the lower category.

To measure or not to measure? Free and independent research is not for measurement – it is guided by rational research-derived principles. However, we live in an open society and there is no longer much to be done about independence in research – funding instruments guide the selections. In general, measurement provides information of the process measured – in this case about the publication process of an individual researcher. Measurements are not completely beneficial, if feedback to improve the process is missing (Figure 4).

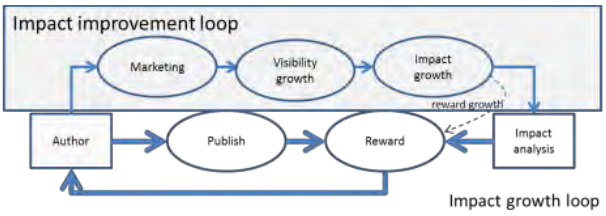


Figure 4. Publication process improvement

Figure 4 illustrates the adoption of the researcher's publication process into measurement – it can also be called "process improvement" as a result of taking into account the policy related to measurement. The lower loop (strong arrow lines) is the "impact growth loop" – measurements guide the author to select the publication forums with more impact. The upper loop (gray background) illustrates the "impact improvement loop," in which the author is active in influencing the impact itself. This will be discussed in more detail by the authors in [4].

Access to different data repositories is also a precondition for successful measurement. Universities need advanced information systems to store and manage the required data. Data supporting measurements must be collected in a systematic way. Some is automatically produced by processes, but unfortunately some must be produced manually. In addition to the ever-growing amount of administrative work, including the time needed to guarantee research funding, it takes time away from productive research. Access to global information systems is also necessary. The role of the digitalization process is important and will be handled by the authors in another paper [4]. It focuses on digitalization as a part of publishing practices, but also briefly discusses new types of learning and organization of university teaching

REFERENCES

[1] J. Beall, Potential, possible, or probable predatory scholarly open-access publishers, 2016. Scholarly Open Access. Available in https://scholarlyoa.com/publishers/. Retrieved on February 16th, 2016.

[2] DOAJ, Directory of Open Access Journal – DOAJ, 2016. Available in https://doaj.org/. Retrieved on February 16th, 2016.

[3] Finlex, Yliopistolaki 558/2009 (Universities Act), 2009. Available in http://www.finlex.fi/fi/laki/alkup/2009/20090558. Retrieved on January 29th, 2016.

[4] H. Jaakkola, H. Aramo-Immonen, J. Henno, J. Mäkelä, "The Digitalization Push in Universities", 2016. Proceedings of the 39th International Convention MIPRO 2016, Mipro and IEEE, 2016 (submitted).

[5] Minedu, Ministry of Education and Culture, University Funding Model 2015. Available in (http://www.minedu.fi/export/sites/default/OPM/Koulutus/yliopistokoulutus/hallinto_ohjaus_ja_rahoitus/liitteet/uni_funding_model_2015.pdf. Retrieved on February 9th, 2016.

[6] Sherpa Romeo, Publisher Copyright Policy, 2016. Available in http://www.sherpa.ac.uk/romeo/index.php. Retrieved on February 16th, 2016.

[7] A. Sivula, J. Suominen, M. Reunanen, "Uusien julkaisukäytänteiden omaksuminen ihmistieteissä 2000-luvulla." Kasvatus & Aika 9(3),2016, pp. 149-171. Available in http://www.kasvatus-ja-aika.fi/site/?lan=1&page_id=720. Retrieved on February 29th, 2016.

[8] Wikipedia, "Basic Research", 2016. Available in https://en.wikipedia.org/wiki/Basic_research. Retrieved on January 29th, 2016.

Use of 'Learning Analytics'

J. Henno*, H. Jaakkola** and J. Mäkelä***
* Tallinn University of Technology, Estonia
** Tampere University of Technology, Pori
*** University of Lapland, Rovaniemi, Finland
jaak@cc.ttu.ee

Abstract – Because of various reasons: cuts in financing, students massive participation in full-time work, rapidly improving technological possibilities etc. the university education is massively moving online and direct classroom contact of teachers with students is replaced with computer network mediated communication, where features of educational processes should be extracted by analyzing data created in this network-assisted education. In the following are presented some experiences and results from such analyze. Because of rather small samples we did not calculate numerical statistics, results are presented visually – this form of presentation is for humans much more understandable than tables with lot of numbers.

I. Introduction

The purpose of education is to prepare young generation for their future life, to give them skills, beliefs, habits which enable them successfully enter the productive life in future society and this goal materializes in student-professor communication. The classroom lecture has been for centuries the prevalent form of university education.

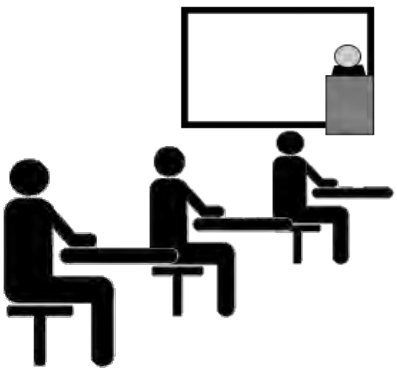


Figure 1. The classical mode of university education

But our society, moods of production and communication are changing very rapidly, especially in Information Technology (IT) and Software Engineering (SE) fields. University education in its traditional form – classroom lecture – is very expensive, especially when considering possibilities provided by modern IT technology. The economic situation has forced recently several well-established universities to severe reductions of their staff (the Helsinki University – nearly 1000 reduction [1], the The University of Copenhagen – ca 7% of staff [2]), certainly there will be more similar news. More and more students are working full-time and instead of direct classroom contact with their professors prefer accessing study materials on-line, especially strong is this tendency in IT and SE, where there are plenty of opportunities for all types of earning. This changes the whole communication structure of education, the classical face-to-face communication between students and professors is replaced with Internet-mediated communication.

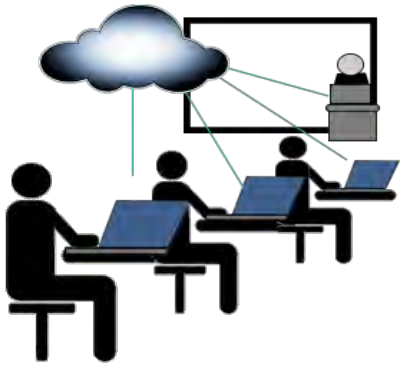


Figure 2. The now-adays mode of university education

Students are actively seeking relevant knowledge not only from of web-pages prepared by their professors, but from the whole net. The previous bi-polar educational communication model *student-professor* is nowadays three-polar: *student-Internet-professor*. And the communication is nearly 100% computer-mediated – students expect everything to be on-line.

Education is based on communication. As in every communication, we often need records to re-access some previous events in this process. The classical mode, classroom lecture based university education forces students to record lectures, make lecture notes. Results at examination depend essentially on quality and availability of lecture notes, so they often become an object of trading.

Computer-mediated educational process eliminates for students need for recording lectures; even more – importance to be present at lecture is reduced significantly, especially if student has full-time job (as most of our students have).

But in computer-mediated communication are recorded nearly every event and this creates 'big data' which could allow to analyze the whole educational process and thus improve understanding and analyze of students behavior and educational communication. This is the new field of 'Learning Analytics' (see e.g. [3]).

II. Analyzing the Educational Process

The possibilities and methods of analyze of data depend first of all on availability of data. Some e-learning organizers, e.g. Learning Management Systems and MOODLE managers do not give course author's free

access to the server. They consider a course as a fixed product which course author should deliver to them and after that author does not have any possibility to make modifications or if modifications are allowed, then author still does not have direct access to server, so that modification becomes a very complicated and time-consuming process. This makes publishing a course on their servers nearly meaningless, especially if it contains exercises, step-to-step explanations of problem solutions and together with course is available all information concerning course everyday life: feedback from student's projects (concrete remarks to concrete students, often with examples – images, video/sound clips – other students learn a lot from this feedback), what was/will be covered in lectures/exercises. Because of students very different backgrounds the presented topics, rate of advance should be often corrected; students need to know about such corrections/changes. Feedback from students creates in real, "living' course often many changes (sometimes several on the same day), news of the field etc. – all this requires, that course author should have free ftp-access to course server. Google corrects parameters (nearly 200, [4]) of its search engine sometimes every day [5] – a group of living students certainly needs the same amount of attention, care and adjustments.

Unfortunately in our market-money oriented society access to course materials should be restricted. In spite of lot of noise about "knowledge should be free" - our practice (and other cases around the web, e.g. [6]) showed, that in practice this is freedom to steal. If course materials are freely available then soon somebody starts selling them, but preparing an media-rich on-line course requires lot of work with images, sounds, scripts, styles, 3D models etc – e.g. in the course "Games and virtual environments" 2015 were used 37 types of support files (i.e other than .htm, see below). Piracy in on-line e-education is a big problem [7], [8], some sources claim, that the piracy rates in e-learning are in China 79% and in Russia 67% [9]. Thus access to course materials is restricted with Unix htaccess passwords, which students themselves create at the beginning of course. But this also allows to follow from Unix/Apache logs every student's activities with course materials – information, what is very difficult to obtain in classical classroom lecture format studies.

In the following are presented some results from analysis of digital data about the first authors courses "Games and Virtual Environments" and "Multimedia" in the Tallinn University of Technology. The main topic in both courses is communication: delivering messages to fellow community members which are noticed, do not vanish in the common and all the time increasing 'bla-bla' noise. Importance of user – generated content is rapidly increasing, this is the most important skill for survival in future in our more and more information-based society. Difference is only in message's form. games are entertaining, artistic tests for our acumen, in 'ordinary' multimedia messages is more important new information, new knowledge, but both are e.g. mixed in edutainment and 'serious' games.

Both courses are elective, aimed primarily to first-second-year Information Technology (IT) undergraduate students, but nearly every year are some participants also from other faculties and levels, e.g. graduate students. Nearly all (> 90%) of participants work and therefore even first year students have often very different skills and preferences – when one is interested in creating a presentation or game for his younger brother or sister, many want to create a commercial product (presentation or game) either for selling itself or for his/her company. Their work makes them often to travel:



Figure 3. Part of map showing accesses to homepage of the course "Games and Virtual Environments" in autumn term 2015.

Therefore both courses have been made as 'student-friendly' as possible: all course materials (tutorials about technology, showing solution of example problems) are online and before every lecture or exercise it is indicated, what will be the topic considered. Thus students can decide, whether they will participate or not. However, this mainly on-line style has not induced low passage rate - the common problem with on-line courses [10],[11].

A. The Unix logs

Analyze was started with downloading the Unix log files covering users access to materials of the courses "Games and Virtual Environments" (delivered on autumn term) and "Multimedia" (delivered on spring term) during academic years 2014, 2015; the courses were intended to Estonian students and all materials are in Estonian.

Both logs were > 7MB text files, e.g. games.log contained 51758 textlines, which each described one request from server:

```
193.40.244.196 - trevor [20/Oct/2014:18:09:00
+0300] "GET /infoinst/conference/courswre/games/
HTTP/1.1" 200 39695
...
```

Every line in the log has the same format (see e.g. [12]); log items are separated by space ' ' and their meaning (in order of appearance in the above example):

`193.40.244.196` - the IP address of the client (remote host) which made the request;

- missing (user-identifier)

`trevor` - userid; this is username, what user created for himself and does not reveal his/her real name; all user data was anonymized;

`[20/Oct/2014:18:09:00 +0300]` – date, time and timezone (+0300 – somewhere from the East: Moscow,

St.Petersburg, Minsk, Tehran – Estonian timezone is +0200)

`"GET /infoinst/conference/courswre/games/ HTTP/1.1"` - the request command (string) to open a file (here - the course homepage)

`200` – server responce (OK, standard response)

`39695` - the size of the returned object (in bytes).

The first task was to clean up this file. First were removed all unsuccessful access attempts (return code `401`), although something can be learned also from these unsuccessful attempts, e.g. some frequent unsuccessful accesses and request for .exe files (e.g. WinSCP.exe) indicated attack attempts, but seemingly there were not anything very serious.

Next were removed logs describing accesses to 'accompanying' files. According to HTTP protocol browser creates for all page objects (.css, images, sounds, script etc.) separate requests for these supportive files, thus one html-page can create in the log multiple lines. The course log for the course "Games and Virtual Environments" indicated, that besides the html documents (files with extension .htm, html) there were altogether 17996 accesses to supportive files of 37 different types (this is the reason why course author wants to keep course materials behind passwords):

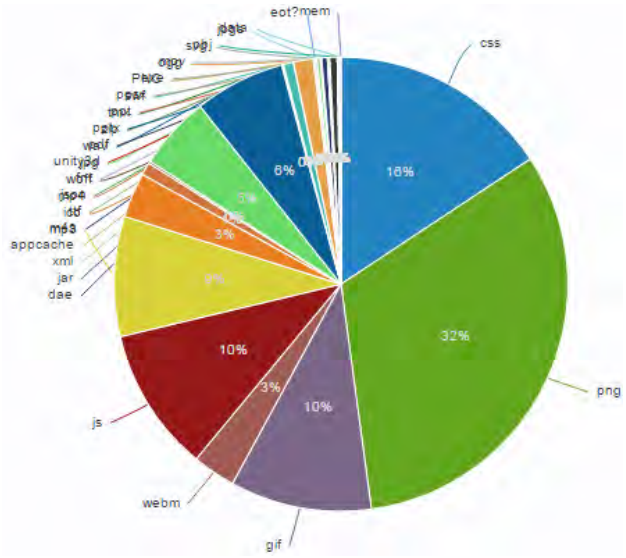


Figure 4. The 37 types of supportive files with thir relative number used in the course "Games and Virtual Environments" 2015

We assumed, that the main interest of user is in a html-document and not in particular image, sound or stylesheet, but this is not exactly true – for instance, when a student wants to look up a game's script he/she may make a request for the Javascript .js script file only or download an image file containing a spritesheet with animation; such requests indicate deeper interaction with course materials and need analyzed further. However, it is difficult to decide whether a request for e.g. a script file was done directly by student or was created by browser when it opened the main html-document, thus requests for supportive files were excluded and only requests for .htm, html files kept; finally the log was separated by dates for the two terms 2014, 2015.

Analyze and visualizing of data was performed with the D3 (Data Driven Documents, [13]) Javascript library, which is currently the most widely used tool for visualizing data in html-document. The Javascript format makes it also easy to add some visualizing scripts also to course homepage and to course materials, so that students can see e.g. their current access rate to every course document.

As always with 'big data', questions actually matter more than answers – if you do not ask right questions you will not get usefull answers [14]. The first issue we wanted to consider was the access rate – how often students actually used the course materials (were logged in). All our students work, thus it is not certain that they study regularly. However, the logs show quire regular and rather high level access to course materials and already from the very beginning of term.

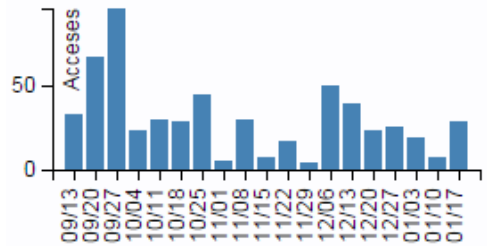


Figure 5. Weekly access rate to course "Games and Virtual Environments" materials in autumn term 2015 by course participants.

The above graph shows weekly access rate of the course "Games and Virtual Environments" materials in autumn term 2015. As seen from this chart, students started to use course materials (i.e. to learn) very actively from the very beginning of the course. For assessment all course participants had to implement two individual projects; the first was assessment of a game (game review) and the second – original game (both projects were individual – for working students group projects create often more problems than the teamwork can solve). The second project – own game – requires quite a lot of creative work and any materials are of little help, therefore the access rate become in the second half of the course lower; interest to course website grow again in the first half of December when participants were anxiously waiting for final results - assessment notes for their games but also to see what others had done – most of presented projects were accessible from the results table (shown below).

Rather similar was access also to materials of the course "Multimedia" 2015. However, we wanted to consider also how the final mark depends on access rate, thus in the next table are presented individual accesses during the first third of the course and students final mark.

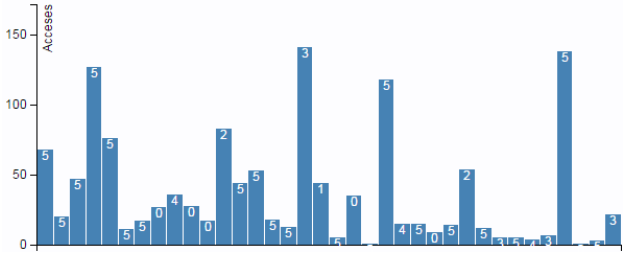


Figure 6. Cumulative individual access rate of participants of the course "Multimedia" in spring term 2015 during the first third of the course time (up to 7.03.2015), tied with final mark of the participant.

It was assumed, that personal access rate of participants is good predictor of final mark already after the first project. However (as seen from the above graph) the relation is not very clear. There may be several reasons, e.g. different preliminary level of participants and availability of many similar materials on www, although not in Estonian language (among participants were many Russian students).

For more detailed study of relation "students-files", i.e. 'who accessed what' was produced a bipartite graph showing number of accesses of students to particular files. The graph (modified from [15]) is in both directions interactive – moving the mouse pointer into a username it shows all files accessed by the student and vice versa, moving pointer into filename shows all students who accessed the file.

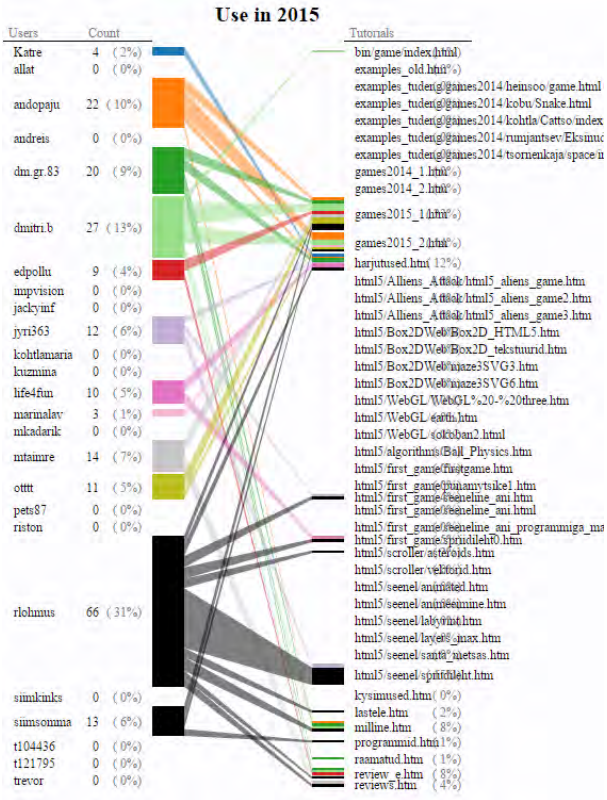


Figure 7. Bipartite graph illustarting binary relation: Students ↔ Files in course "Multimedia" 2015 shows strong grouping tendencies.

This allows to see both the activities of particular students and popularity/use of certain course materials. For instance, from this graph it was clear unpopularity of some material presentation formats, e.g. videoclips – nobody accessed them twice (most of tutorials were accessed repeatedly). This also revealed rather strong preferences of participants – some were very interested in 3D modelling, some – in data visualization, some – in SVG and/or WebGL etc.

The access rate (per week) remained more or less the same for all students and is not a reliable predictor of student's course mark. Below is a graph depicting cumulative (total) access of the course "Games and Virtual Environments" in 2015 during the first third of the course (until the first individual project return date, lighter color) and total access during the term (lighter+darker color) together with course mark.

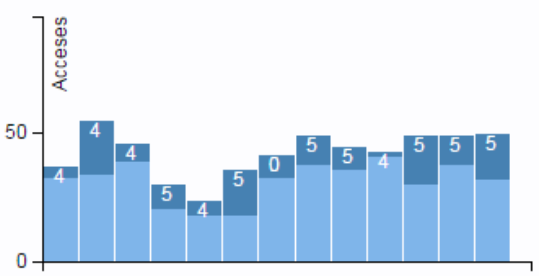


Figure 8. Access of course materials during the first third of the course (lighter color) and total access (the whole bar) together with course mark.

As seen from this picture, students who got "4" worked generally more than those who got "5" and two the most industrious students even did not pass the course (one of the reasons was cheating, presenting materials copied from www and presented as own).

The rather big rate of failures is seen also in the distribution of final marks:



Figure 9. Distribution of final course marks for the course "Multimedia" in spring term 2015 (41 participants).

The failure rate may seem rather big, but it is still much better than the general failure rate of totally on-line courses [16],[17],[18].

For the whole group access rate corresponds to final result. Surprising was a rather large number of "wasted" accesses (33%), which did not result in passing the course. This shows, that students failed not because they did not work – there were other reasons.

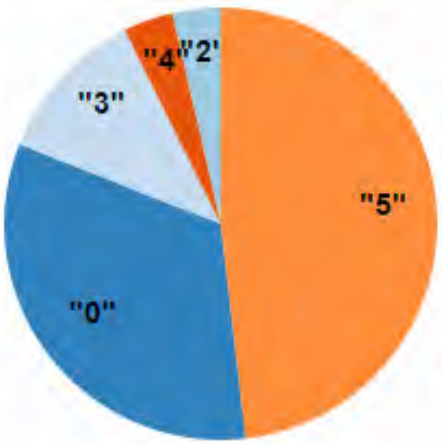


Figure 10. Final mark ("0" – did not pass) and summary accesses to course materials from this group of students.

B. *Analyzis of significance of assessment criterias*

Both considered here courses are courses which aim to develop skills of digital communication. Communication, ability to create messages which are individual, have original content, are technically high-level and beautiful, so that they do not sink in the overall swamp of 'bla-bla', the overwhelming informational noise surrounding us everywhere – this is (one of) the most important skills of our overly connected future.

Both courses were based on two individual projects. According to this aim of courses, in assessment of student projects were separately considered the following criteria:

- Message: originality and importance/weight, contemporaneity of the idea, how clearly the main idea was stated/presented, language of the presentation. From our experience, communication is for many IT students a very 'weak spot'. They believe, that their most important skill is programming. But nobody will know about their value, skills and programs if they cannot advertise those abilities and not to computers, but to people – people decide, what to buy, who to promote; we are (still) governed (mostly) by people.
- Visual appeal, beauty of the presentation;
- Interactivity, programming, links to outside documents;
- own media – own images, sounds, video;
- quality of accompanying document, where has to be explained purpose and target audience of the project, project background (who/where have been implemented similar ideas) and explained sources of used media.

Assessment of all these components was presented to students in a large table, which contained also links to presented projects, thus the whole group could assess each other's projects and see their problems and sometimes the result (the mark of the project) was changes as a result of student's opinions.

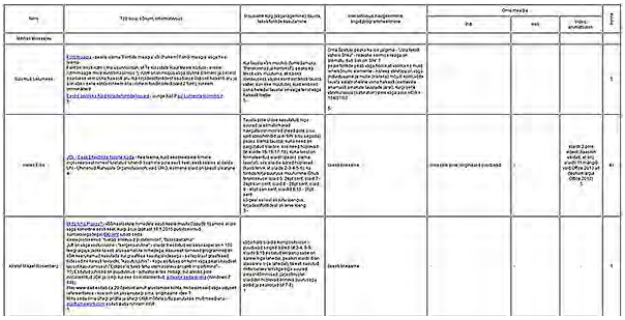


Figure 11. Beginning of the assessment table.

However, the final mark was not always arithmetical average of component assessments – e.g. a game may not have author's own images, video may not be interactive etc, so often some components did not have a mark. But assessments of components definitely were the most important part of the final mark and we wanted to estimate importance of components, their role in the final mark.

Thus we considered a linear model, where the final mark c was calculated as the average of known components $a_1, a_2, ..., a_n$ and an correction b , which was calculated with formula:

$$c*(n+1)-a_1-a_2-...-a_n \ ,$$

thus the final mark become average of component marks:

$$\frac{a_1+...+a_n+b}{n+1}=c$$

The importance of a components was set by

$$\frac{a_1}{c*(n+1)},...,\frac{a_n}{c*(n+1)},\frac{b}{c*(n+1)}$$

For instance, if a project had component marks $a_1=3$, $a_2=4$ but the final mark $c=5$, then significance of marks a_1, a_2 together was only 0.25 + 0.33 = 0.58 – they counted only a bit more than half of the final result.

To estimate overall significance (and quality of assessment), these marks were summed (by table columns) and then the similar way estimated their summary significance:

Results are depicted in the following graph.

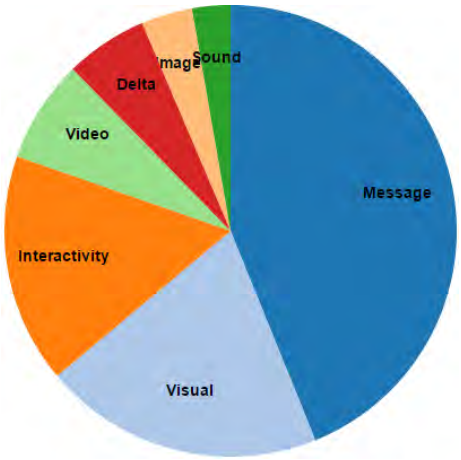


Figure 12. Avarage influence of mark components to the final mark in assessment of course "Multimedia" 2015..

As seen from this graph, the most important was the Message: 43.9%; followed Visual: 20%, Interactivity: 16.5%, Images: 3.7%, sound: 2.7%, Video: 7.4%, Delta (correction): 5.8%. As seen from here, creating a video clip is much more popular than recording a sound or even taking a still picture. The rather essential percentage of corrections to average of average of component assessments is caused by very different nature/format of projects: video clips telling a story, slideshows, (interactive) advertisements, games etc. – it is difficult to use the same set of criteria for assessment of different projects.

C. students attidude to security

In some ways working (especially working in IT field) students are more 'grown-up', understand the necessity of some requirements, e.g. strong passwords. The Internet is full of examples of break-in-s because of very week passwords, e.g. in 2015 passwords "**123456**" and "**password**" were still in the top of the most popular passwords [19]. Although there is some improvement – in 2015 a bit more complex password "**12345678"** come in third and "**12345**" was only fifth – losing two steps in the passwords popularity ranking list.

Both discussed here courses have been password-protected (passwords created by students themselves) for ten years, so now there is some data to analyze.

During these ten years has been repeatedly explained importance of strong passwords and currently passwords of our students are much better, length of most of them is 8 characters or even more. The next graph shows average entropy of student-created passwords to these courses in the last 10 years:



Figure 13. Entropy of passwords for access to courses..

The sharp increase in 2009 was caused by forcing students to use system-created passwords, but many students claimed that 'these are toooo complex", thus the system of computer-generated passwords was abondoned.

III. CONCLUSIONS

The analysis of these two datasets (two log files describing together 98840 HTTP requests) showed, that:

students started to work with course materials from the very beginning of course (Fig. 5,6,8);

access rate come (somewhat) down when they started to prepare their own individual projects (Fig. 5);

students who got better mark accessed materials more frequently (Fig. 9), but access rate and final mark are not very strongly tied (Fig. 6,8);

most students had rather strong preferences about what they were interested in (Fig. 7);

the most important criteria for individual projects is the quality/intensity of message – what the project wanted to say;

it is difficult to convey some messages (i.e. "use STRONG passwords!") if they are not supported by practical experience.

REFERENCES

[1] http://yle.fi/uutiset/helsinki_university_to_shrink_payroll_by_nearly_1000/8628873
[2] http://news.ku.dk/all_news/2016/02/massivelayoffs/
[3] http://chronicle.com/article/This-Chart-Shows-the-Promise/234573/
[4] https://www.searchenginejournal.com/200-parameters-in-google-algorithm/15457/
[5] http://backlinko.com/google-ranking-factors
[6] https://medium.com/@robconery/how-udemy-is-profiting-from-piracy-5638b929ffca#.8rmnsmfwt
[7]
[8] http://onehourprofessor.com/is-udemy-any-good/#ixzz3qjt1y4rC
[9] http://www.troyhunt.com/2015/11/the-piracy-paradox-at-udemy.html
[10] http://www.e-learningcentre.co.uk/blog/2012/04/how-big-a-problem-is-piracy-and-copying-to-e-learning/
[11] https://www.insidehighered.com/news/2013/05/10/new-study-low-mooc-completion-rates
[12] http://www.latimes.com/local/lanow/la-me-ln-online-classes-20140514-story.html
[13] https://en.wikipedia.org/wiki/Common_Log_Format
[14] *https://d3js.org/*
[15] How BuzzFeed Thinks About Data Science. http://www.buzzfeed.com/daozers/how-buzzfeed-thinks-about-data-science#.oy3a9rDBB
[16] http://bl.ocks.org/NPashaP/9796212
[17] http://hechingerreport.org/five-studies-find-online-courses-are-not-working-at-community-colleges/
[18] D.Wilson, D.Allan. Success rates of online versus traditional college students. Success rates of online versus traditional college students. http://www.aabri.com/manuscripts/11761.pdf
[19] M.McArdle. Are Online Courses Failing America? http://www.bloombergview.com/articles/2013-07-25/are-online-courses-failing-america-
[20] http://gizmodo.com/the-25-most-popular-passwords-of-2015-were-all-such-id-1753591514

Smart Immersive Education for Smart Cities

with Support via Intelligent Pedagogical Agents

M. Soliman and A. Elsaadany
Pharos University in Alexandria
Department of Computer Engineering
mohamed.soliman@pua.edu.eg, amr.sadany@pua.edu.eg

***Abstract*—the continued development of traditional education environments into smart ones is gaining tremendous speed with the rapid advances in technology. This paper explores the importance of immersive learning environment as an educational infrastructure for smart cities. It discusses the requirements for the smart city and suggests a bi-directional link between the physical environment and the immersive virtual environment. It gives example scenarios based on linking location services to a virtual world. Based on prior work on intelligent pedagogical agents, it suggests a non-conventional smart interface for smart-city services and in particular for educational needs. The proposed architecture enables the integration of various new application scenarios and potential new services such as contextual navigation support, augmented attendance, locating collaborating peers, and serving learners with special needs.**

Keywords—smart city; immersive education; smart campus; intelligent pedagogical agents; location services

I. INTRODUCTION

3D visualization and other rich abilities of an immersive environment position it as a strong candidate for smart cities. It allows users to immerse into a virtual reality like environment to enjoy remote visual collaboration. While users in such environments are able to collaborate or observe tasks, the user representation as an avatar offers supporting social abilities. Consequently, and based on its affordances [1], the environment can offer visual simulation of different aspects for smart cities from education, energy, and different aspects of urbanization. A virtual world is an example of practical implementations; such as SecondLife [2] or Open Wonderland [3]. However, there is a need for adding further intelligence to the immersive environments by different means.

While the research of immersive environments are directed for different situations that utilize specific environment affordances, few of such work have been directed towards linking potential virtual environment to smart cities models, standards, and services. That is in addition to the need to link smart education to smart city vision, proposals, and standards.

In order to advance steps towards effective use of immersive environments for smart cities in general and for its education in particular, a study of supporting aspects for smart education that is directed for a smart city suggestion should be provided. A conceptual model of how to increase the intelligence of the immersive environment for smart education is consequently needed. It would be a distinguishing factor whether a pedagogical autonomous agent, can communicate with avatars to: 1) provide the intelligent smart services, 2) observe education relevant knowledge and link it to appropriate systems, 3) use it as a pedagogical tool and as a linkage tool for other smart educational services to enable its discovery, interfacing, and intelligent navigation.

There is a strong need to tie different concepts of immersive 3D learning environments, intelligent pedagogical agents, and relevant models to be tested against the requirements for the smart city model. This reshaping serves needs of enriching the city, simulating the environment, and putting immersive virtual learning environments into use.

While the intelligent pedagogical agent concept has shown to provide smart educational services to learners represented as avatars, and acts as a central point of interaction between the environment and the user to achieve pedagogical goals, it can thus do the same in the context of smart cities. The intelligent pedagogical agent is based on supporting intelligent agent framework that realizes the autonomy of the agent based on defined goals that are relevant to the city particular model. It is supported by the smart city model components. The pedagogical agent provides a smart multi-modal interface that allows text-based communication with users. It also provides verbal communication and interface to specific feedback devices to special types of users (citizens).

The objective is thus to introduce a smart educational system that meets the needs of the learners in a smart city. The system will present an interactive immersive learning environment as well as combining a positioning system that provides services based on the location of the learner citizen in the physical smart city.

While an e-learning system is considered as an essential requirement in modern universities, it becomes even more demanded within smart cities in general. Basically, e-learning relies on a software system that allows access to all the materials for the educational process and makes them electronically available to all the students whenever they need and wherever they are. Hence, the e-learning system is expected to resolve several problems that face the traditional academic learning systems. These problems are related to time and place of the classes, the materials of the lectures as well as issues related to lab experiments or the special needs of some students.

The electronic modernization of educational institutions in smart cities must take advantage of the location-based data in order to provide suitable services to the users of the system. In this work, we are considering the available positioning

technologies and methods needed to be linked to the immersive environment offering new possibilities. For example, educational institutions traditionally consist of large areas and many buildings with wide area in-between. As such, there are difficulties that face new students in finding the places, educational materials, services, and peers. A location-based service enables finding places within the space and specifying the name and location to students and the guests, and then streamline personalized educational informational to the e-learning environment accordingly. The system can provide a map of places and link the students to learning activities based on their location, through the immersive environment. With these location abilities, it becomes possible to locate prior virtual peers physically, connect them to educational resources and services on campus and support their learning accordingly.

Combing all the above concepts into a comprehensive smart city educational environment aims at providing quality educational services and empowering the smart city infrastructure while saving cost on the long term. Scalability must thus be considered in the proposed system and the immersive environment must also provide a proper pedagogical-aware and smart-city-aware user interface through an effective utilization of intelligent pedagogical agents.

This paper is organized as follows. Section 2 provides introductory background of smart cities, immersive environments, and intelligent pedagogical agents. Section 3 discusses the immersive environment as a candidate for smart cities by giving the requirements and proposing a conceptual view of the integration of the concepts. Section 4 focuses on the location aspect that helps in linking the physical world to the virtual world. To further illustrate the impact of integrating location services, Section 5 details location-dependent scenarios that are realizable in an immersive environment that serves the smart city concept. Section 6 concludes and provides future work and limitations.

II. BACKGROUND

A logical expansion to the noticed proliferation of ICT in the different areas of life is the smart city view which is recently arising. While various research and industrial work aims at establishing the definition of the smart city and its goals, all agree on the importance of integration and utilization of the current digital world innovations to serve life needs and making it smarter. Six aspects that form the concept of the smart city are identified in [4]: 1) smart economy, 2) smart governance, 3) smart people, 4) smart mobility, 5) smart environment, and 6) smart living; and suggest ICT as a key enabler element. Under each category, various application types are needed. Example application types include the support of smart grid and green energy, transport and mobility, and more. Smart Cities as a system of systems mandate heavy reliance on standards in order to achieve the high degree of interoperability demanded. These systems come from different domains of systems that are of interest to the smart city concept. The main idea is to combine these systems into building the services required by the smart city. The main driving force behind the integration of the corresponding applications is the availability of a pervasive networking infrastructure.

A special dedicated IEEE standard is under development for the architectural framework of the Internet of Things (IoT), namely, the IEEE P2413 standard. IoT is basically a system of interconnected people and physical objects along with ICT technology to build, operate, and manage the physical world via smart networking, pervasive data collection, predictive analytics, and deep optimization [5]. The IoT standard provides a reference model, defines architectural building blocks, and provides development mechanisms for the relevant systems.

The educational aspect has been almost ignored in smart cities vision despite its importance as a pillar for economic development and growth, in addition to its value for the citizen growth. While urbanization is found to be a main focus in smart city development, a big population of the urban city is engaged in either formal or informal learning activities. A considerable space is allocated to learning activities being conducted on campuses, schools, and community colleges. The educational aspects of the physical space should also be considered in the activities of urban design and in the simulation of those spaces. E-learning systems, given their continuing rise, should be considered as a major component for the smart city infrastructure.

A. Immersive Virtual Learning Environments

Immersive virtual environments are distinguished by different properties that support their candidacy for smart cities. They offer rich 3D visualization combined with the ability of immersion by the user who is represented as an avatar. With this, immersion into a virtual reality like environment adds new aspects of learning in 3D with several visualization and simulation supported application types. The immersive environment, as a 3D virtual space, thus supports urbanization and the simulation of space.

Immersive environments are remotely accessible. Being a multi-user system, and with the above features, they support visual collaboration as avatar users can see others in the immersive environment in addition to having verbal or text-based chat. Studies revealed specific immersive environments benefits to learning[1] which are grouped into spatial knowledge representation, experiential learning, engagement, contextual learning, and collaborative learning [1]. Those benefits encourage the utilization for several educational relevant activities day after day. As such, a virtual world implementation can support creating those features practically. An example virtual world is Open Wonderland which is an open source environment that can act as an experimentation platform. On the other hand, SecondLife is an online accessible virtual world environment with a high volume of users. Several of those virtual world environments originated from Massive Multi-Player Online Games (MMOG) to support the high volume of users. Consequently, universities created educational events and virtual classrooms in virtual worlds, such the river city project [6]. With the 3D simulation abilities, it was possible to simulate various physics experiments for education purposes in the Open Wonderland virtual world [7].

As the immersive learning environment proves to provide learning values to education, its potential impact for education in smart cities is essential to be discovered. That can also lead to strengthening the use of the environment resulting from this

inspection. It is expected that the learner user of the immersive environment is a citizen of the smart city; i.e. not only a citizen receiving smart city services but also a citizen contributing to this environment.

B. *Intelligent Pedagogical Agents in Immersive Environments*

Intelligent Pedagogical Agents (IPA) can be viewed as autonomous proactive entities that can support the learner in the immersive environment [8]. The user, through his avatar, can directly interact with the IPA, which can guide him, answer questions, and more. This kind of support can be performed intelligently by means of AI tactics such as intelligent agents, [9]-[11]. For example, the learning goals can be provided to the IPA supporting intelligent agent mechanism that consequently guides the learner to achieving his goals. Another type of intelligence facilitates a non-conventional user interaction that with a multi-modal interface. The IPA can thus have an understanding of the environment, its resources, pedagogical goals, etc. and reason it to provide step-wise guidance through learning tasks. The IPA simplifies the interactions through its interface to act as a central point of interaction with the environment as shown in Figure 1. Furthermore, the IPA should facilitate collaborative learning [12].

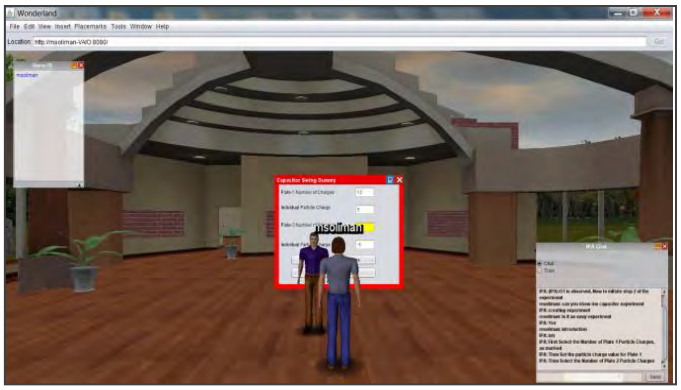


Fig. 1. A Pedagogical agent providing a tutorial in OpenWonderland [8].

These archetypes of interaction the IPA provides are intended to be recorded through different learner and learning-centric models. Assessment, for example, is needed to monitor learner abilities versus progress. The acquisition of such data should be of interest to smart city governance in different aspects. The smart city model suggests integrating vast resources and making an intelligent proposition of services to the user. The IPA as a central point of interaction with the learner-citizen can extend its abilities and integrate with other city models. The trend of employing learning analytics can be also supported in the immersive environment when the IPA monitors and tracks learner activities.

III. IMMERSIVE EDUCATION FOR SMART CITIES

Urbanization and smart cities suggest the use of appropriate technology to simulate the space aspect and the citizen behavioral patterns in relation. Several studies suggest the use of latest developments of virtual realities and immersive environments to support the design of the smart cities given its abilities. Utilizing immersive educational environments consequently suggests inspecting its candidacy for the smart city requirements and infrastructure described below.

A. *Requirements*

A proposed incorporation of immersive environments to serve the smart city concepts imposes the following requirements:

- Strong interoperability no matter what smart city system is being used by the learner citizen. The educational system must provide the students with the same services across the platform of the smart city. A customized adaptable application interface should also be available.
- Mapping between the physical space and virtual space with bidirectional support services. I.e., it is required to answer the question of how the virtual environment supports the physical space and vice versa in a way to improve the smartness of both. Location aware systems should be linked to the immersive environment via utilization of positioning technologies and paradigms in order to serve the learning purposes. When such integration is required for e-learning systems from the physical environment, the use of a 3D virtual environment is thus expected.
- The ability to utilize other smart city infrastructure and its benefits for education. An example is the e-health services that the smart city is expected to provide to the citizens. Those services are also provided to learners in the actual campus with smart services relevant to providing health-aware learning services such as the case with learning for disabilities. Other examples are operating under risk and in cases of safety of learners, hazardous transportation aspects, etc.
- The use of the immersive environment for simulating city infrastructure such as urban design, simulation of the energy grid, and more, can be an input to educating city citizens about the city itself. Demonstrations or illustrations can be simulated and then presented on the user device.
- The ability to capture high volume of interactions, summarize them for proper use, then use them intelligently, and make them available for other management or integration layers (Governance or planning, etc.).
- Scalability is inherent in the need to serve 1) a large base of learning community and 2) a large base of virtual 3D space. Users should be able to co-design the space, which is already possible in various virtual world implementations. This requirement also affects the back-end design of the systems, e.g., the proper database sizing must be taken into the design considerations.
- Intelligence support, i.e., various tactics such as AI should be utilized to facilitate the learning process to be smarter. This is of particular importance to provide highly intelligent services to students.

- Support the different requirements for users with special needs. The beauty of e-services is that it can be provided remotely wherever the user is located; thus attending lectures or interacting with instructors are available to those users on their own devices. In addition, immersive environments can provide virtual support services to aid people with special needs in the physical environment.

B. *Towards Integration: a Conceptual Model*

Smart cities require applications that can provide smart services to the users. These applications range from smart energy to smart government. Figure 2 shows a conceptual model incorporating different smart city and immersive environment concepts. Of particular interest to us is the smart education which we believe it is not only a service given to the students, but also the building block in producing innovation into the industry of the smart cities. The other smart services given in the architecture such as e-health are also important and as a matter of fact, are tied to education and are considered as facilitation services.

In order for the smart applications to perform, there must be some type of networking and sensing infrastructure in the city in addition to a supporting ICT platform. The smart application is hosted on these platforms and requires the input signals of the sensor in order to make the correct decision for the user. These signals are transmitted via the communication networks that are available in the smart city.

The second part of the conceptual architecture is the immersive environment which can provide a plethora of virtual services not only to students but also the rest of the crowd in the city. However, in this paper, we focus on its use in the smart educational environment and as such we use the main components which are the virtual campus model, the learner model and the pedagogical model. These models support the educational services in the architecture such as the tutoring and assessment services, the intelligent navigation support, and the smart online labs and visual simulations.

The Intelligent Pedagogical Agent is the key player that guides and monitors the user into the different application domains. For example, when a user is looking for the direction to a certain lab or classroom, the agent can traverse the campus map on the user device as illustration to the final destination. It acts as a central point of interaction with the user while providing the various educational services. It is thus equipped with intelligent decision support, via an intelligent agent paradigm [11], and multi-modal communication abilities to foster easier communication by different means [8].

Once the immersive environment is well incorporated, it becomes a part of the smart city infrastructure providing immersive education services. It should be noted that the link between the physical space and the virtual space utilizing the smart city technological infrastructure is a distinguishing aspect of potential services that can be given to the smart city citizen. As location services play an important role in linking the physical space to the virtual spaces, it is focused in the following section.

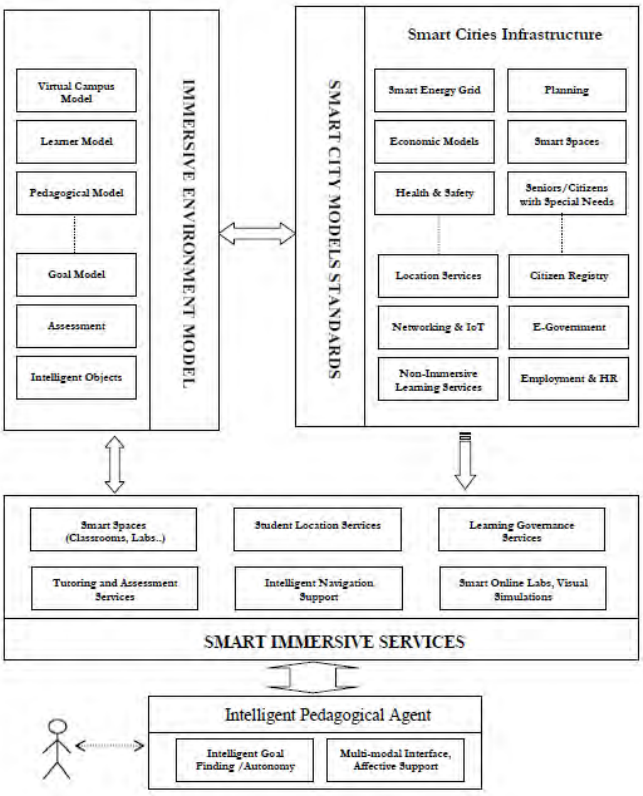


Fig. 2. A Conceptual depiction towards incorporating Immersive Environments in Smart Cities.

IV. Architecture For Location-Smart Education

The main idea here is to provide SMART Educational System Testbed (SMARTEST) that enables the e-services requirements which combines the location-based services and the immersive mapping in a smart campus. Figure 3 shows the location-based services subsystem architecture.

Basically, this subsystem consists of two main parts which are front-end and back-end. The front-end is the user application itself that can be installed on the users' mobile or the immersive interface. It is a graphical interface consisting of some buttons, dropdown menus, text fields, etc. to interact with the user (e.g., on Android mobile device). The back end is hidden from the user and installed on a server to be connected by all mobiles that have installed the user application interface. The application server should be available 24/7 to respond to any client request.

The proposed location-based subsystem has many features that make the educational environment friendly in terms of finding places or people and in solving many related problems. In educational environments that have a large area and many buildings, there are many difficulties that face the students, staff members, and visitors (for example, finding the places and hotspots for the new students and visitors). New students may miss lectures because of not knowing the location or the time of the lectures.

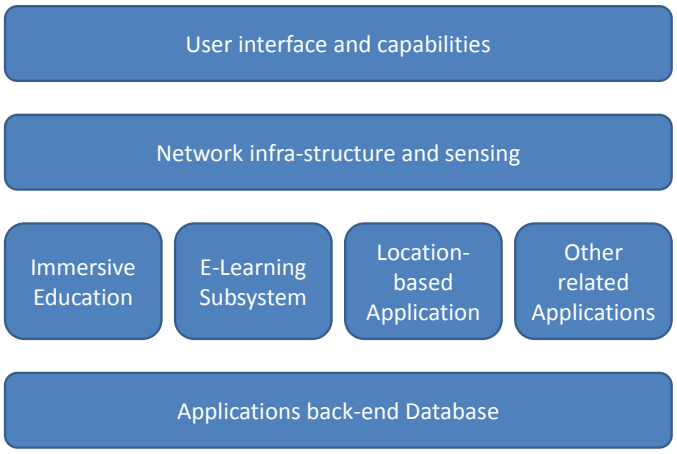


Fig. 3. Location-based services subsystem architecture.

The subsystem is accessible from a mobile application that can track users and find places inside the university campus. The user can see all the buildings inside the university and asks for the directions. The user might be a staff member, a student or a guest. Staff members and students have the full functionality of the system, but guests can just see the buildings and ask for directions. An indoor positioning feature of this system can be used for taking the attendance of the lectures automatically by locating the students inside the lecture hall. This feature saves the time needed for taking the attendance manually.

The positioning system can also let the users find someone on the map and locate him/her. Each user has a profile and a friend list, so the user can track a person if this person in his friends' list (and his location is turned on). This feature is also useful for students to find their professor and check their status (and if they are available or busy). This option is not only useful for students and professor but also for locating other staff members. The subsystem can be used for sending a notification to all or a particular class of staff members and students to notify them of an event or advertisement. This feature can be used instead of posting paper flyers.

The implementation of the location subsystem relies on the use of the already available algorithms for determining the position of the user. These algorithms utilize techniques that rely on the available wireless signals at the location of the user's device [13]-[17].

The main aim of a supporting learning system is to save travel costs and instructional time provided that the system provides quality education and improves the performance of the education for both students and instructors [18]. In the smart education system, the design and implementation must not only provide the normal educational services but also incorporate these services into the overall smart city. For example, the user should get the customized services based on his location which is yet provided by another application in the platform.

The smart system must be designed and implemented such that it provides the user interface based on the user device characteristics. The user with a mobile device screen might get a different interface than a user with a laptop screen [19]. Moreover, for certain applications the user device should allow immersive navigation to take advantage of the integration with location supported applications. The combined smart systems must be tested to prove the new design concepts and the overall integration. The back-end infrastructure must not only support the subsystems integration but also allow the future usage of the subsystems in smart city educational and research institutions.

V. Location-Dependent Scenarios

A. Physical Navigation Support

With the ability to obtain the student device location, it is possible to position the user in the virtual campus offering him the exact location he is in and consequently directions to his target location. The supporting pedagogical agent can also obtain the expected learning activity location, based on the stored student schedule and offer directions to guide the student to his lecture, see Figure 4.

Given the potential use of a positioning system, the physical location coordinates are updated to avatar coordinates in the virtual 3D campus to update the scene location and to provide the guidance dynamically.

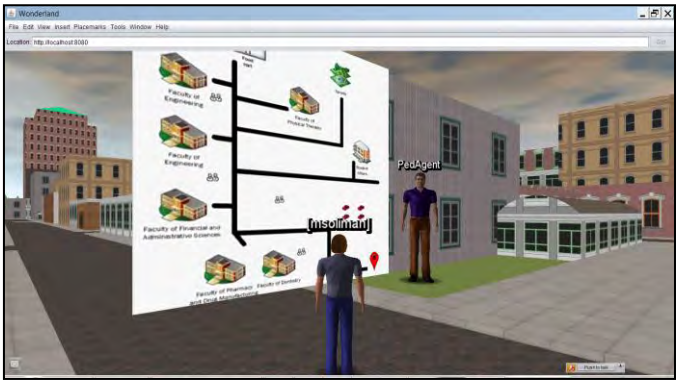


Fig. 4. A Pedagogical agent instantiates a campus map, based on the student physical location, to direct a student to his leacture. An illustration from the Open Wonderland Virtual World.

B. Virtual Navigation Support

The ability to have a 3D model of the campus allows the navigation to be also virtual, which supports the leaner needs through his avatar to navigate virtually to the destination to test the path before going. This is in support to new students who experience the campus for the first time or would like to know the route to a particular classroom. A possible IPA-to-learner interaction is illustrated in Figure 5 as the IPA proposes to the user the ability to navigate virtually first.

This text based interaction is supported through an Artificial Intelligence Markup Language (AIML) [5][20]-[21] to be linked to the virtual campus location model in relation to the student physical location. In this scenario, the student physical coordinates are released from his immersive location coordinates allowing to use the immersive environment teleportation ability [3].

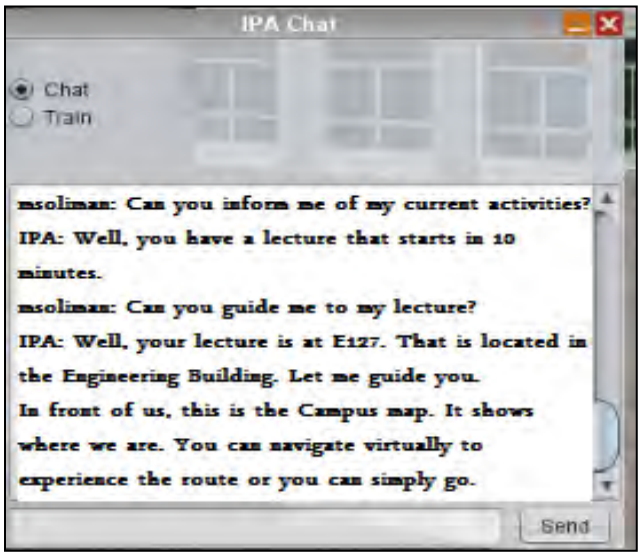


Fig. 5. A Sample text-based interaction between a student and an IPA providing virtual and physical campus navigation support.

C. Augmented Attendance

The immersive environment enables creating virtual classrooms where lectures are given to remote attendees. The instructor and attendees are shown as avatars, the teaching material is a white board or is a presentation that is shared and synchronized among attendee clients. The Immersive environment further enables immersive audio to give the feel of existence in the space. However, face-to-face interactions are still needed in the real campus where traditional lectures occur. Linking the physical reality to the virtual ones creates the possibility to augment the physical attendance with virtual possibilities which is still needed in cases of students with special needs, remote learners, or for special situations that hinder student physical attendance such as illness. In this case, opportunity should be given to attend virtually.

Linking the physical lecture to be casted in the immersive environment can be as simple as having the instructor to use the immersive environment tools. With the positioning services, the physical attendee avatars are in exact position while virtual attendees are shown and tracked as well with the aid of their pedagogical agent that monitors their learning engagement [22].

D. Peer Finding in a Smart Interconnected Campus

Students are expected to use the environment for collaborative learning with peers given the immersive environment affordances for collaborative learning. It has been shown also that an intelligent pedagogical agent has methods to support remote collaborative learning [12]. Finding and meeting the same collaborating peers in campus should support this type of learning. Upon location tracking, finding collaborating peers available in proximity by their location and schedule is supported.

E. Scenario for e-Health, Safety, and Just-in-Time Learning

In emergency situations or when students are in distress, the location of the student can be of great importance to provide immediate help. For example, it helps locating the nearest clinic, directing a medical specialist to the location, or even finding personnel to provide medical first-aid. With the 3D visualization abilities, remote instruction and just-in-time learning can show a simulation of what to do in a particular situation to lower the risk consequences. Immersive environments have been widely investigated to support evacuation scenarios and several safety training in different applications.

VI. Conclusion

This paper studied immersive education aspects to support the smart city concept and its infrastructure. It attempts to answer the question of how the immersive environment can integrate to contribute to the learner citizen of the smart city. Given an interest from different types of users, it offers various new application scenarios and potential new services. The integration into smart city infrastructure promises positive aspects to the learners and the learning community. The paper described the requirements of such integration and the need to strongly interoperate with other citizen services such as health, safety, and mobility in addition to learning systems. This possible integration suggests the need to realize and link the city location services to both the physical and the virtual environment.

Our proposed system architecture is illustrated and the corresponding location aware scenarios are discovered, including contextual navigation support, augmented attendance, locating collaborating peers, and serving learners with special needs. With the physical navigation support, we showed the ability to utilize the exact user location and give the directions to the target location. With the virtual navigation support, the 3D model of the campus allows the navigation to be virtual. This supports the leaner needs (through the avatar) to navigate virtually to the destination and to test the complete path. The text based interaction is supported through linking the virtual campus location model to the student physical location.

The augmented attendance feature enables obtaining the remote/virtual attendance, this in addition to face-to-face interactions in the real campus where traditional lectures occur. Linking the physical reality to the virtual ones creates the possibility to augment the physical attendance with virtual possibilities. Peer finding in a smart interconnected campus is important for collaborative learning and for sessions' scheduling. Finding and meeting the collaborating peers in campus supports many types of collaborative learning based upon location tracking. More services can be used in emergency situations where the location of the student can be of great importance for providing immediate help, such as locating the nearest clinic, directing a medical specialist to the student location, as well as supporting evacuation scenarios and other safety related situations.

With the concept of the smart city, vast heterogeneous services are expected to be available to the user which raises a need for a non-conventional user interface. This paper stresses the opportunity to utilize the concept of intelligent pedagogical agents not only for learning services, but also for all other types of smart city services. This proposes the IPA to act as a central point of interaction between the user and the virtual

environment with multi-modal communication and intelligence abilities.

Definitely, the smart city concept, as a system of systems, mandates further requirements for scaling to larger population and further studying into realizing concrete scenarios of integrating the immersive environment into the smart cities, which is the target for future work.

REFERENCES

[1] B. Dalgarno and M. Lee, "What are the Learning Affordances of 3D Virtual Environments?". British Journal of Educational Technology, Vol. 41, No. 1, 2010, pp. 10–32.

[2] SecondLife, http://secondlife.com/, Online, Accessed Nov. 10, 2015.

[3] OpenWonderland, http://openwonderland.org/, Online, Accessed Nov. 10, 2015.

[4] S. Pellicer, G. Santa, A.L Bleda,. R. Maestre, A.J Jara, and A.G. Skarmeta, "A Global Perspective of Smart Cities: A Survey". Seventh International Conference on Mobile and Internet Services in Ubiquitous Computing, July 2013

[5] IEEE Standards Association (IEEE-SA), Internet of Things (IoT) Ecosystem Study, 2015 IEEE.

[6] The River City Project, Online, Accessed Nov. 15, 2015. http://muve.gse.harvard.edu/rivercityproject/

[7] B. Scheucher, P. Bailey, C. Guetl, and V. J. Harward, "Collaborative Virtual 3D Environment for Internet-accessible Physics Experiments," International Journal of Online Engineering (iJOE), Vol. 5, pp. 65–71, 2009.

[8] M. Soliman and C. Guetl, "Implementing Intelligent Pedagogical Agents in Virtual Worlds: Tutoring Natural Science Experiments in Open Wonderland". IEEE Global Engineering conference, IEEE EDUCON 2013, March 13-15, Berlin, Germany.

[9] M. Soliman and C. Guetl, "Simulating Interactive Learning Scenarios with Intelligent Pedagogical Agents in a Virtual World through BDI-Based Agents". International Journal of Engineering Pedagogy, iJEP, Vol. 3, No. 2, April 2013.

[10] M. Soliman and C. Guetl, "A Survey of Pedagogical Functions of Intelligent Agents in Virtual Learning Environments". Journal of Internet Technologies, Special Issue of Agent Technology in Cyberspace. Vol. 12, No. 6, pp. 995-1005.

[11] M. Soliman and C. Guetl, "Experiences with BDI-based Design and Implementation of Intelligent Pedagogical Agents". International Conference on Interactive Computer-Aided Learning, ICL2012, Sept. 26-28, Villach, Austria.

[12] M. Soliman, "Supporting Collaborative Learning in Virtual Worlds by Intelligent Pedagogical Agents: Approach and Perspectives". In Proceedings of the Second Workshop on Intelligent Support for Learning in Group. Artificial Intelligence in Education, AIED 2013, July 2013, Memphis, Tennessee.

[13] L. Ma, X. Ma, X. Liu, and Y. Xu, "WLAN Indoor Positioning Algorithm Based On Sub-Regions Information Gain Theory". Wireless Communications and Networking Conference (WCNC), April 2013.

[14] E. Lau, B.G. Lee, S.C. Lee, and W.Y. Chung, "Enhanced RSSI-Based High Accuracy Real-Time User Location Tracking System for Indoor and Outdoor Environments". International Journal on Smart Sensing and Intelligent Systems, Vol. 1, No. 2, June 2008.

[15] S. Tennina, M. Di Renzo, F. Graziosi, and F. Santucci, "Locating ZigBee Nodes Using the TI's CC2431 Location Engine: A Testbed Platform and New Solutions for Positioning Estimation of WSNs in Dynamic Indoor Environments". in Proc. of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT 2008), Sep. 2008.

[16] M. Sugano, T. Kawazoe, Y. Ohta, and M. Murata, "Indoor Localization System Using RSSI Measurement of Wireless Sensor Network Based on ZigBee Standard" in Proc. of Wireless Sensor Networks 2006 (WSN 2006), July 2006.

[17] Youngjune Gwon, Ravi Jain, and Toshiro Kawahara, "Robust Indoor Location Estimation of Stationary and Mobile Users". in Proc. of IEEE INFOCOM, March 2004.

[18] R.C. Clark, R.E. Mayer, E-Learning and the Science of Instruction, Wiley, 2011.

[19] K. Abbas, A. Elsaadany, "Adaptive User Interface Development for E-Learning Platform", Fourth International Conference on Informantion Communications Technology, December, 2014, Alexandria.

[20] AIML: Artificial Intelligence Markup Language, http://www.alicebot.org/aiml.html, Online, Accessed Nov. 10, 2015.

[21] Rebecca AIML, http://rebecca-aiml.sourceforge.net, Online, Accessed Nov. 10, 2015.

[22] M. Soliman and C. Guetl, "Evaluation Study and Results of Intelligent Pedagogical Agent-led Learning Scenarios in a Virtual World". Computers in Education, CE, International Convention on Information, Communication Technology, Electronics, and Microelectronics, MIPRO, May 2014, Obatija, Croatia.

Review of source-code plagiarism detection in academia

Matija Novak*
* University of Zagreb/Faculty of Organization and Informatics, Varaždin, Croatia
matija.novak@foi.hr

Abstract: Plagiarism is a big concern in academia and it can be a problem in every course. Plagiarism occurs when someone present others work as their own. Students plagiarize in different areas: homework assignments, essays, projects, etc. In this work focus is on programming courses and plagiarism in programming assignments. While source-code plagiarism detection, is in some way very similar to text plagiarism detection, it is very different in other ways. So, a lot of research is done focusing on source-code plagiarism. Some questions that are researched in this field are: what is considered plagiarism in programming assignments, how to perform plagiarism detection in programming assignments, how to do it automatically, what tool(s) to use, how students cheat in programming courses, how they try to obfuscate cheating, and many other questions. This work is a review of important research papers in the field of source-code plagiarism detection in academia. This paper tries to answer some of the mentioned research questions and give indication to future work.

I. INTRODUCTION

Plagiarism has been a big concern in academia and outside, especially since the popularization of the Internet. Individuals take others' work and present it as their own. In Encyclopedia Britannica the term plagiarism is defined as "the act of taking the writings of another person and passing them off as one's own" [1]. In programming courses students sometimes try to copy source-code parts, or the entire code, from their colleagues, or find similar source-code on the Internet and use it without noting its origin.

A lot of research has been done [2]–[4] that deals with various plagiarism issues in higher education. From the existing studies it can be concluded that academics have differing perspectives on what constitutes plagiarism. For instance, there is disagreement as to whether the reuse of source-code without acknowledgement can be considered plagiarism. The issue of students' inability to understand the boundary between collaboration and collusion was reported in [5] and [6]. According to [7, p. 183]: "Collusion is generally regarded in a softer light than plagiarism".

There are different types of plagiarism in academia. Among those, there are two very similar types that are most commonly mentioned together in the literature [8]: *textual plagiarism* and *source-code plagiarism*. Some similarities between the former and the latter two types are [9], among others: the mean number of characters per line – the mean number of words per sentence; the proportion of 'while' loops to 'for' loops – the proportion of uses of 'their' compared with 'there'; the mean number of lines per subprogram – the mean number of sentences per paragraph.

In spite of many similarities there are "obvious differences between plagiarism written in text and programming languages" [10, p. 22]. Probably the most important difference is that source-code is more structured than natural language, which makes it more likely that two lines are similar. Also, source-code can be easily tokenized, which eliminates the problem of simple variable renaming and other lexical modifications. Simon, Cook, Sheard, Carbone and Johnson established that there are "substantial differences between academic integrity issues in text based and computing assessments" [11, p. 109]. In computing assessments it is more difficult to define plagiarism and collusion; and even if one choses to reference another person's computing work, there are no standard guidelines for doing so. Another problem is that there is a gray area between plagiarism/collusion relating computer code and standard practices like using online tutorials. Simon, Cook, Sheard, Carbone and Johnson concluded that "educating students about academic integrity policies is a necessary but insufficient strategy . . . Academic integrity standards should be accompanied by strategies to imbue students with an understanding and genuine commitment to these standards" [11, p. 113].

The focus of this research is detecting source-code plagiarism. The rest of this paper is structured as follows. Section 2 describes automatic source-code plagiarism detection approach. Section 3 focuses on source-code plagiarism detection engines used to automatically detect similarity between source-code documents. Section 4 describes the source-code preprocessing step. Section 5 gives the conclusion.

II. AUTOMATIC SOURCE-CODE PLAGIARISM DETECTION APPROACH

To prevent detection, students use different *lexical* and *structural modifications* aimed at obfuscating plagiarism [12], [13]. These modifications can vary in complexity, from simple ones like variable renaming to complex ones like replacing control statements. Faidhi and Robinson [14] presented these varying degrees of complexity through six levels of plagiarism. These levels can be matched [9] to either lexical or structural modifications. Consequently, the first two levels can be categorized as lexical modifications and levels four to six as structural modifications. However, it is not clear which category

level three (i.e. change of variable positions) should be placed in.

In programming courses, especially in large student groups and considering many obfuscation methods that are available, only very few cases of plagiarized assignments are found manually. To resolve this problem plagiarism detection engines can be used. Some authors, however, advocate a different approach. For example, [9] and [15] suggested that students should be questioned about their programs and that source-code history should be examined, since automated detection implies a number of issues. Another suggestion [16] is that a student should be given a program and then be asked to develop a new one based on the first one during class. A disadvantage of both of these approaches is that they take too much time in large classes and with multiple assignments during the course. On the other hand, automatic detection can help identify problematic cases in assignments more easily and takes less than either of the aforementioned approaches. To identify plagiarized source-code automatically, there are many plagiarism detection engines which detect similarities in the source-code.

The problem of detecting similar source-code in student assignments is not new and was already researched in the 70's by Ottenstein [17] who used Halstead's metrics [18], [19] to calculate similarity as well as an algorithmic approach for detection and prevention of plagiarism.

Donaldson, Lancaster and Sposato [20] built a plagiarism detection system in 1981 to detect similarity in FORTRAN programming assignments. In this work Donaldson already mentioned several obfuscation methods which have been used by students since [20]: renaming variables, reordering of statements if they do not affect the final result, altering format statements, and splitting single statement into multiple statements.

In its early stage of development, detection was fingerprint based, with occurrences of variables, subprograms, loops, etc. that were counted. Based on these metrics similarity was calculated and plagiarism established. In doing so, it was assumed that "if the difference between two assignments is very small, while at the same time the differences for the majority of the class are large, then perhaps these two assignments deserve closer scrutiny" [20, p. 24]. This approach was characterized as weak [21] and having many limitations so it is not the best technique to calculate similarity. Other metrics have therefore been proposed [22] that can be used to improve the accuracy of similarity detection.

Throughout the years programming languages have changed, giving rise to new problems. For instance, in object oriented programming the reuse of source-code is encouraged and should not be seen as plagiarism. Nevertheless, Cosma and Joy [23] stated that, if its reuse is allowed, the acknowledgement to the original author should be made. Yang, Jiau, and Ssu [24] cautioned that all source-code similarities cannot necessarily be considered plagiarism because there are different causes that contribute to them. Some of these causes are, among others: discussing code fragments or execution results, brainstorming for extra-functionality, discussing the analysis of problem, programming architecture and assignment design, coding style, software reuse as well as assignment requirements and constrains.

All of these causes should be considered when using some plagiarism detection engine. It is important to keep in mind that no tool can be used to accuse somebody of plagiarism. Tools can only suggest that plagiarism may have occurred in some cases if the similarity between student programming assignments is high. As Donaldson, Lancaster and Sposato said: "It is certainly safe to say that neither the detection system described in this paper nor any other detection system will find all occurrences of plagiarism. There is an inherent tradeoff between a highly discriminatory system, which overlooks some instances of cheating, and a less discriminatory one which flags many dissimilar programs" [20, p. 25].

Because of all previously stated, almost all plagiarism detection engines give an overview report and a detailed report. Overview report most often displays similarity (in percentages) or plagiarized pairs with the corresponding similarity. Overview report is the first thing that should be checked by the teacher. Usually, the pairs with higher similarity suggest that plagiarism might be taking place. Teacher should select pairs that have high similarity and examine the detailed report for those pairs. Detailed report, usually, contains parallel view of both submissions indicating the plagiarized parts. Now it is up to the teacher to decide if this parts are enough to consider these pairs as plagiarized. During making decision teacher should take into account all of the before mentioned causes of similarity. These causes increase the similarity although no plagiarism was made. For example it can be that there was a high similarity but most code parts were actually some "getters" and "setters" and not relevant for the core assignment.

III. SOURCE-CODE PLAGIARISM DETECTION ENGINES

In order to automatize plagiarism detection and to increase the chance of finding plagiarized assignments various plagiarism detection engines (further referred to as tools) have been built, including MOSS [25], jPlag [26], Sherlock [13] and YAP3 [27], among others. Plagiarism detection engines differ in many ways. For example, some of them are programming language specific, while others enable detection for various programming languages; some of them are available offline, while others are used as a service; some are open source, while others are not.

Except for before mentioned tools there are many plagiarism detection engines built primarily to detect text similarities. Weber-Wulff with others made in 2004, 2007, 2008, 2010, 2011, 2012 and 2013 extensive investigation of various tools. In most years as in the latest report from 2013 [28] they focused on text similarity. In 2012 [29] they tested 18 systems for effectiveness on text and on programming code. From the tested system the following were ranked as partially useful systems for programming code analysis: jPlag, MOSS, SIM, SPLaT. For program code analysis no system was ranked as useful system or occasionally useful system, all others were ranked as useless systems.

There are other articles like [30]–[32] that compare different plagiarism detection tools and their similarity detection techniques (or algorithms). However, most of them only provide comparison of characteristics, functions and techniques that are gathered from literature and are not based on real objective experiments with different tools used to analyze the same data set. There are few research that compare different tools using experiments like [33] and [34]. Mozgovoy [35] gave an overview of desktop tools used for offline plagiarism detection in a computer program by calculating the complexity of algorithms. From the comparative research in the literature it is evident that no existing tool is perfect. This is because detection algorithm that each tool uses is more or less vulnerable to different obfuscation methods. Among the compared tools the following seem to stand out: jPlag, MOSS, SIM [36], Sherlock, and Plaggie [37]. In addition, there are some newer tools like GATE [38] or GPlag [39], which have not been frequently compared so far. Table I presents some pros and cons of tools that were stated as useful in previously mentioned literature and are mostly compared in the literature.

The most general way of classifying source-code similarity detection tools is, according to [9], [10]: *attribute based* or *structure based*. In the *attribute based* technique – also known as fingerprint based (used in attribute counting technique [10] and ranking measures [10]) – the idea is to create a fingerprint for every document in the collection; in the *structure based* technique (sometimes called content comparison technique) – the structure of the program is represented as tokens which are then analyzed. Another category that can be added in this respect – although it is not strictly related to source-code similarity since all coding syntax is ignored – is *text based,* which nevertheless can sometimes be effective, as shown by Burrows et al. [40].

Another way of classifying plagiarism detection tools is into [41]: *text based* – each statement is treated as a string, and code is considered as a sequence of strings (one option in Sherlock); *token based* – statements are replaced by token (used by Sherlock, jPlag and others); *tree based* – using abstract syntax trees or parse tree (parse tree used by Sim [36]); *graph based* – using program dependency graphs or n-gram diagrams [42], [43]; *hybrid* [44] – combination of a few approaches.

While some of these tools and their algorithms are immune to some obfuscation methods, others are not. Because of that, there is a lot of research that attempts to improve the accuracy of plagiarism detection in different ways: by using new approaches [45], [46], creating new methods for similarity calculation [44] or building new techniques for plagiarism detection [47].

To improve the accuracy of detection, preprocessing techniques are used in some of the tools. For example, Sherlock uses the removal of the comment option. Preprocessing can be useful when design patterns like gang of four (GOF) patterns [48] are used. Since patterns cause a false positive plagiarism detection, better detection results were reported when patterns were removed (de-patternized) from the code [41].

TABLE I. OVERVIEW OF TOOLS

Tool	Pros	Cons
MOSS	Open soruce; Enalbes exclusion of template; Can be used for 25 programming languages; has a Moodle plugin which is usefull if Moodle is used; Side by Side presentation;	Online usage only; has trouble with a lot of the types of plagiarism as it tries very hard to avoid false-positives, thus discarding a lot of information;
JPlag	Open soruce; Avalible offline; Enalbes exclusion of template; Verry fast detection; It gives maximal and avarage similarity and similarity of assignement A to assignement B and vice versa which can be very useful; Side by Side presentation;	Since it parses the results it happens that there are no scores if one small error is in the soruce-code; Supports only 6 programming languages;
SPLaT	Open soruce; Avalible offline; Side by side comparisons; better for text comparisons but quite good for source-code also;	Mainly developed for text comparions;
SIM	Open soruce; Avalible offline; better for text comparisons but quite good for source-code also;	One file is considered as one submission, does not allow group of files as one submission; Does not allow exclusion of template; No graphical interface; tends to compute high similarity scores for cases that clearly do not belong in a top-10; Supports only 7 programming languages;
Marble	Avalible offline; Enalbes exclusion of template; Extandable to other programming languages;	Not open source; No graphical interface; Supports only 5 programming languages;
Plaggie	Open soruce; Avalible offline; Enalbes exclusion of template; Side by Side presentation;	Since it parses the results it happens that there are no scores if one small error is in the soruce-code; Programming language dependent; Returns quite a lot of high scores for cases that turn out to be ‘false alarm’; Supports only Java programming language;
Sherlock Warvick	Open soruce; Avalible offline; Enalbes exclusion of template and other preprocessing of the source-code; It is not programming language dependent; It shows result visualy with graphs; Side by Side presentation;	Tokenized version of detection is realy slow it can take several hours to complete detecion with 50 submissions;

IV. SOURCE-CODE PREPROCESSING

Although changing the detection algorithm can sometimes be useful, to further improve the accuracy of detection it is easier to define a process with various preprocessing steps aimed at 'cleaning' the source-code files from irrelevant parts, as reported in [12], [49].

Durić and Gašević [12] proposed a five step approach (process model) to detect plagiarism:

- Pre-processing – making the detection robust to source-code transformations like: removing blocks of comments, splitting/merging variable declarations, changing variable order, adding redundant statements, etc.
- Tokenization – making the detection robust to source-code transformations like: language translation, renaming of variables, etc.
- Exclusion – creating and removing a tokenized version of the template source-code
- Similarity measurement – detecting similarities
- Final similarity calculation – calculating overall similarity between two sources based on similarities obtained.

Kermek and Novak [49] offered a more detailed 12 step model which deals with multiple files in a single assignment and the problem of having multiple groups with different allowed source-code files. Steps in this model are:

1. Load all student assignments
2. Exclude overhead files – deletes all unused files, leaving only files that should be checked.
3. Merge all files from one assignment – all assignment files of each student are merged into one big file that represents his/her assignment.
4. If allowed source code files exist, go to step 5, otherwise go to step 10.
5. Prepare one file with all the source code whose use is allowed for each group.
6. Choose allowed source code file – it is checked which file among all files given to groups matches best the student assignment and this file is then used further.
7. Compare each student assignment with the file chosen in step 6 – the plagiarism detection engine is run to find similarities between the student assignment and the allowed source code file.
8. Filter student assignment – delete similarities found in step 7 from the student assignment.
9. Iterate steps 7 and 8 until no similarity exists between the file in step 6 and the student assignment. This should be done for every student's assignment.
10. Run the plagiarism detection engine on all student assignments – finding similarities between refined student assignments.
11. Create overview report – with information about similarity in percentage for all pairs sorted by the similarity.
12. Create detailed report – report with concrete file names and source code blocks that are similar between two student assignments.

In the proof-of-concept study [49] it was established that preprocessing steps can be effective in improving the accuracy of plagiarism detection. Also, it was indicated that the usage of this process model led to better plagiarism detection quality compared to the existing tools.

In general plagiarism detection can be performed in four major steps:

- Preprocessing of input data
- Similarity detection
- Visualization of results
- Confirmation of plagiarized cases

Preprocessing as a separate step should be performed in such a way that results obtained during this stage can be used on multiple tools and their algorithms. Accordingly, preprocessing can be further divided into two categories: *independent preprocessing* – wherein obtained data can be used with different tools or algorithms like removal of comments, and *dependent preprocessing* – wherein results can be used only on some tools like tokenization. Dependent preprocessing could be considered as a part of a new tool or algorithm.

During preprocessing multiple techniques can be used. From these, the following techniques found in the literature could be considered independent [10], [41], [49]: removal of white space and comments, removal of imports and common declaration statements, removal of access specifiers, changing data types to synonyms, changing synonym constructs with generic constructs, resorting source-code, exclusion of template code, exclusion of source-code given by teachers, etc.

Tokenization, on the other hand, could be seen as dependent processing. Every tool has its own variation of tokenization. As a result, it is not possible to use the tokenization process results obtained by one tool in another tool.

V. ACCURACY METRICS

Any kind of objective testing evaluation requires appropriate methods. Mozgovoy [35] described two main principles of evaluation. The first is to calculate precision and recall, as used by Whale [50]. The second is to calculate the ratio between highest false match (HFM) and separation (difference between the lowest true/correct match (LTM) and the HFM), as used by Hoad and Zobel [51]. The shortcoming of the second measure is that when HFM is larger than LTM the results can be hard to interpret. However, the benefit of the ratio is that it can

show differences in tool accuracy when precision and recall are equal to one. Since precision and recall are rarely equal to one it is not used very often.

The accuracy measure F1 (calculated from precision and recall) is the most common measure used in the existing research [12], [44], [47], [52]. The results are comparable to some extent but they should be taken with caution since evaluation reported in these studies were performed on different datasets. Furthermore, F1 can also be calculated to give more weight to precision or recall, so this weight needs to be checked when results are compared.

VI. Conclusion

Plagiarism detection has come a long way since its beginnings but no similarity detection engine has so far been proved perfect. Moreover, no similarity detection engine can be used to positively accuse someone of plagiarism. Suspicious cases should always be checked manually by the teacher for example as described at the end of Section 2.

Source-code similarity detection and text similarity detection have many features in common in spite of important differences between them. Using text similarity detection on source-code can sometimes be successful although it is usually not very effective.

There are many source-code similarity detection engines and a large body of research that deals with the comparison between them. In most cases the focus of comparative studies is on characteristics, functions and techniques that are gathered from literature. On the other hand, there are only a few studies that report real objective experiments with different similarity detection engines used on identical data sets. Future research in this field should therefore be directed to fill this research gap. From all the available tools the following seem to stand out in literature, that deals with source-code plagiarism, in terms of accuracy and that are open source: jPlag, MOSS, SIM, Sherlock, and Plaggie. But those tools might not be the best on the market. There are also other tools like GATE and GPlag which so far have not been extensively compared. The comparisons of newer tools should be in focus of future research.

There are different ways to improve the accuracy of similarity detection engines. An interesting method is to use preprocessing techniques which 'clean' the data from all unnecessary source-code parts. Preprocessing is especially useful if the technique can be used in combination with multiple tools.

References

[1] Encyclopædia Britannica Inc, "plagiarism," *Encyclopædia Britannica*, 2015. [Online]. Available: http://www.britannica.com/topic/plagiarism. [Accessed: 10-Sep-2015].

[2] G. Cosma and M. Joy, "Source-Code Plagiarism : a Uk Academic Perspective," *7th Annu. Conf. HEA Netw. Inf. Comput. Sci.*, pp. 116–120, 2006.

[3] D. Zhang, M. Joy, G. Cosma, R. Boyatt, J. Sinclair, and J. Yau, "Source-code plagiarism in universities: a comparative study of student perspectives in China and the UK," *Assess. Eval. High. Educ.*, vol. 39, no. 6, pp. 743–758, Aug. 2014.

[4] F. Culwin, A. Macleod, and T. Lancaster, "Source code plagiarism in UK HE computing schools," in *2nd Annual LTSN-ICS Conference*, 2001.

[5] M. Joy, G. Cosma, J. Y.-K. Yau, and J. Sinclair, "Source Code Plagiarism - A Student Perspective," *IEEE Trans. Educ.*, vol. 54, no. 1, pp. 125–132, Feb. 2011.

[6] R. Barrett and A. L. Cox, "'At least they're learning something': the hazy line between collaboration and collusion," *Assess. Eval. High. Educ.*, vol. 30, no. 2, pp. 107–122, Apr. 2005.

[7] R. Fraser, "Collaboration, Collusion and Plagiarism in Computer Science Coursework," *Informatics Educ.*, vol. 13, no. 2, pp. 179–195, Sep. 2014.

[8] A. M. El Tahir Ali, H. M. Dahwa Abdulla, and V. Snášel, "Overview and comparison of plagiarism detection tools," in *11th Annual Workshop on Databases, Texts, Specifications and Objects, DATESO 2011*, 2011, vol. 706, pp. 161–172.

[9] T. Lancaster, "Effective and Efficient Plagiarism Detection," South Bank University, 2003.

[10] P. Clough, "Plagiarism in natural and programming languages: an overview of current tools and technologies," *Research Memoranda: CS-00-05, Department of Computer Science, University of Sheffield, UK*, pp. 1–31, 2000.

[11] Simon, B. Cook, J. Sheard, A. Carbone, and C. Johnson, "Academic integrity perceptions regarding computing assessments and essays," in *Proceedings of the tenth annual conference on International computing education research - ICER '14*, 2014, pp. 107–114.

[12] Z. Duric and D. Gasevic, "A Source Code Similarity System for Plagiarism Detection," *Comput. J.*, vol. 56, no. 1, pp. 70–86, Jan. 2013.

[13] M. Joy and M. Luck, "Plagiarism in programming assignments," *Educ. IEEE Trans.*, vol. 42, no. 2, pp. 129–133, 1999.

[14] J. A. W. Faidhi and S. K. Robinson, "An empirical approach for detecting program similarity and plagiarism within a university programming environment," *Comput. Educ.*, vol. 11, no. 1, pp. 11–19, 1987.

[15] J. K. Harris, "Plagiarism in computer science courses," in *Proceedings of the conference on Ethics in the computer age -*, 1994, pp. 133–135.

[16] J. Sant, "Code Repurposing as an Assessment Tool," in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, 2015, vol. 2, pp. 295–298.

[17] K. J. Ottenstein, "An algorithmic approach to the detection and prevention of plagiarism," *ACM SIGCSE Bull.*, vol. 8, no. 4, pp. 30–41, Dec. 1976.

[18] M. H. Halstead, "Natural laws controlling algorithm structure?," *ACM SIGPLAN Not.*, vol. 7, no. 2, pp. 19–26, Feb. 1972.

[19] M. H. Halstead, *Elements of Software Science (Operating and Programming Systems Series)*. New York, NY, USA: Elsevier Science Inc., 1977.

[20] J. L. Donaldson, A.-M. Lancaster, and P. H. Sposato, "A plagiarism detection system," *ACM SIGCSE Bull.*, vol. 13, no. 1, pp. 21–25, Feb. 1981.

[21] V. Y. Shen, S. D. Conte, and H. E. Dunsmore, "Software Science Revisited: A Critical Analysis of the Theory and Its Empirical Support," *IEEE Trans. Softw. Eng.*, vol. SE-9, no. 2, pp. 155–165, Mar. 1983.

[22] R. J. Leach, "Using metrics to evaluate student programs," *ACM SIGCSE Bull.*, vol. 27, no. 2, pp. 41–43, Jun. 1995.

[23] G. Cosma and M. Joy, "Towards a definition of source-code plagiarism," *IEEE Trans. Educ.*, vol. 51, no. 2, pp. 195–200, 2008.

[24] F.-P. Yang, H. C. Jiau, and K.-F. Ssu, "Beyond plagiarism: An active learning method to analyze causes behind code-similarity," *Comput. Educ.*, vol. 70, no. 0, pp. 161–172, 2014.

[25] S. Schleimer, D. S. Wilkerson, and A. Aiken, "Winnowing: local algorithms for document fingerprinting," in *Proceedings of the 2003 ACM SIGMOD international conference on on Management of data - SIGMOD '03*, 2003, p. 76.

[26] L. . Prechelt, G. . Malpohl, and M. . Philippsen, "Finding plagiarisms among a set of programs with JPlag," *J. Univers. Comput. Sci.*, vol. 8, no. 11, pp. 1016–1038, 2002.

[27] M. J. Wise, "YAP3: improved detection of similarities in computer program and other texts," in *Proceedings of the twenty-seventh SIGCSE technical symposium on Computer science education - SIGCSE '96*, 1996, vol. 28, no. 1, pp. 130–134.

[28] D. Weber-Wulff, C. Möller, J. Touras, and E. Zincke, "Plagiarism Detection Software Test 2013," Berlin, 2013.

[29] D. Weber-Wulff, K. Köhler, and C. Möller, "Collusion Detection System Test Report 2012," Berlin, 2012.

[30] M. Mozgovoy, K. Fredriksson, D. White, M. Joy, and E. Sutinen, "Fast Plagiarism Detection System," in *Lecture Notes in Computer Science*, vol. 3772 LNCS, Berlin, Heidelberg: Springer-Verlag, 2005, pp. 267–270.

[31] T. P. Gondaliya, H. Joshi D., and H. Joshi, "Source Code Plagiarism Detection ' SCPD et ' : A Review," vol. 105, no. 17, pp. 18–25, 2014.

[32] D. Luke, P. S. Divya, S. L. Johnson, S. Sreeprabha, and E. B. Varghese, "Software Plagiarism Detection Techniques : A Comparative Study," vol. 5, no. 4, pp. 5020–5024, 2014.

[33] J. Hage, P. Rademaker, and N. van Vugt, "Plagiarism Detection for Java: A Tool Comparison," in *Computer Science Education Research Conference*, 2011, pp. 33–46.

[34] V. T. Martins, D. Fonte, P. R. Henriques, and D. Da Cruz, "Plagiarism detection: A tool survey and comparison," in *3rd Symposium on Languages, Applications and Technologies, SLATE 2014*, 2014, vol. 38, pp. 143–158.

[35] M. Mozgovoy, "Desktop tools for offline plagiarism detection in computer programs," *Informatics Educ.*, vol. 5, no. 1, pp. 97–112, 2006.

[36] D. Gitchell and N. Tran, "Sim: A utility for detecting similarity in computer programs," in *SIGCSE Bulletin*, 1999, vol. 31, no. 1, pp. 266–270.

[37] A. Ahtiainen, S. Surakka, and M. Rahikainen, "Plaggie: GNU-licensed source code plagiarism detection engine for Java exercises," in *6th Baltic Sea Conference on Computing Education Research - Koli Calling 2006*, 2006, vol. 276, pp. 141–142.

[38] O. Müller and S. Strickroth, "GATE - Ein System zur Verbesserung der Programmierausbildung und zur Unterstützung von Tutoren," in *Proceedings of the First Workshop "Automatische Bewertung von Programmieraufgaben" (ABP)*, 2013.

[39] C. Liu, C. Chen, J. Han, and P. S. Yu, "GPLAG: Detection of software plagiarism by program dependence graph analysis," in *KDD 2006: 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, vol. 2006, pp. 872–881.

[40] S. Burrows, S. M. M. Tahaghoghi, and J. Zobel, "Efficient plagiarism detection for large code repositories," *Softw. Pract. Exp.*, vol. 37, no. 2, pp. 151–175, Feb. 2007.

[41] A. Asadullah, B. M., I. Stern, and V. D. Bhat, "Design Patterns Based Pre-processing of Source Code for Plagiarism Detection," in *2012 19th Asia-Pacific Software Engineering Conference*, 2012, vol. 2, pp. 128–135.

[42] A. Jadalla and A. Elnagar, "PDE4Java: Plagiarism Detection Engine for Java source code: a clustering approach," *Int. J. Bus. Intell. Data Min.*, vol. 3, no. 2, pp. 121–135, 2008.

[43] S. Sharma, C. S. Sharma, and V. Tyagi, "Plagiarism detection tool 'Parikshak,'" in *2015 International Conference on Communication, Information & Computing Technology (ICCICT)*, 2015, pp. 1–7.

[44] H.-J. Song, S.-B. Park, and S. Y. Park, "Computation of Program Source Code Similarity by Composition of Parse Tree and Call Graph," *Math. Probl. Eng.*, vol. 2015, pp. 1–12, 2015.

[45] T. Ohmann and I. Rahal, "Efficient clustering-based source code plagiarism detection using PIY," *Knowl. Inf. Syst.*, vol. 43, no. 2, pp. 445–472, May 2015.

[46] R. Brixtel, M. Fontaine, B. Lesner, C. Bazin, and R. Robbes, "Language-Independent Clone Detection Applied to Plagiarism Detection," in *2010 10th IEEE Working Conference on Source Code Analysis and Manipulation*, 2010, pp. 77–86.

[47] B. Muddu, A. Asadullah, and V. Bhat, "CPDP: A robust technique for plagiarism detection in source code," in *2013 7th International Workshop on Software Clones (IWSC)*, 2013, pp. 39–45.

[48] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*. Pearson Education, Limited, 2005.

[49] D. Kermek and M. Novak, "Process model improvement for source code plagiarism detection in student programming assignments," *Informatics Educ.*, In press, 2015.

[50] G. Whale, "Identification of program similarity in large populations," *Comput. J.*, vol. 33, no. 2, pp. 140–146, 1990.

[51] T. C. Hoad and J. Zobel, "Methods for identifying versioned and plagiarized documents," *J. Am. Soc. Inf. Sci. Technol.*, vol. 54, no. 3, pp. 203–215, Feb. 2003.

[52] C.-H. Hsiao, M. Cafarella, and S. Narayanasamy, "Using web corpus statistics for program analysis," in *Proceedings of the 2014 ACM International Conference on Object Oriented Programming Systems Languages & Applications*, 2014, vol. 49, no. 10, pp. 49–65.

The comparison of impact offline and online presentation on student achievements: A case study

Esztelecki Péter
University of Szeged, Faculty of Science and Informatics
epeter@inf.u-szeged.hu

Kőrösi Gábor
University of Szeged, Faculty of Science and Informatics
kgabor@inf.u-szeged.hu

Námesztovszki Zsolt
University of Novi Sad
zsolt.namesztovszki@magister.uns.ac.rs

Major Lenke
University of Novi Sad
major.lenke@magister.uns.ac.rs

Abstract - **Nowadays online presentations and educational videos are frequently integrated into various-learning environments and applications, such as MOOCs, global sets of conferences or video-sharing websites. This paper presents the findings of a comparison of online presentations (educational videos) and offline presentations. The total number of student participants in this research was 191, mostly primary and secondary school students from Serbia as well as Hungary, studying both online and offline learning environments within the framework of the course Conscious and safe internet usage. The impact of offline and online presentations was investigated using both pre- and post-presentation questionnaires. Statistical analysis was used to measure the impact of offline and online presentations, in addition to other factors determining student achievements.**

Key words: offline presentation, online presentation, student achievement

I. INTRODUCTION

Who has not heard of the cloud school project founded by Sugata Mitra, introducing a revolutionary school which eliminates the teacher in favour of a projector and Internet connection. Adhering to this spirit and following this line of logic, hundreds of video-based online learning platforms have been appearing on a daily basis. A few years ago Kőrösi (2015) highlighted that E-learning websites would bring a positive and lasting change to education [1]; however, a growing number of experts advocate video-based education and MOOC courses that would trigger a substantial reform [2][3][4][5]. This trend seems to be accelerating rather than slowing down [6], though a rapid change may carry a considerable number of unsolved questions. To meet these requirements, there is a need for conducting empirical studies, and completing case studies, since it has become obvious that traditional, distance, and video-based learning have their

advantages and disadvantages, respectively. The advantages of the first method include being a closed and restricted educational environment led and motivated by a teacher, while the latter two allow learning freed from time and place constraints [7].

The characteristics of the above-mentioned methods must be, however, taken under a close examination in order to reveal if one could produce similar values through keeping the same motivational factors and teacher control with the traditional and video-based learning environment (MOOC). To answer such questions, the authors examined online and off-line opportunities to be incorporated into education based on the same conditions. This paper presents the findings of the comparison of online (educational videos) and offline presentations. The total number of student participants in this research was 191, mostly primary and secondary school students from Serbia as well as Hungary, studying both online and offline learning environments within the framework of the course Conscious and safe internet usage. The impact of offline and online presentations was investigated using both pre- and post-presentation questionnaires. Statistical analysis was used to measure the impact of the offline and online presentations, as well as other factors determining student achievements.

The paper consist of the following sections: Traditional vs. online-video learning, Background, Interpretation of the research results and Conclusion.

The first section examines the comparison of online and offline video learning, the second section describes the theoretical background and the hypotheses, methods and the sample of this investigation. In the third section authors present the results of research and based on this results in the fourth section have been conceptualized the conclusion.

II. TRADITIONAL VS. ONLINE-VIDEO LEARNING

The basis of modern online education does not originate in websites or pdf files but they are constructed using videos. Taking a closer look at video-based teaching, one must realize that it has been around for a significant period of time going back as far as filmstrips studied during World War II as a training tool for soldiers [8]. Despite their 70-year long existence, there are still unresolved questions to be answered. The mystery behind this may be justified by the under-development of the technical background. In those times, the recording of a video series for lectures cost a fortune, while today mobile phones and video sharing platforms provide almost free tools to complete the same task. Despite technical possibilities, it is noticeable how many obstacles still lie in the way of such educational possibilities. Not surprisingly, the United States is a great step ahead in terms of developing online platforms for video-based education involving universities and public institutions. To list but a few, these are Udacity, edX, etc. Furthermore, with the constant spread of this learning system, it will be highly desirable to reveal the differences and overlaps between the traditional and virtual educational methods. This concept is reinforced by the fact that the number of online courses rises exponentially, day by day [6]. In line with the above-stated, [9] attempted to provide a list of the strengths and weaknesses of the traditional educational methods and the online (video based) teaching.

A. Online Education

Pros

1. saving
2. convenience and flexibility
3. more marketable skills and job focus
4. learning digital skills in a digital space
5. keeping up with the times

Cons

For a list of cons regarding online education, all one really needs to do is look at the pros list for traditional education. Online education will most likely not provide these benefits. It also offers limited career options. One will not be able to become a doctor, lawyer, etc. with online education nor will one be able to enjoy the typical college experience. Online education empowers people who are self-directed and motivated but it can be lonely compared to traditional education.

B. Traditional Education

Pros

1. experience
2. network
3. people will recognize the obtained degree
4. provides facilities necessary for certain studies and activities
5. some careers and professions generally require a degree

Cons

1. it is considerably expensive
2. it does not always save one from underemployment
3. it might not provide the skills needs

By observing Rauch's (2015) arguments, one cannot reveal new information regarding this field, since most of the researchers are already well informed [9]. Society cannot easily be convinced by enumerating rational arguments favouring new technological solutions because innovations have always been regarded with a certain suspicion. The success of video lessons cannot be neglected given the fact that studies have proven the efficiency of educational films created long before the Internet era. [10] by using examples also touch upon these facts:

- Watching the television program Blue's Clues has strong effects on developing preschool viewers' flexible thinking, problem solving, and prosocial behaviors.

- Court TV's Choices and Consequences program reduced middle school students' verbal aggression-including tendencies to tease, swear at, and argue with others.
- Viewing Sesame Street was positively associated with subsequent performance in reading, mathematics, vocabulary, and student readiness. A "recontact" study with a sample of 15- to 20-year-olds found that those who had been frequent viewers of Sesame Street at age 5 had significantly better grades in English, science, and mathematics; read more books for pleasure; and had a higher motivation to achieve.

A number of researchers argue supporting this line of thought. Video is a rich and powerful medium being used in E-learning. It can present information in an attractive and consistent manner. [11] The use of video is only beginning to meet the needs of today's and tomorrow's learners. Video can help educators address the challenge of different learning styles and enhance the way in which today's children and youth access, absorb, interpret, process and use information [12]. One of the greatest strengths of television and video is the ability to communicate with viewers on an emotional, as well as a cognitive, level. Because of this ability to reach viewers' emotions, video can have a strong positive effect on both motivation and affective learning. [13]. Although the impact of video and multimedia technologies in educational outcomes is a field of ongoing research, the pedagogical impact of video can be summarized by three key concepts:
1) Interactivity with content (the learner relates to visual content, whether verbally, by note taking or thinking, or by applying concepts)
2) Engagement (the learner connects to the visual content, becoming drawn in by video, whether on-demand or real time)
3) Knowledge transfer and memory (the learner may remember and retain concepts better than with other instructional media) because video combines many [12].
It is, however, not enough to highlight all the positive aspects because as [9] points out, one must consider the differences between the traditional and the new online educational methods, also observed by others. Ya Ni's (2013) opinion ought to be stated here, namely that: An important component of classroom learning are social and communicative interactions between a student and a teacher, and a student and a student [14]. Nyíri (2009) enumerated four important deficiencies: Firstly, due to the shift from a personal communication to virtual communication, cognitive losses may appear somewhere in the process [15]. Secondly, information carried over by paper-based or printed texts regarding cognitive qualities are different from textual information appearing on the screen. Thirdly, a study must clarify the question to what extent can information be carried by texts in a digital environment supplemented by information mediated by images. Lastly, one must check an obvious but hard-to-analyze phenomenon, specifically, to what extent different personality types differ in solving tasks in a virtual environment.
A MOOC course would touch upon these questions and deficiencies in an attempt to combine the possibilities of both the traditional and the video-based education. Finally, to verify all the listed assumptions, they must be put to practice in real life circumstances, whose efficiency must be further confirmed by empirical and case studies.

III. BACKGROUND

A. Hypotheses

(1) It was assumed that the achievements of both groups would increase during the test process, in accordance with an increase in the level of acquired knowledge.

(2) It was assumed that the experimental (online) group would achieve a higher average of points, over three tests, control (offline group).

(3) It was assumed that the average total score of experimental (online) group would be significantly higher than the average total score of the control (offline group).

(4) It was assumed that the time spent in an online environment is directly proportional to achievements in the case of both groups.

B. Methods

The online course was held in February and March 2015 and the offline course was organized in January 2016. The course was called "Conscious and safe internet usage" and it consists of the following modules:

- 1. module - Digital footprint
- 2. module - Conscious and safe internet usage
- 3. module - Online bullying

After each module, the students filled in some tests, for the experimental (online) group it was online tests, while the control (offline) group received paper-based offline tests. The tests were identical, all sheets contained 3 x 10 questions and the maximum score was 20.

C. Participants

A total of 191 students participated in this investigation . The participants were divided into two groups:

1.) 120 students in the experimental (online) group
2.) 71 students in the control/offline group (Table1.).

TABLE I. THE DISTRIBUTION OF PARTICIPANTS INTO TWO GROUPS

	N	%
experimental/online group	120	63
control/offline group	71	37
total	**191**	**100**

The average age of the participants was 19. The youngest participant was 11 and the oldest was 63 years old. The average age in the experimental/online group was 23,while in the control/offline group it was 13 (see Table2.).

TABLE II. THE AGE OF THE SAMPLE

	Mean	Minimum	Maximum	N
age of experimental/ online group	22.9	12	63	120
age of control/offline group	12.9	11	15	71
average of both groups	19.2	11	63	191

The countries of residence of the students were Hungary, Serbia and Romania with the following distribution: 19% of the participants were from Hungary, 80% from Serbia and there was one participant from Romania (see Table3.).

TABLE III. RESIDENCE OF THE SAMPLE (COUNTRY)

country	N	%
Hungary	37	19
Serbia	153	80
Romania	1	1
total	**191**	**100**

IV. INTERPRETATION OF THE RESEARCH RESULTS

The members of the experimental/online group reached an average of 15,15 points during the first test, 16,22 point during the second test and 18,73 during the third test. In the control/offline group the average points was 12,34 during the first, 12,37 during the second and 13,97 during the third test (see Table4.).

TABLE IV. RESULTS OF THREE TESTS

	experimental/online group (N=120)				control/offline group age (N=71)			
	Mean	**Std. Dev.**	**Min.**	**Max.**	**Mean**	**Std. Dev.**	**Min.**	**Max.**
first test	15.15	3.2	2	20	12,34	3,4	2	20
second test	16.22	2.9	8	20	12,37	3,4	2	18
third test	18.73	2.2	8	20	13,97	3,8	0	20
total	50.10	5.5	28	60	38,68	8,5	14	54

A significant increase was proven based on the paired samples t-test values between the experimental (online) group means of the first and second test values ($t = -2.8$, $p = 0.006$), the second and third ($t = -8.5$, $p = 0.001$), and the first and third ($t = -11.3$, $p = 0.001$) tests (see Table5.).

TABLE V. RESULTS OF PAIRED SAMPLES T-TEST IN THE EXPERIMENTAL / ONLINE GROUP

	t	df	Sig. (2-tailed)
first test - second test	-2.817	119	0.006
second test - third test	-8.582	119	0.001
first test - third test	-11.287	119	0.001

Examining the same values, using paired samples t-test, in the control (offline) group there were no significant differences between the results of the first and second tests ($t = -0.64$, $p = 0.94$). On the other hand, there was a statistically sustained increase in the student achievements between the second and third ($t=-3.69$ $p=0.001$) and the first and third test ($t=-4.02$ $p=0.001$) (see Table6.).

TABLE VI. RESULTS OF PAIRED SAMPLES T-TEST IN THE CONTROL/OFFLINE GROUP

	t	df	Sig. (2-tailed)
first test - second test	-0.64	70	0.94
second test - third test	-3.69	70	0.001
first test - third test	-4.02	70	0.001

Comparing the performance of the experimental and control groups in the first ($t = 5.7$, $p = 0.001$) and second ($t = 8.3$, $p = 0.001$) and third ($t = 9.7$, $p = 0.001$) tests and the total score ($t=10.1$ $p=0.001$), there were significant differences. The experimental (online) group in each case achieved better results than the control (offline) group (see Table 7).

TABLE VII. RESULTS OF TWO-SAMPLE T-TEST

	F	Sig.	t	df	Sig. (2-tailed)
first test	0.580	0.447	5.754	189	0.001
second test	1.575	0.211	8.276	189	0.001
third test	23.885	0.001	9.717	98.243	0.001
total	15.963	0.001	10.076	105.488	0.001

The time spent online for all groups was 2.3 hour. In the experimental (online) group this value was 2.4 and in the control (offline) group it was 2.2 hour. Most of the members of the experimental and control group spent 1-3 hours in an online environment (see Table 8).

TABLE VIII. TIME SPENT ONLINE

	experimental/online group (N=120)		control/offline group age (N=71)
time	**N**	**%**	**N**
less than 1 hour	15	12,5	9
1-3 hour	63	52,5	46
4-5 hour	26	21,7	11
more than 5 hour	16	13,3	5
average (hour)	2,36		2,17

Authors could not prove a statistically significant relation between the time spent online and student achievements.

V. CONCLUSION

Based on the results, the first hypothesis that assumed both groups' average performance would increase during the three measurements was proven, the average performance increase in the case of the experimental (online) group was greater than that of the control (offline) group had.

The assumption of the second hypothesis regarding the experimental group was proven; the group working in an online environment achieved significantly better results in all three measurements compared to the previous measurement. Regarding the control group there was no performance increase between the first and the second measurements although the third measurement showed significantly higher points than the first and second ones. This means that the program is also useful in an offline environment, however, the performance increase takes more time.

Both the experimental and the control groups' accumulated results based on the three measurements showed significant difference in favor of the experimental group. Thus according to the third hypothesis, at the end of the program the experimental group achieved better results than the control group.

The fourth hypothesis that assumed that time spent in an online environment is directly proportional to the performance could not be proven for either group.

This analysis showed that learning free from time and place using online video-based learning is more effective than offline learning method, therefore it is not surprising that more and more universities provide space for online training courses. Although the results show positive changes, many people, including the majority of the ministries of education, harbor doubts about the effectiveness and efficiency of this method, nonetheless numerous applications demonstrate that there is a need of this kind of education.

REFERENCES

[1] Kőrösi, G. (2015), Tanár nélküli tanulás a jövő iskolájában? Az online videó oktatási jellegű alkalmazásának lehetőségi és buktatói, VII. Nemzetközi XIII. Országos Interdiszciplináris Konferencia, PTE Grastyán Endre Szakkollégium, Pécs, 146-153.

[2] Námesztovszki, Zs., Kôrösi, G., Esztelecki P., Vinkó A., Kovács C. (2015), Tapasztalatok És Következtetések Egy Kísérleti MOOC Kapcsán, VII. Oktatás-Informatikai Konferencia, ELTE Pedagogikum Központ, ELTE Pedagógiai És Pszichológiai Kar, Budapest, 302-314.

[3] Allen, E., Seaman, J. (2014), Grade change: Tracking online education in the United States. Babson Survey Research Group and Quahog Research Group, LLC.

[4] Fini, A. (2009), The technological dimension of a massive open online course: The case of the CCK08 course tools. International Review of Research in Open and Distance Learning, 10(5), 74-96.

[5] Stewart, B. (2013), Massiveness + openness = New literacies of participation? Journal of Online Learning and Teaching, 9(2), 228-238.

URL: http://jolt.merlot.org/vol9no2/stewart_bonnie_0613.htmU

[6] Szeghegyi, Á., Szoboszlai V., Velencei, J. (2014), Informal Post-Experiential Learning, Acta Polytechnica Hungarica, Vol. 11, No. 4, 249-250.

[7] Esztelecki, P., Kőrösi, G. (2015), Idegennyelv-tanulás megvalósítása online eszközökkel, VI. BáthoryBrassai Nemzetközi Tudományos Konferencia, Óbudai Egyetem, Óbuda. 91-96. o.

[8] Hovland, C.I., Lumsdaine, A.A. & Sheffield, F.D. (1949), Experiments on mass communication. Princeton, NJ: Princeton University Press.

[9] Rauch, J. (2015), Online Education vs.Traditional Education: The Pros and Cons, SkilledUp, San Jose

[10] Cradler, J., Freeman, M., Cradler, R. (2005), Research Basis for the Schlessinger Media Programs Grades K-12, Educational Support Systems, San Mateo, USA, URL: https://www.bcps.org/offices/lis/safari/Cradlerpaper62105_32_June22[1].pdf

[11] Zhang, D., Zhou, L., O. Briggs, R., Nunamaker J.F.Jr. (2006), Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness,Information & Management, Elsevier, vol 43 15–27.p. URL: http://er.educause.edu/articles/2013/1/the-mooc-model-challenging-traditional-education

[12] Greenberg A. D., Zanetis J. (2012), The Impact of Broadcast and Streaming Video in Education Report commissioned by Cisco Systems Inc., Wainhouse Research, LLC. URL: http://www.cisco.com/web/strategy/docs/education/ciscovideowp.pdf

[13] Cruse, E. (2011), Using educational video in the classroom: Theory, research and practice.

URL: http://www.libraryvideo.com/articles/article26.asp

[14] Ni, Y.A. (2013), Comparing the Effectiveness of Classroom and Online Learning: Teaching Research Methods,Journal of Public Affairs Education 19(2), 199–215.p.

[15] Nyíri, K. (2009), Virtuális pedagógia – a 21. század tanulási környezete, Oktatáskutató és fejlesztő intézet, Magyarország, Budapest URL: http://www.ofi.hu/tudastar/iskola-informatika/nyiri-kristof-virtualis

Digital Competences for Teachers: Classroom Practice

Martina Filipović Tretinjak*, Vesna Anđelić**
*School of Electrical Engineering (Elektrotehnička škola), Zagreb, Croatia
**Agency for Vocational Education and Training and Adult Education Zagreb, Croatia
martinafilipovic.tretinjak@gmail.com
vesna.andelic@asoo.hr

Abstract - Digital competence is the set of knowledge, skills and attitudes that should be acquired by all people in order to ensure critical and creative use of ICT and digital media for the purpose of achieving goals related to work, learning and/or leisure. In order to make learning and teaching effective, all teachers should become familiar with the following digital competence areas: information (e.g. identify, locate, retrieve, store, organize and analyze digital information), communication (e.g. communicate through online tools, taking into account privacy, safety and netiquette), safety on the internet (e.g. properly manage personal protection, data protection, digital identity protection), problem solving (e.g. identify digital needs and resources, solve conceptual problems through digital means, creatively use technologies, solve technical problems) and content creation (e.g. create and edit new content (from word processing to images and video), integrate and re-elaborate previous knowledge and content, apply intellectual property rights and licenses). This paper describes various methods for the implementation of digital competences in the classroom learned through the DIGICOMP portal. The DIGICOMP portal is an e-learning portal created under the project entitled Digital Competences for Teachers (DIGICOMP).

I. INTRODUCTION

Digital competence can be defined as the set of knowledge, skills and attitudes to ensure critical and creative use of ICT and digital media for the purpose of achieving goals related to work, learning and/or leisure. In order to confidently and critically use ICT in teaching, cooperate and communicate with the assistance of ICT and to use various tools for creating e-content, all teachers need to become familiar with the following digital competence areas: information (e.g. identify, locate, retrieve, store, organize and analyze digital information), communication (e.g. communicate through online tools, taking into account privacy, safety and netiquette), safety on the internet (e.g. properly manage personal protection, data protection, digital identity protection), problem solving (e.g. identify digital needs and resources, solve conceptual problems through digital means, creatively use technologies, solve technical problems) and content creation (e.g. create and edit new content (from word processing to images and video), integrate and re-elaborate previous knowledge and content, apply intellectual property rights and licenses). The definition for each competence area is given in Table I.

TABLE I. DIGITAL COMPETENCE AREAS

Competence	Definition
INFORMATION	Ability to identify, locate, retrieve, store, organize and analyze digital information, judging its relevance and purpose.
COMMUNICATION	Ability to communicate in digital environments, share resources through online tools, collaborate with others through digital tools, interact with and participate in communities and networks and raise cross-cultural awareness.
CONTENT CREATION	Ability to create and edit new content (from word processing to images and video), integrate and re-elaborate previous knowledge and content, produce creative expressions, media outputs and programming and deal with and apply intellectual property rights and licenses.
SAFETY	Ability to properly manage personal protection, data protection, digital identity protection, security means and safe and sustainable use.
PROBLEM SOLVING	Ability to identify digital needs and resources, make informed decisions on most appropriate digital tools according to the purpose or need, solve conceptual problems through digital means, creatively use technologies, solve technical problems and update own and other's competence.

The main focus of this paper is to describe the use of specific portals and programs for the implementation of digital competences in the classroom learned through the DIGICOMP portal [1]. The portal can be accessed via the following URL: http://moodle.digital-competences-for-teachers.eu/index.php.

DIGICOMP portal is an e-learning portal created under the project entitled Digital Competences for Teachers (DIGICOMP) [2]. The DIGICOMP project is

carried out as part of the Erasmus+ program. The Agency for Vocational Education and Training and Adult Education (Croatia) is the project coordinator. The project is carried out in collaboration with partner institutions from Slovenia (PIA, informacijski sistemi in storitve d.o.o. and Aeroform d.o.o.), Spain (Fundación Metal Asturias), Lithuania (Kauno Regioninis Inovaciju Centras) and Turkey (Konya İl Milli Eğitim Müdürlüğü).

All competence areas' definitions are shown in Tables I to VI taken from the document: Methodology for peer evaluation, DIGICOMP project.

Most of the mentioned activities require computers with access to the internet in the classroom; therefore, these activities can be used for courses that already have in-class computer-based activities. All activities are intended for not more than 14 students per classroom.

Some activities (e.g. subject forum or homework and/or project submission) require from students to have computers with access to the internet at home. However, these activities can be modified for the use in the classroom, on school computers.

II. INFORMATION

A detailed description of the competence area information is shown in Table II.

TABLE II. COMPETENCE AREA: INFORMATION

Competence	Definition
Browsing, searching and filtering information	To access and search for online information, to articulate information needs, to find relevant information, to select resources effectively, to navigate between online sources, to create personal information strategies.
Evaluating information	To gather, process, understand and critically evaluate information.
Storing and retrieving information	To manipulate and store information and content for easier retrieval, to organize information and data.

A. Exploring a specific course topic

Whether students are working on their homework, researching a specific topic in the classroom as a part of the project-oriented learning or group work activities, they need to know how to find information on the internet, effectively use the resources found, evaluate and store information.

Let's take, for example, a classroom activity called the information seeker. In this activity, students are required to create a presentation on the specific course topic. First, teacher creates a folder on Google Drive (folder is shared with students) and divides students into three (or more) groups. One topic and one search engine (e.g. Google, Bing, Yahoo) is assigned to each group. Students then conduct the search following and implementing the six-stage (initiation, selection, exploration, formulation, collection and search closure) Information Search Process (ISP).

After the students have completed the information search, they present the information that was found through the process using Google Docs (e.g. Google Slides). The created document is stored in the shared folder and can be easily accessed by other students from this or other classes. Also, in their presentation, students need to evaluate all information from the internet used in answering these evaluation questions concerning: audience (Who was the resource written for? Who was the website created for?), accuracy (Is the information accurate or valid?), bias (What is the resource's point of view? What is the website's point of view?), credibility (Who is the author and publisher?) and currency (What is the publication date of the resource? When was the website last updated?). At the end of this activity, each group presents their work to the rest of the class.

B. Storing and organizing homework and class assignments using Google Drive storage

On their Google Drive storage, students have to create a folder that will contain all of their homework and class assignments created during the school year. The main purpose of this activity is to teach students to store and organize their work.

Teachers from different subjects can easily keep track of students' progress. Since the folder can contain a student's work for different subjects, it can also be used as a student's portfolio.

III. COMMUNICATION

A detailed description of the competence area communication is shown in Table III.

In all the mentioned activities, whether students are posting a forum content or communicating to each other through different online tools and platforms, students are encouraged to follow the netiquette rules.

The set of digital competences related to the communication competence area can be applied in the following way:

A. Subject forum

The teacher creates a subject forum (e.g. to create a subject forum, he/she uses ProBoards [3], Forumotion [4] or Lefora [5]). Afterwards, the teacher posts a question related to the current teaching unit. Each student is required to post an answer to that question (e.g. the answer can be in the form of a text, a picture or a video). After posting an answer, each student then needs to posts a comment for two other answers.

B. Distance teamwork

Using a teamwork project, students learn to collaborate and to meet deadlines through various online tools. Students are divided into teams (one team can be composed of students from different classes.). The teacher then creates one document, Google Sheet, and shares with all students. The tasks for each team are written in a separate tab of the Google Sheet document.

A task can vary from working on one shared document (e.g. Google Docs) to answer and post answers on the subject's forum.

In the last tab, the teacher creates a table to monitor the progress of each team. Each team leader must fill out the table with information on what team member has completed a given task and when (date). Each team member needs to complete one team task.

All students can see the tasks for each team and can monitor their own and other teams' progress. All teams have a time limit (e.g. five days) for each task. The team that finishes all tasks in time is given an opportunity to download an example for a written or oral subject's exam.

TABLE III. COMPETENCE AREA: COMMUNICATION

Competence	Definition
Interacting through technologies	To interact through a variety of digital devices and applications, to understand how digital communication is distributed, displayed and managed, to understand appropriate ways of communicating through digital means, to refer to different communication formats, to adapt communication modes and strategies to the specific audience.
Sharing information and content	To share with others the location and content of information found, to be willing and able to share knowledge, content and resources, to act as an intermediary, to be proactive in the spreading of news, content and resources, to know about citation practices and to integrate new information into an existing body of knowledge.
Engaging in online citizenship	To participate in society through online engagement, to seek opportunities for self-development and empowerment in using technologies and digital environments, to be aware of the potential of technologies for citizen participation.
Collaborating through digital channels	To use technologies and media for team work, collaborative processes and co-construction and co-creation of resources, knowledge and content.
Netiquette	To have the knowledge and know-how of behavioral norms in online/virtual interactions, to be aware of cultural diversity aspects, to be able to protect self and others from possible online dangers (e.g. cyber bullying), to develop active strategies to discover inappropriate behavior
Managing digital identity	To create, adapt and manage one or multiple digital identities, to be able to protect one's e-reputation, to deal with the data that one produces through several accounts and applications.

C. *Homework and/or project submission*

To learn and practice the use of email in business communication, students can be encouraged to submit their homework and/or class projects via email following the rules for email etiquette.

IV. CONTENT CREATION

A detailed description of the competence area content creation is shown in Table IV.

TABLE IV. COMPETENCE AREA: CONTENT CREATION

Competence	Definition
Developing content	To create content in different formats including multimedia, to edit and improve content that s/he has created or that others have created, to express creatively through digital media and technologies.
Integrating and re-elaborating	To modify, refine and mash-up existing resources to create new, original and relevant content and knowledge.
Copyright and licenses	To understand how copyright and licenses apply to information and content.
Programming	To apply settings, program modification, program applications, software, devices, to understand the principles of programming, to understand what is behind a program.

A. *Creating a video project*

Working together on creating a video, students establish their professional skills like collaboration, communication and critical thinking. Also, students deepen their knowledge about a specific course topic.

In this activity, each team's task is to develop and share a created video. First, the teacher prepares several topics, shares them with students via Google Drive and divides students into teams. Each team gets one topic.

For creation of the video project, students can use WeVideo [6], a cloud-based video creation platform, VideoToolbox [7] or Windows Movie Maker [8]. Each team is required to upload their work on YouTube and share a link to the created content with other teams. Every team needs to watch all videos, and decide on what video is the best and why. The team leader sends a review to the teacher. If several videos have the same number of votes, the teacher decides which team has won.

B. *Creating mind maps*

By creating a mind map, students can recall information more easily, explore and organize ideas and concepts.

At the end of each teaching unit, students can create a mind map on a topic that they have learned in the unit. To create mind maps, students can use various platforms such as XMind [9] and MindMap (Google Chrome plug-in) [10]. Also, students can be given a task to find already created mind maps and to complete them with additional information. Mind maps can be shared via Pinterest or the subject's web site.

C. Students creating course content for other students

Have students engage in creating course content to improve their learning. For example, students can be asked to write quiz questions and answers. Questions can be used as a repetition of the subject's topics.

Also, students can create video content on how to work with different computer programs: installing and configuring a program, solving a specific task and programming. This can be useful for subjects that teach programming (e.g. sql, java, c language) or working in a specific program (e.g. MySQL, MS Office, CADdy++ Electrical). Students can upload their videos to YouTube and share a link with other students via email or virtual classroom like Edmodo [11].

V. SAFETY

A detailed description of the competence area safety is shown in Table V.

TABLE V. COMPETENCE AREA: SAFETY

Competence	Definition
Protecting devices	To protect own devices and to understand online risks and threats, to know about safety and security measures.
Protecting personal data	To understand common terms of service, active protection of personal data, understanding other people's privacy, to protect self from online fraud and threats and cyber bullying.
Protecting health	To avoid health-risks related with the use of technology in terms of threats to physical and psychological well-being.
Protecting the environment	To be aware of the impact of ICT on the environment.

During the school year, students will create different accounts to access various social media sites, virtual classrooms, cloud storage and other platforms.

At the beginning of the school year, before students engage in creating their email, social media and other profiles, they should learn about their digital identity, footprints and how to create a strong password.

A. Exploring digital footprints

In this activity, one public person is assigned to each student. Students have to search the web and collect all information that they can find about that person.

Then they divide the found information in three groups: data that can be used by third parties (identity theft risk), data that can be seen in a negative context (by a prospective employer) and data that can be seen in a positive context (by a prospective employer). After that, students can try to find information about themselves.

At the end of this activity each student will answer the following question: "Are you monitoring your digital footprints and why?".

B. Creating a safe password

Before students create their email account for school purpose, they are asked to find the most common passwords used and explore the rules for creating a strong password.

Students create one strong password. Each student shares the created password with the rest of the class answering the following question: "What makes this password strong (or weak)?".

VI. PROBLEM SOLVING

A detailed description of the competence area problem solving is shown in Table VI.

TABLE VI. COMPETENCE AREA: PROBLEM SOLVING

Competence	Definition
Solving technical problems	To identify possible technical problems and solve them (from troubleshooting to solving more complex problems).
Identifying needs and technological responses	To assess own needs in terms of resources, tools and competence development, to match needs with possible solutions, adapting tools to personal needs, to critically evaluate possible solutions and digital tools.
Innovating and creatively using technology	To innovate with technology, to actively participate in collaborative digital and multimedia production, to express oneself creatively through digital media and technologies, to create knowledge and solve conceptual problems with the support of digital tools.
Identifying digital competence gaps	To understand where one's own competence needs to be improved or updated, to support others in the development of their digital competence, to keep up-to-date with new developments.

A. Collaborative problem solving

As an introduction to the teaching unit, students can participate in a problem solving activity. The teacher creates one Google Document that contains a problem which needs to be solved and shares the document with students.

Each student explores the web and uses online tools to visualize the problem (e.g. mind maps) and then organize activities that will lead to the solution. Students put their answers in the document. Then they need to read all answers from other students and agree on the solution to the problem. To agree on the solution, students can collaborate exclusively through the use of collaboration platforms (e.g. Slack [12], Trello [13], Google Apps [14]).

B. Updating digital skills

At the beginning of each school year, the teacher creates a list of online platforms that will be used for learning, creating content and collaborating. During the school year, students need to add to the list other

programs, online platforms or applications they discovered while working on various assignments.

VII. Conclusion

Each day we notice that many students as well as our colleagues lack digital competences. Digital competence is one of the eight key competences for lifelong learning. The other seven are: Communication in the mother tongue; Communication in foreign languages; Mathematical competence and basic competences in science and technology; Learning to learn; Social and civic competences; Entrepreneurship; and Cultural awareness and expression.

In order to become active participants in our digitalized society, we need to engage ourselves in lifelong learning.

The DIGICOMP portal is an easy to use e-learning portal that can help you learn different digital competences which can be used in everyday life and in the classroom. The main advantages of learning through the DIGICOMP portal are time and money savings (course participants can access materials from home or any other location at any time), coursework can be updated whenever needed, class work can be scheduled around personal and professional work, participants of the course can learn in their own pace.

All educators have a responsibility to develop and with time improve their digital competences and to include adopted knowledge, skills and attitudes in the classroom.

References

[1] DIGICOMP LOGIN. [Online]. Available: http://moodle.digital-competences-for-teachers.eu/index.php. Accessed: Feb. 1, 2016.

[2] DIGICOMP. [Online]. Available: http://www.digital-competences-for-teachers.eu/. Accessed: Feb. 1, 2016.

[3] ProBoards. [Online]. Available: https://www.proboards.com/create-free-forum. Accessed: Feb. 2, 2016.

[4] Forumoton. [Online]. Available: http://www.forumotion.com/. Accessed: Feb. 2, 2016.

[5] Lefora. [Online]. Available: http://www.lefora.com/. Accessed: Feb. 2, 2016.

[6] WeVideo. [Online]. Available: https://www.wevideo.com/highered. Accessed: Feb. 2, 2016.

[7] Video Toolbox. [Online]. Available: http://www.videotoolbox.com/ . Accessed: Feb. 2, 2016.

[8] Windows Movie Maker. [Online]. Available: http://windows.microsoft.com/hr-hr/windows/movie-maker. Accessed: Feb. 2, 2016.

[9] XMind. [Online]. Available: http://www.xmind.net/. Accessed: Feb. 2, 2016.

[10] Mind Map. [Online]. Available: https://www.mindmup.com/. Accessed: Feb. 2, 2016.

[11] Edmodo. [Online]. Available: https://www.edmodo.com/. Accessed: Feb. 2, 2016.

[12] Slack. [Online]. Available: https://slack.com/. Accessed: Feb. 2, 2016.

[13] Trello. [Online]. Available: https://trello.com/. Accessed: Feb. 2, 2016.

[14] Google apps. [Online]. Available: https://apps.google.com/ Accessed: Feb. 2, 2016.

Introducing collaborative e-learning activities to the e-course "Information systems"

Martina Asenbrener Katic, Sanja Candrlic and Martina Holenko Dlab
Department of Informatics, University of Rijeka, Rijeka, Croatia
masenbrener@inf.uniri.hr, sanjac@inf.uniri.hr, mholenko@inf.uniri.hr

Abstract - Progress in the development of information and communication technology and its wide availability offers many new possibilities to teachers designing e-courses. Besides blending face-to-face with online learning, trends in design assume blending various technologies (tools) as well as individual with collaborative learning. This paper presents changes in the design of undergraduate e-course *Information systems* that will be conducted in accordance with these trends, previous experience on the e-course and students' attitudes. By introducing additional e-learning activities, that will be used for formative and summative assessment, this research aims to increase students' motivation for learning the subject matter as well as to improve their final results (grades).

I. INTRODUCTION

Information and communication technology (ICT) constantly progress and its availability and applicability grows. As in other parts of life, it also offers many new possibilities for teaching process and learning process. In the context of course development, ICT provides the environment needed for transition from courses held in traditional classrooms to e-courses held online. At university level, teachers often combine face-to-face (f2f) learning in traditional classrooms with online learning. This approach is known as blended (hybrid or mixed) learning [1], [2]. According to the latest research in the field, beside combining online and traditional learning environments, blended learning assumes blending different technologies as well as different pedagogies in order to ensure effective learning [3].

Among technologies that can support realization of e-courses, Learning Management Systems (LMSs) have significant role. However, student-centered didactic models (such as case studies, project-based learning, inquiry-based learning, etc.) can also be enhanced by Web 2.0 tools which facilitate information sharing, interoperability, user-centered design and collaboration [4], [5], [6]. According to Lee and Woods [7], learning environment supported by ICT, that enables students to access and share knowledge and resources with one another, supports and stimulates the learning process and enhance learning performance. Therefore, e-course designers should consider the introduction of collaborative learning activities during which students will share and distribute expertise in order to achieve collective understanding and knowledge [8].

In the process of the design of e-course that includes collaborative learning it is not sufficient to use the same assignments (tasks) as in traditional classroom or to assign them to groups of students. Teachers have to be creative and plan various activities for students in order to enable acquiring learning objectives. Process of transition from traditional to student-oriented e-learning is demanding for the teacher [9] and it often requires education of the teachers and appliance of a certain methodology or a model [10], [11].

Collaborative e-learning activities, also referred as e-tivities [12], are oriented towards the completion of a given task. They emphasize interaction of student with available ICT but also with peers (colleague students). During the e-course development, it should be taken into account that e-tivities can be used for summative, but also for formative assessment. Formative assessment can be very useful for recognizing students' difficulties in learning, acquiring new knowledge and understanding new concepts. Formative assessment has potential to assure feedback on learning success to the student, but also to the teacher. According to the feedback student receives during formative assessment assignment, he/she will re-evaluate his/hers work and learning style or engagement. But this feedback is also very important for the teacher – based on the dynamics and quality of students' learning process, the teacher will adapt the teaching process accordingly [13].

Numerous examples of implementation of collaborative e-tivities can be found in literature. Some of them are based on appliance of a single method, such as collaborative writing or creating wikis [14], [15], mind-mapping [16], creating e-comics [17] etc., and some deal with combining different collaborative activities [18], [19], [20]. With the help of ICT, collaborative activities can be implemented even in global collaborative learning environment in which students from different countries collaborate on solving assignments, although they are not even sharing the same time zone [21]. It is indicated that collaborative activities encourage learning [22]. Participants involved in collaborative learning activities show higher level of motivation and through social involvement, higher level of responsibility for e-tivity success [12].

This paper presents a part of the ongoing research whose goal is to design contemporary course that follows trends in blended learning. After designing didactic model for *Information systems* course that included self-assessment tests for formative and summative assessment [23], the next step is to design collaborative e-learning

activities (e-tivities) that will ensure learning in an interesting but also effective way. In order to test students' attitudes towards collaborative e-tivities a survey was conducted. Students' opinions were taken into account during the process of planning new activities.

The paper is organized in the following way. After introduction in Section I, Section II gives an overview of the course structure. Section III describes research methodology. In Section IV authors analyze results for key grading elements of the course and propose course improvements. Section V gives conclusions and plans for future research.

II. COURSE STRUCTURE

A. Course context and content

The course *Information Systems* is a course held in winter semester at the Department of Informatics, University of Rijeka, Croatia. *Information systems* course is a mandatory course for students of Single major program of informatics, and elective for students of Double major program of Informatics, Mathematics, Physics and Polytechnic. Two teachers (professor and teaching assistant) participate in the realization of the course.

Some of the topics of the course are fundamentals of information systems (IS), models and stages of IS development, methodology of IS development, analysis and design of IS, development and software testing, implementation, application and maintenance of IS.

By completing this course, students gain basic knowledge in the field of information systems design that is necessary in the education of future professionals in this area [24]. The *Information systems* course stands as an introduction to a group of courses in the field of information systems. These courses, held later during the study, broaden subject meter, cover topics in more detail and ensure deepening of students' knowledge and experience. In addition to the theoretical knowledge that students gain during the *Information systems* course, students participate in a practical assignment that ensures application of theoretical knowledge. Students get the opportunity to know information systems of the real world.

The course combines (blends) traditional approach to teaching in the classroom, individual work outside the classroom and e-learning principles, using the learning management system MudRi. MudRi LMS is a customized solution adapted to the needs of University of Rijeka, based on the open source product.

B. Course activities

During this course, students are obligatory to attend lectures in f2f environment during which theoretical concepts are presented by professor. In addition, students should attend exercises where they have opportunity to solve practical assignments with the help of a teaching assistant. The course subject matter is divided into three main modules. Within each module self-assessment tests are available in the LMS during a predefined time period. After each module students write an online test – theory exam. During the whole semester, students work on a seminar paper which they submit for evaluation at the end of the semester.

Course activities workflow is shown in Fig 1. The most of activities is included in the summative assessment. Number of points for each activity is shown in the brackets next to the activity name. The final grade is formed according to the following scale:

- A - 80-100 points - excellent (5),
- B - 70-79.9 - very good (4),
- C - 60-69.9 - good (3),
- D - 50-59.9 - sufficient (2),
- E - 40-49,9 - sufficient (2),
- F – 0 - 39,9 – insufficient (1)

Students that get grade F have to retake the course in the next academic year.

Detail description of each activity follows.

Class attendance and class activity - During the semester teachers observe students' class attendance and try to estimate the level of students' activity during the f2f classes. According to their participation in class activities, students are rewarded with up to 5 points.

Self-assessment tests - During the semester students can solve self-assessment tests. Self-assessment tests are used to help students in learning but are not obligatory. Tests are carefully prepared to ensure formative assessment [23]. Students use them for review of lessons learned within each module. Each self-assessment test is available within a specified period, which is in advance known to the students. Students can access each self-assessment test only once and they can solve it from any chosen location (not necessary in the classroom).

Students can receive up to 3 points for each self-assessment test, depending on its correctness. On each self-assessment test students have to achieve a minimum of 40% points. Otherwise, points are not summed up with their other course points. The number of self-assessment tests in each module is different (it varies among one to two self-assessment tests per module). During the semester five self-assessment tests are available – students can receive up to 15 points for this activity.

Online tests – Theory exams - Students solve three online test – theory exam. Theory exams that are taken at the end of each module are in the form of online test with different question types. The tests cover general theoretical knowledge on concepts related to design and development of information systems (IS) and IS classification, IS users and IS application.

Previous research enabled teacher to define an optimal ratio of questions of objective versus subjective questions and the score they carry [25]. Online tests are set up following these rules: each test consists of 70% of objective question types (multiple choice, true/false and connections) and 30% subjective question types (short answer and essay questions).

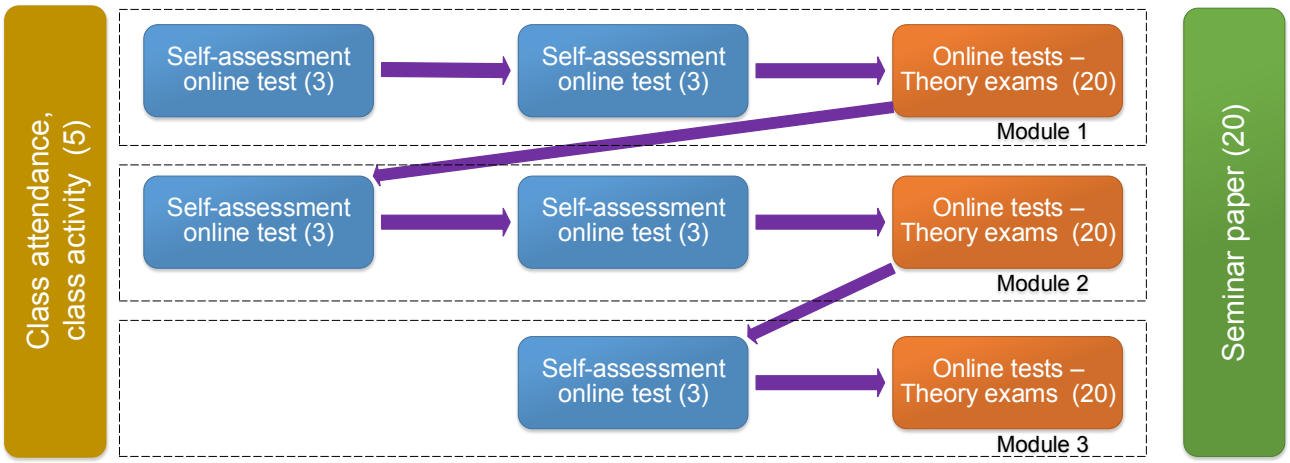


Figure 1. Activities workflow of the *Information system* course

In each test objective questions carry 50% of test score, and subjective questions carry the same percentage of test score. The passing threshold for theory exams is also 40%. In addition, students have to achieve a minimum of 40% at least on two out of three theory exams. All three tests are taken in the classroom under teachers' supervision. At the beginning of the semester, students are familiarized with the date of each theory exam.

Seminar paper - Seminar paper is a result of practical knowledge on analysis of information systems in different business organizations. During f2f exercises of the course students try to determine subsystems of information systems and to recognize relations among them. Students analyze documents that are used within the chosen information system and they analyze data on documents. They analyze the purpose and the structure of each data.

Students solve this assignment individually, over a longer period. Seminar paper is submitted at the end of the semester. During the preparation of the seminar students must respect defined deadlines.

Consultations - At any time, students can approach the teacher in order to ask a question or clear any remaining issue. They can use e-mail to communicate with the teacher, but also forum available within the MudRi LMS system. If they prefer f2f communication or their question requires a particular attention, students can attend f2f consultations which are held in the predefined time. Students can attend consultations individually, in pair or in a group.

III. METHODOLOGY

Process of introducing new activities to the course workflow included analysis of the course results for past several years as well as a survey regarding students' attitudes towards existing model and possible changes in the form of additional activities for formative assessment (collaborative e-tivities). During the course results analysis, the scores that students achieved in theory exams, seminar paper and final grade were analyzed. The analysis included the results from academic year 2011/2012 to 2014/2015.

Participants of the survey were students that attended the *Information systems* course in 2014/2015. Out of 112 students that attended the course, 87 students filled in the survey (77,67 %). The survey was taken at the beginning of the semester, during the first introductory lecture. Prior to the survey, the students were introduced with the course structure and with the points each activity carries. After the short insight, the survey was given to the students in the paper form and they filled in their answers in the classroom.

The survey is divided in three parts. In the first part students expressed their previous experience on collaborative e-tivities in which they participated during previous courses. In the second part of the survey students expressed their attitude towards given statements using the Likert scale. All 13 statements are shown in Table 1. The last part of the survey offered the comment section.

A. Survey results and discussion

Among collaborative e-tivities in which students participated before, most often listed answers were writing seminar papers and preparation presentations (individually or in groups).

Results of the central part of the survey are presented in Table 1. Percentages for scale elements are shown together with average result and standard deviation. Students are satisfied with the current model according to which they can collect up to 100 points during the semester (without the final exam).

In their opinion the course curriculum should, in addition to self-assessment tests for formative assessment before theory exams, contain other activities in order to allow them to collect points. They believe that e-tivities such as writing wikis, drawing diagrams and similar can also serve as a good way for preparing for the exams.

Some of them are less decisive weather they can learn something from their colleagues during an activity or not, achieve better success (grade) with the use of collaborative e-tivities, as well as regarding the use of Web 2.0 tools for the realization of tasks during e-tivities. However, most of students believe that such learning model could have a positive impact on their motivation.

The students are satisfied with the existing activities, but also see the benefits of introducing new ones that would preceded the theory exams and be included in the grading. During the implementation process, it will be necessary to point out the benefits and best practices of collaborative e-tivities to show to the students how they can profit from the knowledge of their colleagues and the usage of different tools.

B. *Course results analysis*

The Table 2 shows average values of the theory exams results, points for the seminar paper and final course results for the period from the academic year 2011/2012 to 2014/2015.

By analyzing these data, it can be observed that the final grades as well as students' points on theory exams and for seminar paper are in most cases decreasing (in average). The introduction of tests for self-assessment in 2014/2015 [23] ensured improvement of the tests scores. However, results are still quite low (average grade is E or D). Therefore, it can be concluded that additional elements of formative assessment should be introduced.

Since new didactic model aims to improve course results, but also to provide learning environment that will be in line with trends in the development of e-courses, suggested changes are twofold: improvements of self-assessment tests and introduction of new collaborative e-tivities for formative assessment.

C. *Suggested improvements of self-assessment tests*

During learning and revision of subject matter for each module students could benefit from multiple attempts for the self-assessment tests. The question remains, should the number of attempts be limited or not. Questions for self-assessment tests are randomized, and multiple access would require a question bank with a significantly larger number of questions, in order to avoid questions to re-appear in tests to many times. Therefore, in the beginning, the number of attempts should be limited to three. The number of attempts can be increased after question bank is amended.

TABLE I. QUESTIONNAIRE RESULTS

No	Statement	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	Av	SD
1	In your opinion, grading model without final exam is better for the course than the model where maximum of 70 points can be collected during semester and maximum of 30 points on final exam.	0,00	0,00	3,45	28,74	67,82	4,64	0,55
2	In your opinion, in addition to self-assessment tests for formative assessment, the course curriculum should contain other assessment activities (self-assessment tests, practical tasks, e-tivities).	2,30	6,90	12,64	33,33	43,68	4,10	1,03
3	In your opinion, assessment activities like self-assessment tests, practical tasks, e-tivities, organized before theory exams, are useful not only for collecting points but to get insight into your current level of knowledge.	1,15	2,30	9,20	43,68	43,68	4,26	0,81
4	You find important that before each theory exam there is at least one activity that can enable you to get feedback.	0,00	2,30	13,79	42,53	41,38	4,23	0,77
5	In your opinion, self-assessment tests can assure adequate feedback before theory exams.	3,45	2,30	11,49	35,63	44,83	4,19	0,98
6	In your opinion, in order to ensure adequate feedback before theory exams, self-assessment tests should be combined with practical tasks, especially for parts of the subject matter that cannot be automatically evaluated (e.g. diagramming).	4,60	10,34	24,14	32,18	28,74	3,70	1,13
7	In your opinion, participation in e-tivities such as creating wikis, diagrams and similar practical tasks can be efficient way of mastering the learning materials.	3,45	9,20	34,48	27,59	25,29	3,62	1,07
8	In your opinion, participation in e-tivities such as creating wikis, diagrams and similar practical tasks can be effective way of mastering the learning materials.	1,15	3,45	13,79	50,57	29,89	4,06	0,83
9	In your opinion, part of the course e-tivities should be designed to assure collaborative learning (in pairs or groups).	4,60	5,75	24,14	40,23	22,99	3,73	1,04
10	You would like to participate in collaborative e-tivities because you believe that you can learn from your peers.	5,75	8,05	26,44	32,18	26,44	3,66	1,13
11	You would like to use popular Web 2.0 tools that enable collaboration (Google Docs, tools for drawing diagrams, brainstorming tools, …) for e-activities.	4,60	12,64	26,44	36,78	19,54	3,54	1,09
12	In your opinion, the participation in collaborative e-tivities can positively influence your final grade for the course.	1,15	5,75	34,48	33,33	25,29	3,76	0,94
13	In your opinion, the participation in collaborative e-tivities can positively influence your motivation during the course.	0,00	4,60	29,89	39,08	26,44	3,87	0,86

TABLE II. COURSE RESULTS

	2011/2012	2012/2013	2013/2014	2014/2015
Number of students	54	56	63	112
1st theory exam:	65,59%	55,48%	46,46%	52,78%
2nd theory exam:	59,41%	58,96%	40,27%	57,13%
3rd theory exam:	48,61%	50,71%	48,52%	49,34%
Seminar paper:	89,31%	82,2%	82,8%	81,85%
Final grade:	68,29	64,54	61,10	62,53

Several attempts for solving self-assessment tests also impose the decision on which attempt should be taken into consideration for grading. Suggested method is *Highest grade* (the final grade is the best grade in any attempt).

Authors also suggest to lower the number of points for each self-assessment test, from currently 3 points to 1 point, regardless of the number of attempts, but only if at least one attempt has at least 40% correctness. In that case, student could receive 0 or a number of points from the segment [0.4, 1] for each self-assessment test.

In the current model, the date for solving self-assessment tests is predefined, but the new didactic model should allow flexible access to self-assessment test throughout the duration of the entire module.

D. Suggested collaborative e-tivities for implementation

It is suggested to introduce a couple of additional collaborative e-tivities to encourage efficient learning and ensure formative assessment through forum discussions and collaborative writing.

One forum discussion should be introduced. By using forum in the LMS, students could share their knowledge on information systems topics. Discussion should be initiated by teachers to support students' critical thinking and expressing their attitudes. Active participation should be included in the grading (up to 5 points).

Students will be divided (based on their own preferences) in groups that consist of 4-5 members in order to write a shorter paper on a given subject using Google Docs - Web 2.0 tool for collaborative writing. Students will be able to suggest a subject according to their interest (but aligned with course subject matter) or they will choose one of the predefined topics. At the end of the e-tivity, students will share documents among teams. It is expected that this e-tivity will contribute to a better understanding of theoretical concepts and therefore is planned before 1st theory exam. Students could gain up to 10 points on this e-tivity.

Another collaborative e-tivity will be group analysis of several documents from an information system, as an introduction to individual seminar paper. By using a Web 2.0 tool Wikispaces that supports collaborative writing in a form of wiki, students will learn the basic concepts, analyze documents and determine subsystems of information system and their connections. Additional support to team collaboration and communication in Wikispaces is given through the discussion forum that is visible to team members. Students could gain up to 15 points on this e-tivity.

Activities workflow and number of points for each course activity after introducing new collaborative e-tivities is shown in Figure 2. Current course activities (class attendance and activity, online tests at the end of each module and seminar paper) should remain the same. However, it can be seen that the grading for these activities should be slightly changed.

IV. CONCLUSIONS AND FUTURE WORK

This paper describes a part of an ongoing research focused on developing a contemporary blended learning course. After implementation of online self-assessments in order to ensure formative assessment, introduction of additional activities is proposed.

Course results analysis showed the need for additional activities for formative assessment that will precede theory exams in order to ensure that students achieve better course results and consequently a higher level of learning outcomes. In addition, conducted survey indicated that students realize the benefits of collaborative learning and activities for formative assessment. Based on that, new didactical model for the e-course Information systems which includes several collaborative e-tivities is proposed.

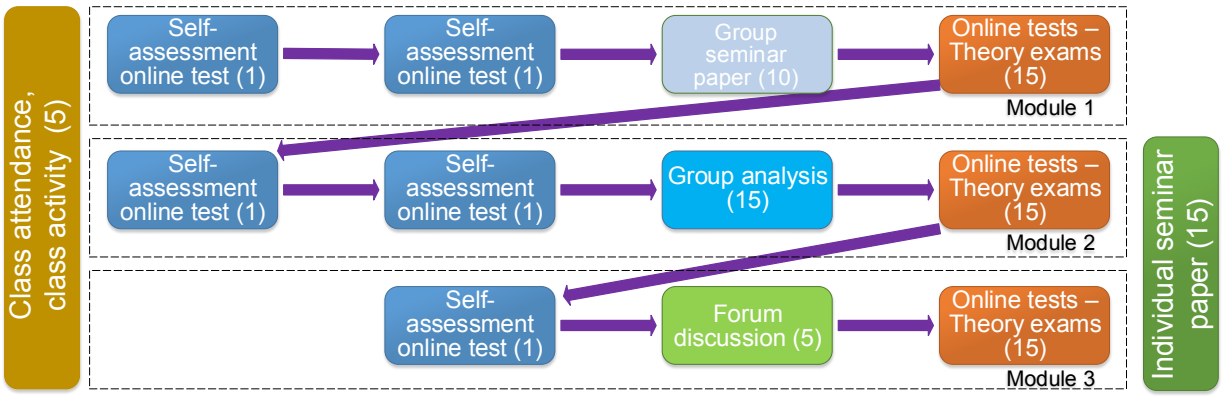


Figure 2. Activities workflow of the Information system course after introducing new collaborative e-tivities.

There is an evident increase in the number of students participating in this course so the increase in workload of teachers that will need to review the results of activities and give feedback should be taken into account. Therefore, the number of new activities in not too big (three new e-tivities are included in the workflow). These activities are in line with the course content and precede theory exams in order to fulfill the role of formative assessment. Planned activities assume collaboration among students and therefore will be performed using adequate tools – discussion forum in the LMS and tools available on the Web (Web 2.0 tools).

Further research will include evaluation of the didactic model using a combination of a comparative study and a survey. Evaluation will be performed during the following academic year. Since proposed improvements are in line with the current trends in the development of e-courses as well as with students' opinions, it is expected that introduced e-tivities will positively influence students' motivation for learning and their results on main grading components - theory exams and seminar paper.

ACKNOWLEDGMENT

The research has been conducted under the project "Extending the information system development methodology with artificial intelligence methods" (reference number 13.13.1.2.01) supported by University of Rijeka (Croatia).

REFERENCES

[1] C. R. Graham, W. Woodfield, and J. B. Harrison, "A framework for institutional adoption and implementation of blended learning in higher education," *Internet High. Educ.*, vol. 18, pp. 4–14, Jul. 2013.

[2] N. Hoic-Bozic, M. Holenko Dlab, and E. Kusen, "A Blended Learning Model for 'Multimedia Systems' Course," in *Advances in Intelligent Systems and Computing, Workshop on Learning Technology for Education in Cloud (LTEC'12)*, 2012, pp. 65–75.

[3] N. Hoic-Bozic, M. Holenko Dlab, and V. Mornar, "Recommender System and Web 2.0 Tools to Enhance a Blended Learning Model," *IEEE Trans. Educ.*, vol. 59, no. 1, pp. 39–44, Feb. 2016.

[4] D. Benito-Osorio, M. Peris-Ortiz, C. R. Armengot, and A. Colino, "Web 5.0: the future of emotional competences in higher education," *Glob. Bus. Perspect.*, vol. 1, no. 3, pp. 274–287, May 2013.

[5] N. Hoic-Bozic, M. Holenko Dlab, and J. Mezak, "Using Web 2.0 tools and ELARS Recommender System for E-Learning," *Int. Conf. e-Learning*, vol. 14, pp. 207–212, 2014.

[6] J. Silva, A. Astudillo, and E. Fernandez, "Student-centered virtual anvironment proposal," *Int. Educ. Res. J.*, vol. 1, no. 5, pp. 64–68, 2015.

[7] S. E. Lee and J. Woods, "Using contemporary topics and Internet resources to stimulate student-centred learning," *Australas. J. Educ. Technol.*, vol. 26, no. 6, pp. 775–790, 2010.

[8] M. Romero, M. Usart, M. Ott, and J. Earp, "Learning through playing for or against each other? Promoting collaborative learning in digital game based learning," in *20th European Conference on Information Systems*, 2012.

[9] J. Gregory and G. Salmon, "Professional development for online university teaching," *Distance Educ.*, vol. 34, no. 3, pp. 256–270, 2013.

[10] G. Bubas, A. Coric, and T. Orehovacki, "Strategies for Implementation of Web 2.0 Tools in Academic Education," in *Proceedings of the 17th European University Information Systems (EUNIS) International Congress*, 2011, pp. 1–17.

[11] G. Salmon and P. Wright, "Transforming Future Teaching through 'Carpe Diem' Learning Design," *Eduation Sci.*, vol. 4, no. 1, pp. 52–63, 2014.

[12] G. Salmon, *E-tivities: the key to active online learning*. Psychology Press, 2002.

[13] L. Dunn, C. Morgan, M. O'Reilly, and S. Parry, *The student assessment handbook: New directions in traditional and online assessment*. Routledge, 2003.

[14] R. Aztalos, "The use of a wiki at a college in Hungary as a tool to enhance personal learning.," in *CALL Design: Principles ad Practice, Proceedings of the 2014 EUROCALL Conference*, 2014, pp. 18–22.

[15] T. A. Thomas, "Developing team skills through as collaborative writing assignment," *Assess. Eval. High. Educ.*, vol. 39, no. 4, pp. 479–495, 2014.

[16] C. O'Reilly, "Mind Mapping: Overcoming Problems of Writer Identity and Convention for Academic Writing by Student Collaboration," *All Irel. J. Teach. Learn. High. Educ.*, vol. 7, no. 3, 2015.

[17] F. Lazarinis, A. Mazaraki, V. S. Verykios, and C. Panagiotakopoulos, "E-comics in teaching: Evaluating and using comic strip creator tools for educational purposes," in *Computer Science & Education (ICCSE), 2015 10th International Conference on*, 2015, pp. 305–309.

[18] J. Mezak, N. Hoic-Bozic, and M. Holenko Dlab, "Personalization of e-tivities using Web 2.0 tools and ELARS (E-learning Activities Recommender System)," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2015, pp. 669–673.

[19] T. Orehovacki, M. Konecki, and D. Radosevic, "Web 2.0 technologies in university education.," in *31st MIPRO International Convention on Computers in Education*, 2008, pp. 25–29.

[20] M. U. Paily, "Creating constructivist learning environment: Role of 'Web 2.0' technology," *Int. Forum Teach. Stud.*, vol. 9, no. 1, pp. 39–51, 2013.

[21] C. Susilawati, "Blended learning extension in a global collaborative learning environment," in *21st Annual Pacific-Rim Real Estate Society Conference*, 2015.

[22] B. M. Hall, "Designing Collaborative Activities to Promote Understanding and Problem-Solving," *Int. J. e-Collaboration*, vol. 10, no. 2, pp. 55–77, 2014.

[23] M. Holenko Dlab, M. Asenbrener Katic, and S. Candrlic, "Ensuring formative assessment in e-course with online tests," in *10th International Conference on Computer Science & Education (ICCSE)*, 2015, pp. 322–327.

[24] M. Pavlic, M. Marinovic, and S. Candrlic, "Model of Basic Training for Information Systems Developers," *8th Int. Symp. Oper. Res. Slov.*, 2005.

[25] S. Candrlic, M. Asenbrener Katic, and M. Holenko Dlab, "Online vs. Paper-Based Testing: A Comparison of Test Results," in *Proceedings of the 37th International Convention MIPRO 2014*, 2014, pp. 775–780.

A Curriculum for Unified Embedded Engineering Education

Ivan Kastelan, Miodrag Temerinac
University of Novi Sad
Faculty of Technical Sciences
Novi Sad, Serbia
ivan.kastelan@rt-rk.uns.ac.rs

Abstract - Tremendous progress in semiconductor technology enabled broad proliferation of computer-based solutions in almost all fields, from medical instrumentation up to spacecraft technology. This development opened the question about a new engineering profile being able to act in research and development of embedded systems. In this paper a contribution to the structuring of undergraduate embedded engineering education is given. The curriculum is largely based on the results of the FP7 project "Embedded Engineering Learning Platform" (E2LP) and proposes a unified approach to undergraduate education in which most courses covering the field of embedded engineering use the same platform for laboratory exercises. The paper concludes with the case study where this concept is partially applied, at the Faculty of Technical Sciences in Novi Sad.

I. INTRODUCTION

The E2LP project [1] concerns a novel development of a unified learning platform for embedded system design, which would serve as a general educational framework for future embedded system engineers.

This approach targets the lab education efficiency with the idea to use a single comprehensive platform for the complete curriculum. The main intention is to make the educational process more efficient and to introduce more interaction between the education and further embedded system research and development, which facilitates an optimal solution for a specific problem. The main E2LP objective is to efficiently educate future engineers using up to date technologies capable of coping with current challenges in real-time embedded computer engineering field.

Focusing the necessary changes in computer engineering and computer science education, we are meeting the next dilemma – what are generic principles which should be covered in the embedded engineering education for variety of using fields. An intuitive approach is to ask industry about needed engineering profiles where educated engineers will act for their lifetime. A joint work between academy and industry certainly improves engineering education programs as it is mentioned in [2] and [3]. But, this approach is facing an issue – the imbalance between technology cycle times and working life times. When technology cycle times, like in the past, are longer than the people working times, industry needs could perfectly define requirements for education which takes more than 10 years (about 12 years for basic and high school education plus about 5 years engineering education). Today, technology cycles are even shorter than the engineering education time, which means that the industry needs today could only express educational needs in the past [4]. Therefore, an interactive approach between industry and academy with improving loops inside education times will be needed. A concurrent evaluation of systems and processes mentioned in [5] could be a right answer to this challenge. Even an early impact during basic and high school education could help to overcome this race between education and development [6], [7] and [8].

The main contribution of the E2LP project is the unified learning platform which is aimed to be used in the majority of courses within the embedded computer engineering curriculum. Through its extension boards and its functional extensions for assisting students (augmented reality interface and remote laboratory) [9], the platform represents a complete solution for laboratory exercises in computer and embedded engineering related courses. With E2LP platform, students are not required to spend many hours learning how to use multiple platforms and go through tutorials; instead they can concentrate immediately on the material of the course they are taking. This significantly reduces the time effort and educational overhead for preparation for the course and extends available time for learning important concepts and reaching new skills.

During the evaluation year, platform was extensively used in multiple universities and evaluation data was gathered from students and teachers using the platform. In addition, augmented reality and remote laboratory went through the pilot-usage phase and were fine-tuned for the final release. One of the major outputs of the final year, which is based on the work in the complete project, from methodology definition, through implementation and evaluation data is the proposal for the embedded computer engineering curriculum based on laboratory exercises with E2LP platform. This paper gives the proposal and presents how it fits with existing curricula at partner universities, as well as how it satisfies learning objectives for an embedded computer engineer.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 317882 and Ministry of Science, Education and Technological Development of the Republic of Serbia, project nº TR32030.

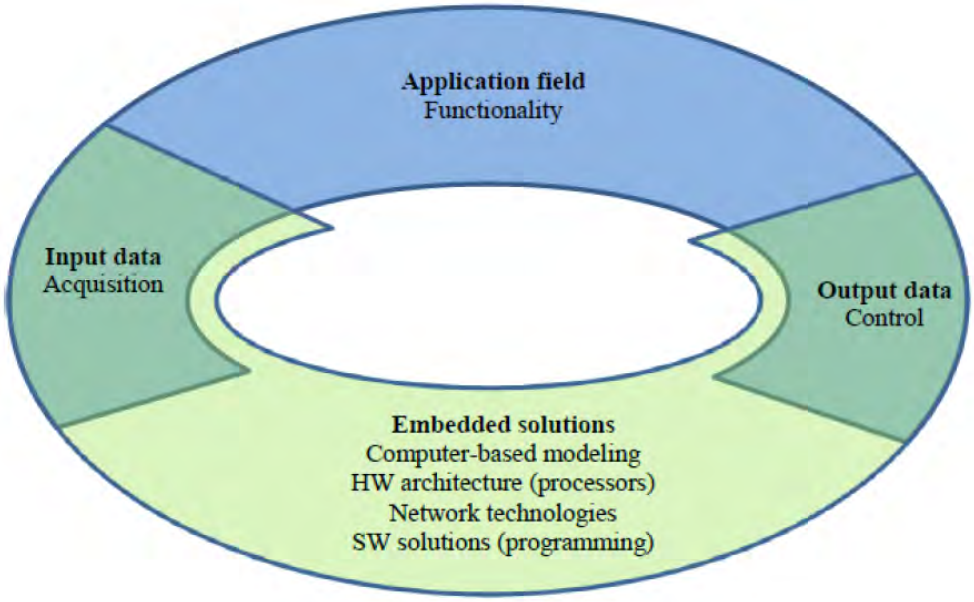


Figure 1. Generic model of embedded solutions

II. LEARNING OBJECTIVES

Based on the analysis of the requirements for the career and professional skills of an embedded computer engineer, after consulting the ACM/IEEE curriculum requirements for computer engineering [10], E2LP project defined the following learning objectives which need to be satisfied with the curriculum and the laboratory tools:

LO #1. Algorithms and programming (software aspects),

LO #2. Embedded microprocessors and computer architectures (hardware aspects),

LO #3. FPGA digital system design, system integration and verification (system aspects),

LO #4. FPGA accelerated computing (improvement aspects),

LO #5. Digital signal processing (audio, video and data) and its real-time implementation (processing aspects),

LO #6. Networks and interfaces (communication aspects).

Furthermore, each exercise within the course should fit into one of the following categories [11]:

1. Exercises
   a. Remembering (Bloom's Taxonomy Level 1),
   b. Understanding (Bloom's Taxonomy Level 2),
2. Problems
   a. Applying (Bloom's Taxonomy Level 3),
   b. Analyzing (Bloom's Taxonomy Level 4),
3. Projects (Bloom's Taxonomy Level 5)
   a. Problem solving,
   b. R&D support.

However, it is important to note that engineering education is not only about developing students' knowledge and skills in the cognitive domain, as mentioned above. Computer engineering education has also to do with fostering students' meta-cognitive competences, for example awareness to fast learning and reflection on new knowledge sphere, and promoting motivational aptitudes and innovational thinking. When designing the curriculum for embedded engineering education, a generic model of embedded solutions would be very helpful to structure the embedded engineering education. Variety of embedded applications makes this abstraction very demanding. Our approach, shown in Fig. 1, could be a useful contribution in this direction.

The first step of embedded system design is modeling the given problem including required functionality and available input/output interfaces. Additionally, the frame conditions (costs, power consumption, security etc.) are considered by the design of an appropriate hardware platform. This platform, depending on requirements and conditions, can vary from the simplest processor architecture (8-bit microcontroller for car window engine) up to complex multi-core solutions (system-on-chip for TV set-top boxes). In distributed applications (e.g. cloud computing) the corresponding networking concept has to be added. The required functionality should be implemented in software (SW) programs. Variety of the used hardware (HW) platforms covers programming from the lowest level (assembler and HW drivers) up to the highest level using complex SW tool environment (Windows, Linux, Android, iOS, etc.). Finally, each embedded solution has to be tested and verified. The assessment of embedded solutions includes methodology and appropriate tools. Additional requirement is that the assessment of solutions provides some metrics which can be used in a feedback improvement loop.

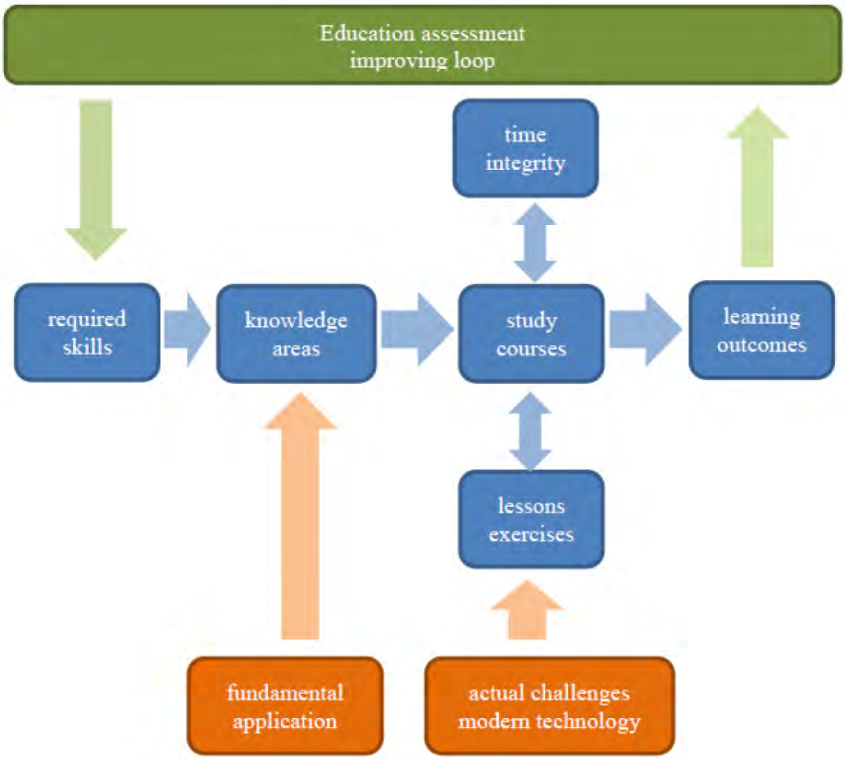


Figure 2. Embedded engineering education process

III. Methodology of Curriculum Design

We propose a structural approach for engineering curriculum using formal expressions of all identified items in the education process. Additionally, specifics of embedded engineering education will be considered.

The engineering education process should be structured using formal approaches. In this section we propose some improvements in the formal approach of engineering education considering specifics of embedded engineering. The embedded engineering education process could be formally expressed using the flow chart in Fig. 2.

Definition of an engineering study curriculum is based on the set of required skills and deviated knowledge areas. They are covered by study courses using lesson-exercise structure. The course structure has to fulfill two time integrity conditions: inter-dependence (sequential building up of knowledge) and limited student load (study time limitations). The improving loop should be closed using some assessment metrics for continuous adaptations of the study curriculum. Specific requirements for embedded engineering education could be expressed in two additional steps: the knowledge areas should include basic understanding of application field (interdisciplinary approach and reusing approach) and lessons should be refreshed by actual challenges accompanied with exercises using modern technology. The set of measurable learning outcomes in education assessment should be extended with soft skills (project management, teamwork, innovation capacity).

Structural approach requires an appropriate numeric metrics. We are proposing a three-degree metrics:

- F – familiarity (understanding knowledge),
- U – usage (implementation capability), and
- A – acquired knowledge (deep understanding, capability of innovative reusing).

This metrics is used in definitions of all matching matrices, between skills and knowledge areas, between knowledge areas and study courses and between study courses and learning outcomes.

The required skills are coming from industry and often they are not well-structured. Identified in company operations (marketing, development, sales & consumer support) they could be generally sorted in three groups:

- P – predefined (established) engineering profiles like software engineering, computer engineering, mechanical engineering, chemistry etc.,
- S – specific knowledge like car engine, sensors, fluid theory etc., and
- C – concrete methods and tools like programming in Linux environment, CAN protocols, GNU editors, etc.

The list of knowledge areas are coming for university in a structured way. They are already identified and well-defined areas from research and education. Generally, they could be grouped in three main groups:

- B – basics covering high-abstract-level knowledge like mathematics, physics, etc.,

- K – fundamental knowledge like communications, computer architectures, programming, imaging, etc.,
- E – elective fields covering specific knowledge, methods and tools like network protocols, operating systems, compilers, etc.

The matching matrix between skills and knowledge areas contains dependencies expressed in the proposed metrics values. In practice, the skill list includes summarized needs coming from more companies.

A study curriculum is defined by study courses. They should be defined as learning-sustainable units respecting previous education experience. Courses are subdivided in lessons accompanied with appropriate practical exercises. The course construction is primary allocated at universities. But, regarding specifics of embedded engineering education, industry needs and experiences should be also incorporated in the definition of study courses. This influence of industrial experiences is especially important for incorporation of soft skills (teamwork, project management) in engineering education.

The matching matrix between knowledge areas and study courses indicates coverage of a knowledge area in a study courses. The achieved learning outcomes are indicated regarding lessons and exercises of all study courses. Standardly, each course yields more learning outcomes. This dependence is defined by a matching matrix between courses and learning outcomes.

Finally, the assessment matching matrix defines coverage of requested skills by achieved learning outcomes (0 – not covered, 1 – covered but not by requested degree and 2 – covered by requested degree). This matrix could be used to check efficiency of the education process regarding the requested engineering profiles. The average of the assessment matching matrix values could be used as the global efficiency indicator of an education program for the given skill requirements.

Such structured approach in construction of engineering curricula additionally requires a sustainable timing order of defined courses, interdependences between the defined courses must be timely considered. Also, necessary efforts (hours per lesson, hours per exercise) must be aligned with a normal load of students (approximately 40 hours per week). These frame conditions define the timing of a study program (semesters and courses).

IV. Case Study – Curriculum of Embedded Engineering based on Unified Education

In this section, we propose a curriculum for 4-year undergraduate Bachelor of Science (BSc) studies of embedded computer engineering which is based on using E2LP platform as a laboratory tool in most of the courses, following the main idea of the E2LP project. We follow the methodology of curriculum design outlined in the previous section, starting with the definition of required skills for an embedded computer engineer, from which we define knowledge areas which need to be covered, courses and learning outcomes.

The following skills are identified for an embedded engineer:

P (predefined):

1. computer engineering
2. computer science

S (specific):

1. programming
2. hardware design
3. system design

C (concrete):

1. C/C++
2. Java
3. VHDL
4. Assembly programming
5. Linux (embedded Linux)
6. Android
7. MATLAB

The following knowledge areas are identified for an embedded engineer:

B (basic):

1. mathematics
2. physics
3. social sciences

K (fundamental knowledge):

1. computer systems
2. software
3. embedded systems

E (elective):

1. networks
2. digital signal processing
3. system software

As a conclusion, table 1 gives curriculum outline across four years of undergraduate studies. For each course, number of credits is given, with accordance to the European Credit Transfer System (ECTS) where 1 credit corresponds to 25-30 hours of work, including lectures, recitations, laboratory exercises, readings, homework, studying and exams. Total number of credits is 240 for the entire curriculum, 60 per year.

Table 2 is the assessment matrix which checks how the learning outcomes match with skills required by an embedded engineer (A matrix). This matrix uses 0-1-2 values.

TABLE I. EMBEDDED ENGINEERING CURRICULUM

Course / Semester	E2LP	1	2	3	4	5	6	7	8
English language 1		3							
Calculus 1		9							
Discrete mathematics and linear algebra		9							
Programming languages and data structures		9							
English language 2			3						
Physics			9						
Fundamentals of electrical engineering			9						
Computer architecture	used		9						
Calculus 2				8					
Object-oriented programming				8					
Fundamentals of system control				8					
Electronics				6					
Social aspects of technological development					3				
Probability and stochastic processes					5				
Digital system design	used				8				
Operating systems					8				
Topics of hardware design	used				6				
Computer system design	used					6			
Real-time system software 1						6			
Embedded system modeling and simulation						6			
Digital signal processing 1	used					6			
Computer networks	used					6			
Laboratory of computer engineering	used						6		
Real-time system software 2	used						6		
Digital signal processing 2	used						6		
Wireless networks	used						6		
Programming apps for Android	used						6		
Advanced real-time software								7	
Algorithms in digital signal processing								8	
Advanced embedded systems laboratory	used							8	
Advanced computer networks								7	
Engineering of computer based systems									8
Multimedia architectures and systems	used								8
Practice work in industry									2
Bachelor of science thesis									12

TABLE II. ASSESSMENT MATRIX IN CURRICULUM DESIGN

A matrix		LO #1	LO #2	LO #3	LO #4	LO #5	LO #6	Broad ed.
P	Computer engineer.	2	1	2	1	2	2	0
	Computer science	2	0	0	1	1	1	0
S	Program.	2	2	0	0	2	1	0
	Hardware design	0	0	2	2	0	1	0
	System design	1	1	1	1	1	2	0
C	C / C++	2	2	0	0	2	1	0
	Java	1	1	0	0	1	1	0
	VHDL	0	0	2	2	0	1	0
	Assembly	2	2	1	1	1	1	0
	Linux	2	0	0	0	1	1	0
	Android	2	0	0	0	1	1	0
	MATLAB	0	2	0	0	0	1	0

E2LP platform is used during the majority of the curriculum. Most courses from learning objectives (1)-(6) use the platform. E2LP start-up kit [12] contains a starting set of laboratory exercises for each of the 13 courses which use the E2LP platform.

In summary:

- 13 courses use E2LP platform,
- 10 courses use PC and general-purpose programming tools (C/C++, Java, MATLAB),
- 3 courses have laboratory exercises outside computer science and engineering,
- 7 courses in broader education category are without laboratory exercises.

V. CONCLUSIONS

In this paper we proposed a curriculum for an undergraduate study program for embedded engineering. This curriculum is highly based on laboratory exercises on E2LP platform, one of the main products of E2LP project. Majority of the courses in the field of computer engineering which have laboratory exercises use E2LP platform and all its benefits in reducing the time for tutorials and preparation and concentrating on the course materials from the very beginning of the course. Some courses use the benefit of education continuity and re-using the results from the labs of the previous course which is an added value of the E2LP platform, e.g. computer system design is a continuation of digital system design.

The proposed curriculum covers all 6 learning objectives identified for embedded engineering. Additionally it gives enough broad education required during undergraduate studies. Besides labs on E2LP platform, it assumes PC labs with general-purpose programming tools mostly for topics which are closer to computer science than computer engineering. Out of 23 courses which have lab exercises in the computer science and engineering field, 13 use the unified platform and 10 use general-purpose programming tools. Remaining 10 courses are either theoretical or with physics-based labs.

This proposal may serve as a guide on curriculum design as it provides a methodology for the curriculum design which is aware of both academic and industry needs. This curriculum, combined with the basic set of exercises from E2LP project, presents a complete solution for computer engineering education which may be used to help in designing the curriculum or be used entirely or in part as a source of materials for laboratory exercises in courses which can be run on E2LP platform.

The study program for computer engineering and computer communications of the Faculty of Technical Sciences in Novi Sad, Serbia (FTN) follows some of the proposals set in this paper. This paper goes further to propose more embedded oriented curriculum which uses the same platform in majority of the courses in the computer engineering field. Credit points are proposed following the European ECTS credit system, for a total of 240 credit points for the complete program.

Even though it is impossible to apply the curriculum immediately, due to the accreditation procedures, FTN will use this proposal to guide the changes for the next accreditation process. Some other universities (University of Banja Luka and University of Osijek) already confirmed to use parts of the materials for laboratory exercises to apply them in their current study programs in computer science and engineering.

REFERENCES

[1] I. Kastelan, J. R. Lopez Benito, E. Artetxe Gonzalez, J. Piwinski, M. Barak, M. Temerinac: "E2LP: A Unified Embedded Engineering Learning Platform", Microprocessors and Microsystems, Elsevier, 2014, Vol. 38, Issue 8, pp. 933-946

[2] J. C. Cunha, J. P. Amaro, L. Marques, "A joint academy-industry initiative for the development of an engineering program", 1st Int. Conf. of the Portuguese Society for Engineering Education, 2013

[3] M. Pokkiyarath, R. Raman, K. Achuthan, B. Jayaraman, "Preparing global engineers: USA-India academia & industry led approach", IEEE, 2014

[4] D. Ktoridou, N. Eteoleous, "Engineering education: time to reform the fragmented, content-overloaded curricula context?", IEEE Global Engineering Education Conference EDUCON, 2014, pp. 377-380

[5] G. Karsai, F. Massacci, L. J. Osterweil, I. Schieferdecker, "Evolving embedded systems", IEEE Computer, May 2010, pp. 34-40

[6] G. MacBride, E. L. Hayward, G. Hayward, E. Spencer, E. Ekevall, J. Magill, A. C. Bryce, B. Stimpson, "Engineering the future: embedded engineering permanently across the school-university interface", IEEE Transactions on Education, Vol. 53, No. 1, Feb. 2010, pp. 120-127

[7] R. Bruce, D. Brock, S. Reiser, "Teaching programming using embedded systems", IEEE, 2013

[8] M. K. Banavar, D. Rajan, A. Strom, P. Spanias, X. Zhang, H. Braun, A. Spanias, "Embedding Android signal processing apps in high school math class – an RET project", IEEE, 2014

[9] J. R. Lopez Benito, E. A. Artetxe Gonzalez, M. Anastassova, F. Souvestre: "Engaging Computer Engineering Students with an Augmented Reality Software for Laboratory Exercises", IEEE Frontiers in Education Conference (FIE), 2014, pp. 1-4

[10] ACM/IEEE-CS Joint Task Force on Computing Curricula, "Computer engineering curricula 2004'', 2004

[11] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, D. R. Krathwohl: "Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain". New York: David McKay Company, 1956

[12] E2LP start-up kit – catalogue of laboratory exercises, http://moodle.e2lp.org/ (accessed on 19 Feb 2016)

Individual versus collaborative learning in a virtual world

Paul Pürcher[1], Margit Höfler[1], Johanna Pirker[2], Lisa Tomes[2], Anja Ischebeck[1], Christian Gütl[2, 3]
[1] University of Graz/Department of Psychology, Graz, Austria
[2] Graz University of Technology/Institute for Information Systems and Computer Media, Graz, Austria
[3] Curtin University, Perth, Western Australia

E-Learning in general and, specifically, learning in virtual worlds has become more and more popular in recent years. In the present study we investigated whether learning is more successful if students learn in a virtual world individually or collaboratively. We also assessed the motivation of the participants and the usability of the virtual environment. The findings showed a tendency that learning was more successful when participants learned collaboratively than individually. Furthermore, their interest (current motivation) decreased during the learning session when they learned individually and increased when they learned in a group. This experiment shows that, in virtual worlds, collaborative learning seems to be more effective and more engaging than individual learning.

I. INTRODUCTION

Nowadays learners can choose between varieties of new learning environments that allow them to become familiar with the learning material. Many of these new learning environments are related to mass media and e-learning [1]. Much research has already been conducted regarding the potential facilitation of learning through the use of these new learning techniques. In general, it has been shown that e-learning makes learning easier in this new era of education [2] and that it has changed the learning behavior of students dramatically, towards more self-directed activities [3]. However, the traditional teaching-learning-models are usually implemented in e-learning settings without any adaption to the specific possibilities of e-learning environments [4].

An important factor pertaining to the learning outcome is whether learning happens individually or in a cooperative/collaborative setting. Previous research has indicated that there are benefits of collaborative learning [5], especially if all group members contribute during the activity and are invited to share their own experience and knowledge. Such collaborative learning (CL) typically leads to an improved learning performance and an increase of knowledge, in comparison to individual learning [5]. Furthermore, CL may even enhance the prosocial behavior and sympathy between group members [6, 7] as well as the social self-efficacy of the learners [8].

However, in contrast to a "typical" collaborative learning setting in which learning occurs face-to-face, in an e-learning setting, several additional factors have to be taken into account. For instance, in e-learning settings it is difficult to ensure that each group member understands the objectives in the same way, since the face-to-face communication habits, which many people are accustomed to, cannot be applied in such situations [9]. Hence, online group-learning may lead to insufficient group organization and this can lead to serious problems regarding learning [10, 11]. On the other hand, the online learning environment may also result in a change of the social relations in the group, of the expectations which the learners have towards themselves [9] and may also result in an equal or better participation rate [12]. For instance, in a recent study we had psychology students work collaboratively on a short paper using an online co-writing wiki system. During the assignment, continuous (visual) feedback about the progress of the work and the activities of each group member was provided, and the group members could also assess their own work activities and those of their peers. The findings showed that the students found the feedback tools useful for having an overview about the actual status of their paper and the contribution of team members [13].

As mentioned above, research has indicated the benefits of collaborative learning, in comparison to individual learning [5] which are not only reflected in high learning performance, but also in improved prosocial behavior and increased sympathy between group members [6, 7]. The aim of the current paper is to have a closer look at these findings, and to test and extend them to the field of e-learning in 3D virtual worlds (VW). Such VW have become more and more popular for online learning in recent years. VW enable users to work and learn in a collaborative online environment. They foster interactions with other avatars (social interactions, collaboration, roleplaying scenarios), the virtual environment (creation, exploration, and visualizations), and support the use of different learning models. VWs support new forms of distance learning and together with the different possibilities for interaction and new auditory and visual channels, VWs are considered to be a "knowledge and social tool" [14, 15].

In the present paper, we were interested in whether learning in a VW is indeed more effective when learners work collaboratively rather than individually, and whether learners that work collaboratively are more motivated during the assignment. To this end, we created a VW

simulating a pyramid and objects known in Ancient Egypt (i.e. statues of Egyptian Gods, Archeological Tools, etc.). The virtual world was created using the virtual world framework Open Wonderland (http://openwonderland.org). Open Wonderland is a VW, which is highly extensive and allows the integration of new modules and functionalities. Users are represented as avatars and can interact and communicate via text-chat, VoIP, or gestures [16].

In our study, we had participants either work collaboratively or individually on an assignment in the VW. Based on the theoretical findings shown above, our hypothesis was that the learning outcomes of the participants were better when they worked on the assignment collaboratively rather than individually. Furthermore, we also assessed whether the motivation of our participants changed while working on the assignment. Here, we expected that participants who learned collaboratively would be more motivated than participants who learned individually.

II. METHODS

A. Design

The only independent variable we used was the learning condition: Participants learned either individually or in groups of two to three people in the VW. Regarding the dependent variables, we measured the learning outcome of the participants (in terms of the number of correctly solved questions during a multiple choice quiz at the end of the assignment in the VW) as well as the participants' current motivation before and after the learning assignment. Finally, we also investigated the usability of the VW as learning tool in general.

B. Participants

16 participants (8 male, 8 female; mean age: 21.4, *SD* = 2.0; range: 19 to 27 years) took part in the study; half of the participants were assigned to the individual condition, the other half to the group condition. The assignment to the learning conditions was randomized. Participants gave informed consent and received class credits for their participation. The experiment was approved by the local ethics board and carried out at the University of Graz.

C. Stimuli & Procedure

Virtual world: The virtual world was implemented using the open-source framework Open Wonderland. Different modules were implemented to support the explorative and collaborative scenarios. This included an item module (to collect and display items in the world), a quiz tool (to assess knowledge), an item board (to share collected information), and the role manager. The role manager is connected to the item module which makes certain items only visible to certain roles. This capability was designed to support a fair and balanced scavenger hunt scenario and makes collaboration to succeed necessary. The virtual world was set up in the style of Ancient Egypt (desert, pyramid; see Fig. 1).



Figure 1. Screenshot of the pyramid in the virtual world (VW). Participants had to enter the pyramid to collect information from certain objects.

Participants were supposed to learn about a myth from ancient Egyptian mythology (the Osiris Myth). Most 3D models (pyramid, statues, collectables) used for the scenario were from the Google 3D warehouse [17].

P*reparation:* Each participant was assigned a role (scientist, adventurer, priest, or historian) when he or she entered the VW. This role indicated which objects the participant had to search for in the pyramid. We provided 8 objects in total (5 statues, 3 other objects, i.e. statues of Egyptian Gods, tools for digging in the sand etc.; see Fig. 2 for an example). The participants' task was to find the randomly placed objects within and around the pyramid.

Each of these objects provided information about Ancient Egypt (e.g. the statue of the Egyptian God Osiris provided information about this god) and the participants were instructed to memorize this information for an upcoming quiz at the end of the assignment. Critically, when a learner learned individually, he or she was able to find all objects necessary to complete the quiz successfully, whereas in the collaborative learning condition each group member could find only a subset of the objects. That is, the group members had to exchange the information about each object they had found in order to complete the final quiz (which was carried out individually) successfully.



Figure 2. One of the objects in the virtual world that participants had to find during their task. These objects included certain information that could be uncovered when the learner right-clicked on it.

Procedure: Before the participants started the learning assignment, we asked them about their attitude towards working in groups in general. After that, their current motivation towards the upcoming assignment was assessed. To this end, we used the FAM Questionnaire [18] in the German language. The FAM is a questionnaire that tests for motivational aspects. It consists of the four subscales "Herausforderung" [Challenge; e.g., "The exercise is a real challenge for me"], "Interesse" [Interest; e.g.: "I like riddles like this one"], "Erfolgswahrscheinlichkeit" [Probability of success; e.g.: "I think I can face the diffi-

culty of his exercise"], and "Misserfolgsbefürchtung" [Fear of failure; e.g.: "When I think about the exercise, I'm a little worried"].

As most of the participants have never worked in a VW before, each participant received a handout in which the basic functions of the VW were described. As explained above, the participant's task was to search for a variety of objects in the virtual world (the actual number depended on the character which the participants were assigned to, but usually, each participant – regardless of the learning condition – had to find approximately five objects) and learn the information provided by each object.

In the individual-learning condition, each participant worked alone in order to complete the assignment. That is, all of the objects had to be found and the information had to be learned by a single person. In the collaborative-learning condition, the participants worked collaboratively on the task. That is, they were allowed to support each other during the search (e.g., they could tell the other group members where individual objects could be found). When working collaboratively, participants sat in the same room, but worked on separate computers with dividers in between. We also provided headsets for the online-collaboration. However, due to permanent technical problems with the integrated communication tool of the VW, the participants had to talk orally to each other during the task in the VW.

Participants were allowed to spend a maximum of 10 minutes in the VW. When the time was up or when all objects had been found, the participants were required to go to a trigger point in the VW (a car placed near the entrance of the virtual pyramid) in order to start the final quiz. The quiz consisted of six multiple choice (MC) questions (see Fig. 3). Each participant had to work on the MC Quiz individually, regardless of whether he or she was part of the individual or group condition.

After completing the quiz the participants pressed a button to indicate that they finished the assignment. We then re-assessed their current motivation towards the VW exercises, again using the FAM [18].



Figure 3. The MC Quiz used in the virtual world.

However, this version of the FAM differed slightly, regarding some formulations of sentences, from the first FAM the participants had to fill in before the virtual world task (e.g.: "I think I won't succeed in this exercise" before the assignment changed to: "I think I haven't succeeded in this exercise" afterwards). Finally, the participants were asked to fill in the SUS Usability Scale [19] to assess the usability of the VW learning environment.

III. RESULTS

In order to test the learning performance of the participants, we first compared the average learning performance of individual learners (*M*: 3.50 out of 6, *SD*: 1.20) to the average learning performance of collaborative learners (*M*: 4.13 out of 6, *SD*: 1.64; see Fig. 4). The findings suggest that collaborative learners were slightly more successful than individual learners. However, a *t*-test for independent samples revealed that participants that learned individually did not differ significantly from participants learning under the collaborative condition, $t(14) = 0.60$; $p = 0.28$.

The comparison of the participants' current motivation before and after the assignment revealed differences in the four subscales over the two time points of assessment (i.e., before and after completing the assignment in the VW).

Both in the individual and collaborative condition, participants showed a decline in the subscale of "challenge" (individual: before: *M*: 5.25, *SD*: 0.44, after: *M*: 4.78, *SD*: 0.65; collaborative: before: *M*: 5.00, *SD*: 0.52, after: *M*: 4.72, *SD*: 0.70). This suggests that the participants found the assignment to be not overly difficult. The subscale "fear of failure" also showed a decline for both conditions (individual: before: *M*: 2.75, *SD*: 1.06, after: *M*: 2.50, *SD*: 1.20; collaborative: before: *M*: 3.60, *SD*: 1.68, after: *M*: 3.18, *SD*: 1.59). This may mean – in accordance with the reduced challenge – that the participants had less fear of failure regarding the task after working in the VW, than before the task.

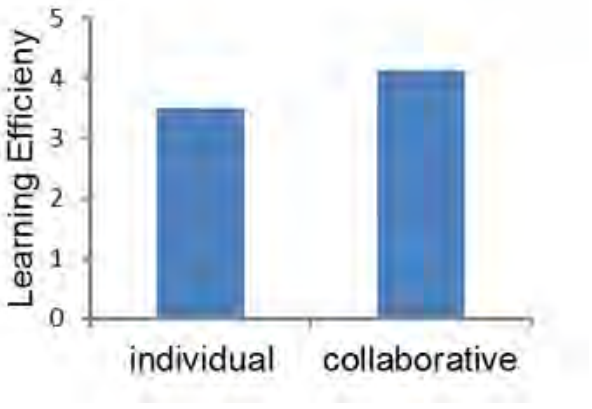


Figure 4. Mean learning efficiency for students in the individual and collaborative learning condition.

The subscale "probability of success" showed an increase in both groups (individual: before: *M*: 3.75, *SD*: 0.44, after: *M*: 4.22, *SD*: 0.63; collaborative: before: *M*: 3.62, *SD*: 0.46, after: *M*: 4.34, *SD*: 0.60). This suggests that the participants became convinced during working in the VW that they would complete the task successfully. An interesting result was achieved for the subscale "interest": It showed a slight decline in individual condition (before: *M*: 5.43, *SD*: 1.15, after: *M*: 5.28, *SD*: 0.95) and an increase in the collaborative condition (before: *M*: 4.65, *SD*: 1.33, after: *M*: 4.98, *SD*: 1.73, see Fig. 5). This suggests that there is at least a tendency showing that collaboratively working participants were more interested in the assignment after working in the VW, whereas individually working participants showed a decrease of interest after they had completed the task. However, this tendency was statistically not significant. ANOVAs performed for each subscale did not reveal reliable differences (all tests: $p > .05$).

The Usability Scale (SUS) [19] revealed that participants in the individual learning group rated the usability of the virtual world higher (*M*: 79.06, *SD* = 6.81) than participants of the collaborative learning group (*M*: 67.66, *SD* = 20.03), although a t-test revealed no statistical difference, $t(14) = 0.92$; $p = 0.18$. A possible reason for the slightly better evaluation of the VW by the individual as compared to the collaborative learners might be that it was caused by the technical challenges in the collaborative condition – i.e. the communication had to be done orally with the participants sitting in the same room due to problems with the communication tools in the virtual world (see above).

IV. CONCLUSION

In our study we showed that participants that learned collaboratively in a VW tended to achieve a better learning outcome as compared to participants that learned individually. Furthermore, the current level of motivation changed while the participants were working in the VW.

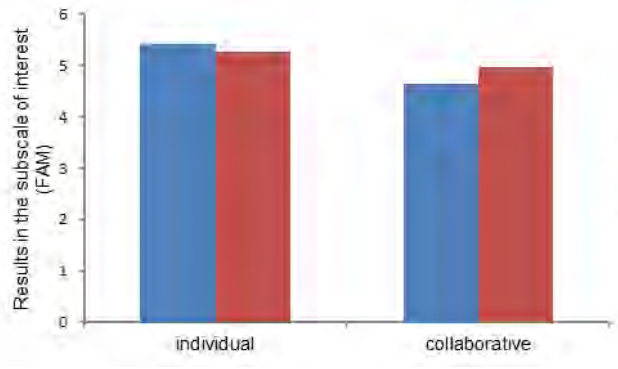


Figure 5. FAM Subscale "interest" [Interesse] for individual and collaborative learning students at the two FAM assessments (interest 1: assessment before the exercises in the virtual world (blue); interest 2: assessments after the exercises in the virtual world (red)) [18].

In particular, three of the subscales for motivation brought up consistent results for both (the individual and collaborative learning) groups: "challenge" (decline), "probability of success" (increase) and "fear of failure" (decline). This suggests that the assignment was not too difficult for the participants. Note that the learning content provided (Ancient Egypt) was not specifically selected for university students, but for the general public. That is, in a "real" (university) learning setting, the learning content should match the skills of the learners. However, as we were not interested in the learning outcome per se, but only in the differences between individual and collaborative learning, the difficulty of the learning content was not of primary interest for us. Interestingly, however, the subscale of "interest" shows a decline in the individual learning condition and an increase in the collaborative learning condition. Previous research reports increased prosocial behaviour and sympathy between group members in collaborative learning groups [6, 7]. This may be one reason why the interest levels of collaborative learners increased during the task, whereas the interest levels of individual learners decreased. That is, the important factor of prosocial behaviour and sympathy within the group may boost the interest of the collaborative group members in the task.

As our sample was rather small (and due to this there is a lack of statistical evidence), future research should have a closer look at the motivation of participants learning under collaborative and individual learning conditions, especially at the variation in motivation and/or interest for the learning assignment when learning individually vs. collaboratively. From a technical point of view, the study revealed some important points that should be considered in future versions of the VW. For instance, the assignment of the roles (scientist, adventurer, priest, and historian) to the students was not particularly flexible (i.e. it was difficult to change roles individually), and the items in the VW could not be added, removed or displaced easily by the experimenter without technical support. This potentially makes the learning environment not as attractive for "real-world" use when, for instance, teachers without a comprehensive technical knowledge want to create new learning environments. In addition, we had difficulties controlling the time stamps for how long the participants worked in the VW (i.e., we could not retrieve when each participant actually started working in the VW or ended working etc.). Such information would of course allow deeper analysis of how the learners experience and explore the learning environment. Furthermore, as mentioned above, the integrated communication feature of the VW did not function properly. This forced us to allow the group members in the collaborative learning group to communicate orally. They were placed in the same room with a divider between each participant during the learning assignment, which prevented visibility between participants. This communication problem effected collaborative learners only, which may also be reflected in the lower evaluation of the usability for these participants.

All in all, this study revealed some important and interesting points about individual versus collaborative learning in a VW. E-learning and especially e-learning within a VW, is an important component of future education and

it is therefore very important that future research focusses on how it can support learners both in individual and collaborative learning.

ACKNOWLEDGMENT

We would like to thank Christian Poglitsch for his technical support when conducting the study.

REFERENCES

[1] D. Bodemer, and W. Echterhoff, E-Learning. In M. A. Wirtz (Ed.), Dorsch - Lexikon der Psychologie Retrieved on 24.02.2015 from https://portal.hogrefe.com/dorsch/e-learning/, 2015.

[2] F. Golband, A. F. Hosseini, R. Mojtahedzadeh, F. Mirhosseiniand, and S. Bigdeli, The correlation between effective factors of E-learning and demographic variables in a post-graduate program of virtual medical education in tehran university of medical sciences. Acta Medica Iranica, vol. 52, 2014, pp 860-864.

[3] H. Gruber, and E. Stamouli, Intelligenz und Vorwissen, in Wild, E. & Möller, J. (Ed.), Pädagogische Psychologie (1.th ed). Berlin, Heidelberg: Springer Verlag, 2009, pp. 27-47.

[4] H. Astleitner, Qualität von web-basierter Instruktion: Was wissen wir aus der experimentellen Forschung? In F. Scheuerman (Ed.), Campus 2000. Lernen in neuen Organisationsformen. Münster: Waxmann, 2000, pp. 15-39.

[5] U. Kessels, and B. Hannover, Gleichaltrige. In Wild, E. & Möller, J. (Ed.), Pädagogische Psychologie. Berlin, Heidelberg: Springer Verlag, pp. 283-304, 2009.

[6] E. Aronson, and S. Patnoe, The jigsaw classroom: Building cooperation in the classroom (2nd ed.). New York: Addison Wesley Longman, 1997.

[7] F. Borsch, J. Jürgen-Lohmann, and H. Giesen, Kooperatives Lernen in Grundschulen: Leistungssteigerung durch den Einsatz des Gruppenpuzzles im Sachunterricht. Psychologie in Erziehung Und Unterricht, 49, pp. 172-183, 2002.

[8] S. Drössler, M. Jerusalem and W. Mittag, Förderung sozialer Kompetenzen im Unterricht. Implimentation eines Lehrerfortbildungsprojektes. Zeitschrift Für Pädagogische Psychologie, 21, pp. 157-168, 2007.

[9] R. Jucks,M. R. Paechter, and D. G. Tatar, Learning and collaboration in online discourses. International Journal of Educational Policy Research and Practice, vol. 4, pp. 117-146, 2003.

[10] H. R. Pfister, H. R., and M. Mühlpfordt, Supporting discourse in a synchronous learning environment: The learning protocol approach. Proceedings of the Conference on Computer Support for Collaborative Learning: Foundations for a CSCL Community, pp. 581-582, 2002.

[11] S. Herring, Interactional coherence in CMC. Journal of Computer-Mediated Communication, vol. 4, 0-0, 1999.

[12] S. G. Straus, Getting a clue: the effects of communication media and information distribution on participation and performace in computer-mediated and face-to-face groups. Small Group Research, volume 27, pp. 115-142, 1996.

[13] G. Wesiak, M. Höfler, M. AL-Smadi, M., and C. Gütl, CSCL in non-technological environments: Evaluation of a wiki system with integrated self- and peer assessment. InPACT 2015: International Psychological Applications Conference and Trends, Ljubljana. pp. 107-111, 2015.

[14] V. Chang, C. Gütl, S. Kopeinik, and R. Wiliams, Evaluation of collaborative learning settings in 3D virtual worlds. International Journal of Emerging Technologies in Learning (iJET), 4(2009), pp. 6-17, 2009.

[15] J. Pirker, C. Gütl, J. W. Belcher, and P. H. Bailey, Design and evaluation of a learner-centric immersive virtual learning environment for physics education. In Human Factors in Computingand Informatics. Springer Berlin Heidelberg, pp. 551-561, 2013.

[16] J. Pirker, C. Guetl, and P. Weghofer, Application scenarios of interactive science fiction prototyping in virtual worlds for education. EAI Endorsed Transactions on Future Intelligent Educational Environments, vol. 1, pp. 1-9, 2014.

[17] L. M.Tomes, Exploratory and Social Learning in 3D Virtual Worlds (Master's thesis). Graz University of Technology, Austria, unpublished, 2015.

[18] F. Rheinberg, R. Vollmeyer, and B. D. Burns, FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen. Diagnostica, vol. 47, pp. 57-66, 2001.

[19] J. Brooke, SUS - A quick and dirty usability scale. In I. L. McClelland (Ed.), Usability evaluation in industry, Taylor & Franci, pp. 189-194, 1996.

Preparation of a hybrid e-learning course for gamification

D. Kermek*, D. Strmečki*, M. Novak * and M. Kaniški*
* Faculty of organization and informatics, Varaždin, University of Zagreb, Croatia

Usage of game mechanics, dynamics, aesthetics and game thinking in a non-game systems is called gamification and recently becomes very popular in academia. In e-learning systems gamification is mainly used to increase user's motivation, experience and engagement. When using gamified design elements in e-learning, we must consider various types of learners and learning objectives that have to be set for a course. For a successful implementation, the cooperation of experts in fields of education, technology, psychology, pedagogy and design is required. The paper presents our first attempt to convert an e-learning course and its grading systems into gamified version using various gamification design elements (e.g. points, badges, leader boards, levels, progress tracking, feedback, and social engagement). Some of them are supported by popular LMSs, usually through some plugin that must be installed. One of the issues many teachers encounter is the limitation of gamified elements in official LMS that they use and how to get the most out of it.

I. INTRODUCTION

A. Gamification

The agreed definition of gamification in literature states that is the usage of game mechanics, dynamics, aesthetics and game thinking in a non-game systems. Kapp provided a broader definition for gamification in learning as the usage of game-based mechanics, aesthetics and game thinking to engage people, motivate action, promote learning, and solve problems [1]–[3]. Mechanics are the game components that make the game (or a non-game system) fun (for example: tasks, rules and feedback). Dynamics imply learner's interactions with those mechanics and aesthetics affect learner's feelings during those interactions [4]. Gamification has become a very popular subject in the e-learning research community during the last couple of years.

The common objective for using gamification in various types of software systems is to increase user's motivation, experience and engagement [4], [5]. Learning is a process that requires high motivation. Thus, we apply gamification design elements to learning management systems (LMSs) in order to motivate and engage its users, aka learners. Gamification can encourage productiveness by using people's competition instinct and stimulating their sense of pride and achievement [6]. However, gamified design elements must be applied cautiously to LMSs, considering the course objectives and all types of learners and their learning habits. Some learners do not like to compete with others, while other may have a very high intrinsic motivation and consider additional motivation a waste of time [6]. Thus, when applying gamification to learning, a cooperation of experts from various fields is highly recommended (including education, technology, psychology, pedagogy and design) [7].

Before its penetration into LMSs, gamification was successfully applied into various entertainment and business applications. Many successful examples of such application like *American Airlines Frequent Flyer Program*, *Foursquare* and *Nike Plus* can be found in literature [4]–[6], [8]–[10]. The application of gamified design elements in LMS courses requires extra caution because of its pedagogical and psychological components. Commonly used gamified design elements in non-game systems include levels, points, badges, trophies, leaderboards, challenges, time based activities, progress tracking, feedback, freedom to fail, social engagement, etc. [2], [6], [9]. We will shortly discuss the most common of them and their suitability for application in LMSs.

Levels and progress tracking are interlinked gamified design elements because levels are used to indicate progress. In LMSs, levels can be understood as a set of topics and lessons or milestones that a learner achieves by completing tasks [2]. Progress bars are used as percentage based guide to learners, providing them with a type of feedback. Good feedback should contain information on what the learner has done and how he can advance in the future [6].

Points are already used in most LMSs to grade learner's participation and work. In addition to levels, they can also be used to indicate progress on the course. However, in a gamified LMS courses, points can also be used to indicate learner's status (by using points gained from other learners on forum, wiki, etc.) and for unlocking some additional content or awards (badges, trophies) [2], [4].

Badges and trophies are the most common types of achievements / awards used in gamified systems. However, it is not enough to simply include achievements into a LMS course. Learners should have the ability to view and share their achievements. A preview of all available achievements is also recommended as an additional motivating factor (*catch them all*) [4]. Most of, achievements must be meaningful and not easy to obtain [2].

Leaderboards are one of the most commonly used gamified design elements because of their simple implementation and support by popular LMSs. However, providing a single course leaderboard may not motivate

many learners. A better approach is to use multiple leaderboards for each activity and a cumulative leaderboard for all activities [2].

Freedom to fail enables learners to learn from their mistakes and previous experiences. Gamified LMS courses should provide multiple attempts in quizzes and tests whenever possible [2]. Equally important, learners should receive positive instructional feedback and be encouraged to try again after some time [6].

B. *Related work*

During the last couple of years, as gamification became a popular research subject in learning, several attempts were made to gamify various e-learning courses. The conducted studies differ in their gamification implementation techniques and show quite different results. Some of them were very successful in motivating and engaging learners, while others experienced poor results and even worse feedback from learners. We will shortly discuss a few course gamification researches.

Adrián Domingues et al. [5] made an experiment gamification plugin for their LMS and found that students who completed the gamified experience got better score in practical assignments, but performed poorly on written assignments. However, their plugin had several shortcomings. It was not fully integrated with the LMS and learners had to manually upload screenshots which were evaluated afterwards, so no instant feedback was provided [5].

Kay Berkling and Christoph Thomas [11] analyzed the unsuccessful gamification of a software engineering course. A custom made Java online interactive application was used to gamify the course. Their survey showed that only 18% of the students play games daily and around 70% of students explicitly did not think of challenge, autonomy or goal seeking as part of why they play games. Poor results are contributed to the lack of application's esthetical appeal, a strong culture of traditional schooling and the student's perception of gamification as an unnecessary hindrance towards studying for the exams [11].

Maria-Blanca Ibanez et al. [12] identified positive effects of gamification in their study of gamifying a C programming language course. They have designed and implemented a so called *Q-Learning-G* gamified learning platform. Authors have collected and analyzed both qualitative and quantitative data and found that students continued their study even after reaching the learning goal. Student's participation was beyond requirements as they continued to master unexplored topics and thus have potentially improved their knowledge [12].

Sasko Ristov et al. [13] found very positive results of gamifying the practical project in the hardware-based course *Microprocessors and Micro-controllers*. Authors claim that gamification improved the course significantly, resulting in more than 60% of students achieving the maximum grade (on a difficult and demanding course). In addition, both genders have improved the average project grade, as well as the average final grade [13].

Most of the conducted studies used their own implementations of gamified LMS, application or plugin. In this paper we will focus on gamified design elements available in popular LMSs (out of the box or via plugins). The remainder of this paper is divided into three sections. Section II presents the gamified design elements available in popular LMSs and their characteristics. Section III presents our experience and findings on converting a classic learner's evaluation model into a gamified one. Finally, section IV outlines the conclusions and gives recommendations for future work.

II. GAMIFICATION OF EDUCATION

A. *Complete gamified platform*

The simplest way to use gamification is through complete gamified platform. Badgeville is one company that provides complete gamified platform and they defined gamification as "a proven business discipline that takes the techniques that make games engaging, fun, and compelling, and applies them to technology investments" [5]. "The company characterizes gamification as a behavior platform for community engagement." [14] Other way is to use one of the existing LMSs as ones described in next subsection.

B. *Gamification in popular LMSs*

Many LMSs have been upgraded with some form of gamification. LMSs are used by academic institutions and companies to manage, track, and deliver courses and training programs. In a ranked list of the top 20 most popular LMS by Capterra [15], 10 systems are listed as including gamification. In this research we analyzed the top three LMS focused on K-12 or higher education. Table 1 shows these top three LMSs, partially based on [16], in additional with their game mechanics support which were selected by a combination of their total number of customers, active users, and online presence.

TABLE I. GAME MECHANICS SUPPORT IN LMSs

Game Mechanics	LMS		
	Edmodo[1]	***Moodle***[2]	***Blackboard***[3]
Points	No	Yes[4]	No
Levels	No	Yes[4]	No
Leaderboards	No	Yes	No
Badges	Yes	Yes	Yes
Challenges / Quests	Maybe	Yes	Maybe
Onboarding	Yes	Yes	Yes
Social Engagement loops	No	No	Yes

Results in [16] show that all of the reviewed LMSs offer intangible rewards in the form of badges and some

[1] www.edmodo.com
[2] www.moodle.org
[3] www.blackboard.com
[4] As plugins

measure of progress. None made use of points or levels, which are equally commonly found in gamification literature [1], [2], [5], [11], [17]–[19]. The lack of points and levels is harder to explain. One explanation gives [20]: "… that students are already being given either points or grades for their assignments and that the addition of another set of points might be confusing." It is also difficult to reward levelling up in a LMS, because we need to consider gamified and not gamified students if such a grouping exists.

Moodle is an open source LMS enabling many educators of: companies, schools, faculties (including faculty of the authors) and universities; to create their own courses with dynamic content that extend learning anytime and anywhere. [21] Since it is so popular and our faculty uses Moodle as a supporting system our gamification will be based on Moodle.

C. *Moodle as an gamified platform*

Moodle community continuously upgrades the system with various learning elements and recently one of them is gamifications. For applying gamification in Moodle one could use various possibilities that Moodle is offering like: groups, forums, quizzes, progress bar, badges, workshops, activity completion, conditional access, questioners, etc. [4] In [23] many features that can be used to gamify the Moodle can be found.

"Structural gamification focuses in the application of game elements in the learning process in order to motivate learners, taking into account that we do not modify the content." [22] Game-elements used in the gamified environment should motivate students and engage their learning. Game-elements should not distract the students and must not be used just because they exist. These game-elements must be carefully chosen, by the teacher, according to learning types and learning outcomes, and always there should be a reason why some feature is used and how it effects the student.

According to [22] features that can be used like the ones described above can be divided into four groups: progressive learning, social, instant feedback and rewards. Every category can be used to have different effect on students.

Administrators can, except from existing features, install Moodle plugins that improve gamified experience for example: Level up!, Progress Bar, Leaderboard, etc. These plugins can further improve the gamified feeling and maybe better motivate and engage students.

III. CONVERTING A CLASSIC INTO GAMIFIED LEARNER'S EVALUATION MODEL

The basis for gamification of the existing course is a score board that is applied during the past few years. Table II shows the elements of the score board, and their proportion in the final grade. Some of these activities such as "Lecture Activity" have so far not been directly marked but were based on the notes of teachers as well as the fulfilment of commitments by the students (e.g. the entry of student information on the server, filling out the initial and final survey, participation in discussions on forums and etc.). During initial design of gamified course version we started with the model described by Urh et al [7].

In the new approach the intention is that all activities will be directly graded which will affect the learning path of each student. Some other activities for example homework were quantified but not properly qualitative marked (similar to places in competitions: 1st, 2nd and 3rd). Specifically in this case plan is to set aside for each homework three best solutions which are very likely graded with the highest grade. For most activities the element activity completion is included, respectively fulfilment of minimum criteria, so that students are visually informed of their fulfilment. Also by some activities it is required that the student manually confirm the activity completion while in other activities it is resolved automatically after the completed activity.

TABLE II. ELEMENTS OF THE SCORING SYSTEM, AND THEIR PROPORTION IN THE FINAL GRADE

Activity	Points
Attendance (lectures, seminar, laboratory exercise)	5.0
Grade of Homework 1, 2,3,4,5	25.0
Seminar	5.0
Midterm 1	7.0
Midterm 2	8.0
Laboratory Exercise and Seminar Activity	3.0
Lecture Activity	2.0
Project	45.0
TOTAL	**100.0**

To achieve a higher level of competitiveness among students, which is also one of the fundamental characteristics of gamification, a new plugin was installed by the name "Level up!" and the following blocks were included: Progress Bar, My Latest Badges and Activity Results. Most activities are limited in their availability on the basis of a time schedule in the course which is to students visible in their calendar and progress bar. For activities with complex availability conditions, i.e. laboratory exercise groups with different homework submission times. Deadlines are set for completeness on the basis of various groups in which the students are involved.

Moodle unfortunately does not display these complex activities in the calendar. So when viewing the progress bar students can see which activities are carried out in time, in which they are late and which awaits them in the future. To further stimulate students highlighted activities (e.g. Midterm) have been selected and they display 10% of the best students in the activity results block. The activity results block will be manually activated and deactivated depending on chosen activities in order to prevent filling out the entire space with that type of block (more elements of that block type).

"Level up!" block is configured to award students with points based on their active or passive usage of the Moodle system in the course (e.g. read notification, created forum post, committed homework, resolved questionnaire, reading teaching materials, watching videos

and etc.). The goal of using this block is to encourage interest among students to use all of the available teaching materials and carry out variety of activities. Depending on how many points students gathered they can progress to the next level and grain a stronger "title" or "rank".

The most demanding part in the preparation of the "Level up!" block was to setting up rules on witch experience points are been obtained. Rules vary from simple ones to some that contains a large number of conditioned activities or events. In addition, analysis and simulation needs to be performed for sharing experience points with the new rules and corresponding predefined rules. Also the possibility of repeating the same actions within a certain time frame needs to be considered. Real situations tends to a normal distribution (Gaussian) of commitments by students which should be considered during the preparation of the system.

As an example, on the highest level, we should have no more than 10% of students. If the activity is one of the homework assignments then the points are given after the homework has been graded. Complex examples of rules usually use the operator "ANY" which allows the connection of multiple events to obtain the exact same number of points. This form of connecting events is not sufficiently restrictive because it does not provide the connection of events occurring at various time frames. To resolve this problem activities such as "collect points" are defined which are based on the completion of various different activities. For example (Figure 1), it is necessary to connect activities at the beginning of the course (reading learning outcomes, selecting a group for laboratory exercise, fulfilling the initial survey, etc.) and when all activities are completed a new activity (type survey) opens named "collect beginning points" and can be accessed only once.

This is a classic example of mapping characteristics from computer games, which is the main aim of gamification. Specifically in our course seven levels have been foreseen where the transition to a higher level requires twice as many points in relation to the previous level. With the transition to a higher level students acquire a stronger badge instead of previous one.

In order to promote additional competitiveness students can collect badges in a variety of ways taking also into account that in some areas the number of badges is limited. Conquering of badges is enabled through a new block named my latest badges.

Badges that are directly related to the quantitative form of evaluation activities, such as "collect badges", were created in a different way. For complex criteria it is not possible automatically receive a badge. Activities "collect badges" follow the same pattern as previously described activities "collect points". So, for example, the activity "collect badge for the maximum points of two neighboring homework assignments" becomes visible only when the student has 100% of points from two neighboring homework assignments (e.g. homework one and homework two or homework two and homework three).

Badges that are related to the qualitative form of evaluation cannot be automated in any way, and they are assigned manually by the teacher. For example, the badge best homework (gold badge) is assigned to the student who made the best homework according to the opinion of teachers. We plan to use so-called "cumulative badges" which are earned based on a large number of quantitative badges and several qualitative badges. For example, "best of the best homework assignments" which is obtained when a student earns one of following set of badges for homework assignments: at least three gold badges; two gold badges and at least one silver and at least one bronze badge; one gold badge and at least two silver and at least one bronze badge; one gold badge and at least of one silver and three bronze badges; at least three silver and two bronze badges. As one can imagine they could only be manually assigned by the teacher.

Badges related to the completion of one or more activities serve as motivator to students carrying out their activities. Badges related to qualitative and quantitative form of evaluation serve along with motivation also to achieve certain benefits. For example, if a student has a specific combination of badges he or she can get some privilege, e.g. one compensated absence from lectures.

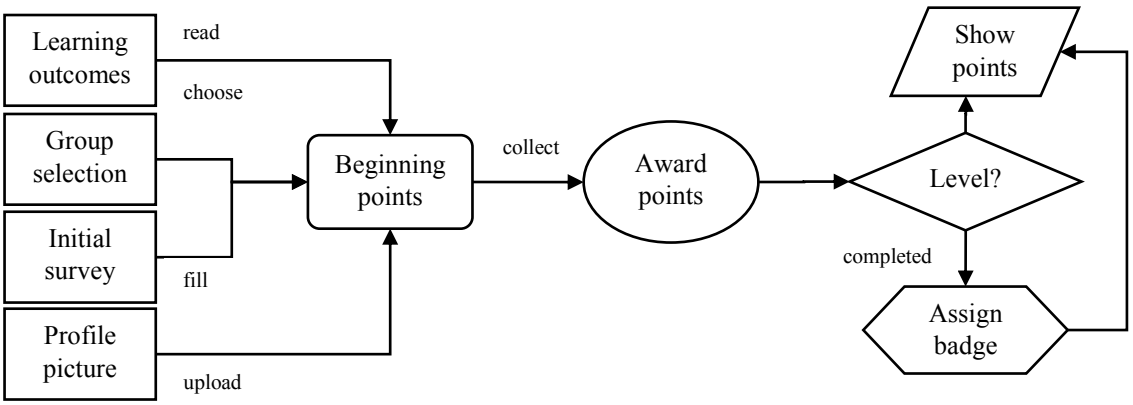


Figure 1 Example of complex rule for experience points

TABLE III. COMPARISON OF EXPERIMENTAL AND CONTROL GROUP

Activitity	Experimental group			Control group		
	Average score	Median	Standard deviation	Average score	Median	Standard deviation
Test of initial competencies	64.4%	65.00%	14.20%.	63.80%	70.20%	20.80%
1. self-assessments test	60.07%	64.44%	21.47%	54.60%	56.67%	20.35%
2. self-assessments test	69.44%	75.00%	19.25%	65.97%	70.00%	24.11%
1. homework	68.38%	73.33%	24.44%	66.31%	73.33%	24.44%
2. homework	70.20%	77.50%	24.85%	68.72%	77.50%	27.04%

Students will be at the beginning of the course introduced with basic elements and rules of the gamified course. Respectively students will be introduced to a system which evaluates their tasks and awards them with points and badges which has been designed and develop upfront. Intentionally students will not be provided with information on all types of badges, how they can be acquired and what are the benefits. This approach rises curiosity and engagement of students during the course, positively influences on they work results, and enables them to discover different badges and its benefits.

The highest form of stimulation can be obtained in the case of acquisition of the highest level in the block "Level up!" and the combination of several cumulative badges. This stimulation exempts a student from final project and awards him / her with maximal points.

In a case that some students especially do not want to participate in gamified version of course we need to assure that they will not see or get notification of any earned points or badges. To implement such a dual course mode we have to prepare some simple adjustments like adding new groups to separate gamified and non-gamified students. More demanding task is adding new role for non-gamified students because that can only do Moodle administrator. If it is not properly setup then that role is visible in other courses and could confuse their teachers and students. New role has arch type of usual student role and therefore inherits all permissions that has his parent role. Further, the administrator must prohibit for new role all permissions that are connected to badges, experience points etc. That works almost as we planned but in current version of Moodle there is a problem with excluding non-gamified students (with new role) from getting automatic badges, so simple badges will be implemented as quantitative badges.

IV. PRELIMINARY RESULTS

As we mentioned previously, the goal of this paper is to present what we have done to gamify the course, what obstacles we have faced in that endeavor and what support for gamification one can find in the most popular LMSs. The gamification of the course has been planned for roughly one year and preparation have been undertaken merely 3 months before the course started in summer semester and some adjustments have been still performed as the semester is in progress. At the present time in a middle of the semester we have some preliminary results concerning gamification and student's achievements.

At the beginning of the semester students chose to be part of an experimental group (knowing nothing what is all about) or be in a regular (control) group. We set 50% as maximum members for experimental group with first-come principle and students got involved very enthusiastically so we got even numbers for both groups.

Table 3 presents data from short analysis with focus on results of few activities like test of initial competencies, self-assessments tests, and homework assignments.

Test of initial competencies students could fill just ones and it showed that the students from the experimental group achieved on average score 64.40% and the students from the control group had 63.80%. Median with 70.20% was higher for the control group while the experimental group had 65.00%. Also standard deviation was higher for the control group 20.35% and for the experimental group was 14.2%. According to the presented data we can conclude that the experimental group is more homogeneous then the control group and with slightly higher competencies.

The students have tree opportunities to fill a first self-assessments test and we shell discuss only the best results. The experimental group had on average score 60.07% with 64.44% median and 21.47% standard deviation. The control group had on average 54.60% with 56.67% median and 20.35% standard deviation. The students from the experimental group did more test tries (76) then from the control group (64).

For second self-assessments test the students have only one opportunities to fill it. The experimental group had on average score 69.44% with 75.00% median and 19.25% standard deviation. The control group had on average score 65.97% with 70.00% median and 24.11% standard deviation. The students from the experimental group did more test tries (30) then from the control group (24). Both homework assignments had very similar data like second self-assessments test. The experimental group had better results roughly 5% for average and median. But the standard deviation is smaller for the experimental group.

According to presented data we can conclude that students from the experimental group are more interested in performing optional activities and achieve better results in mandatory activities. We do not have complete data due to semester is still in progress but we see that the gamification has some positive influence on the students from the experimental group.

V. CONCLUSION

Gamification has been present in higher education in recent years. Some gamification pilot projects, as mentioned in the introduction, showed significant improvements in student achievements and engagements compared to usual education model. Gamification could be based on dedicated gamification software but in our context it is used together with eLearning system Moodle.

The literature like [23] gives only very simple examples of gamification by using badges and earning experience points. Complex examples with complex rules and features are not presented. Our effort is directed toward real gamified course with complex rules and features that should fulfill all learning outcomes but also give students more interesting and challenging work during the course.

The novelty of our approach is the combination of simple and complex rules as a way of rewarding students by using multiple features. These features are: "Level up!", simple and cumulative badges, progress bar, scoreboard, activity constraints (time restrictions, visibility constrains, constraints based on grades and / or groups), activity completion, etc.

Based on preliminary results we can conclude that students from the experimental group are more interested in performing optional activities and achieve better results in mandatory activities. We are planning to analyze the final results after the course finishes. Some of them will be compared to our data from previous years when gamification was not used and to the results from other researches especially those that were done in the higher education and courses that deal with ICT.

REFERENCES

[1] A. Amriani, A. F. Aji, A. Y. Utomo, and K. M. Junus, "An empirical study of gamification impact on e-Learning environment," in *Proceedings of 2013 3rd International Conference on Computer Science and Network Technology, ICCSNT 2013*, 2014, pp. 265–269.

[2] Enders Brenda, *Gamification, Games and Learning: What Managers and Practitioners Need to Know*. The e-learning Guild, 2013.

[3] K. M. Kapp, *The gamification of learning and instruction: game-based methods and strategies for training and education*. John Wiley & Sons, 2012.

[4] G. Zichermann and C. Cunningham, *Gamification by design: Implementing game mechanics in web and mobile apps*. " O'Reilly Media, Inc.," 2011.

[5] A. Domínguez, J. Saenz-de-Navarrete, L. De-Marcos, L. Fernández-Sanz, C. Pagés, and J.-J. Martínez-Herráiz, "Gamifying learning experiences: Practical implications and outcomes," *Comput. Educ.*, vol. 63, pp. 380–392, Apr. 2013.

[6] I. Glover, "Play as you learn: gamification as a technique for motivating learners," in *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 2013, vol. 2013, no. 1, pp. 1999–2008.

[7] M. Urh, G. Vukovic, E. Jereb, and others, "The model for introduction of gamification into e-learning in higher education," *Procedia-Social Behav. Sci.*, vol. 197, pp. 388–397, 2015.

[8] G. I. Bíró, "Didactics 2.0: A Pedagogical Analysis of Gamification Theory from a Comparative Perspective with a Special View to the Components of Learning," *Procedia - Soc. Behav. Sci.*, vol. 141, pp. 148–151, 2014.

[9] C. I. Muntean, "Raising engagement in e-learning through gamification," in *Proceedings of 6th International Conference on Virtual Learning ICVL*, 2011, vol. 2011, no. 1, pp. 323–329.

[10] O. Wongso, Y. Rosmansyah, and Y. Bandung, "Gamification framework model, based on social engagement in e-learning 2.0," in *Technology, Informatics, Management, Engineering, and Environment (TIME-E), 2014 2nd International Conference on*, 2014, pp. 10–14.

[11] K. Berkling and C. Thomas, "Gamification of a software engineering course and a detailed analysis of the factors that lead to it's failure," *2013 Int. Conf. Interact. Collab. Learn. ICL 2013*, pp. 525–530, 2013.

[12] M.-B. Ibanez, A. Di-Serio, and C. Delgado-Kloos, "Gamification for engaging computer science students in learning activities: A case study," *Learn. Technol. IEEE Trans.*, vol. 7, no. 3, pp. 291–301, 2014.

[13] S. Ristov, N. Ackovska, and V. Kirandziska, "Positive experience of the project gamification in the microprocessors and microcontrollers course," in *Global Engineering Education Conference (EDUCON), 2015 IEEE*, 2015, pp. 511–517.

[14] R. Conaway and M. C. Garay, "Gamification and service marketing," *Springerplus*, vol. 3, no. 1, p. 653, 2014.

[15] Capterra, "Top 20 most popular LMS." [Online]. Available: http://www.capterra.com/learning-management-system-software/#infographic. [Accessed: 29-Jan-2016].

[16] J. Broer and A. Breiter, "Potentials of Gamification in Learning Management Systems: A Qualitative Evaluation," in *Design for Teaching and Learning in a Networked World*, Springer, 2015, pp. 389–394.

[17] J. C. Paiva, J. P. Leal, and R. Queirós, "Odin: A Service for Gamification of Learning Activities," 2015, pp. 194–204.

[18] I. V Osipov, E. Nikulchev, A. A. Volinsky, and A. Y. Prasikova, "Study of gamification effectiveness in online e-learning systems," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 2, pp. 71–77, 2015.

[19] J. J. Vleeshouwer, "Gamification in higher education : factors influencing the usage intensity of a gamified E-learning application." 2015.

[20] C. Gráinne, K. Tomaz, R. Christoph, K. Johannes, and E. Lavoué, "Design for Teaching and Learning in a Networked World," in *10th European Conference on Technology Enhanced Learning, EC-TEL 2015*, 2015, vol. 9307.

[21] Moodle, "Moodle features." [Online]. Available: https://docs.moodle.org/30/en/Features. [Accessed: 29-Jan-2016].

[22] H. Pastor Pina, R. Satorre Cuerda, R. Molina Carmona, F. Gallego Durán, F. Llorens Largo, and others, "Can Moodle be used for structural gamification?," 2015.

[23] N. Denmeade, *Gamification with Moodle*. Packt Publishing, 2015.

Implementation of Fundamental Ideas into the Future Managers´ Informatics Education

L. Révészová
Faculty of Economics Technical University of Košice
Department of Applied Mathematics and Business Informatics, Košice, Slovakia
libusa.reveszova@tuke.sk

Abstract – Informatics is a relatively young discipline which is very dynamically developing at present. Since its implementation and applications are important in all areas it has become a regular school subject. Probably because of unpredictable development and impact in the future, there are many questions concerning the content and methods of informatics education. The paper intends to find the way how to prepare proper informatics education for future non-professionals in Informatics, so that it isn´t just about "clicking the correct buttons" in the current packets of programs/applications - and therefore not getting obsolete just after a few months/years. Being aware of the increasing pedagogical challenges posed by using and teaching ICT, we at the Faculty of Economics, Technical University of Košice started to improve our Informatics courses. The education we propose is practical and real-life oriented. Our aim is also to capture successful experiences of learning and teaching from both academic and practice areas.

I. INTRODUCTION

Computer Science - Informatics is a dynamically developing discipline; paradigm changes are constantly announced. Although computer science has been a regular subject at school for a relatively long time, there is still more and more discussion on how to teach computer science and what to teach: Should computer science education be oriented more towards its applications or more towards its fundamentals or more towards its social effects [18]?

Nowadays informatics education is essential for all of us - citizens of information and knowledge society. That is why Informatics is essential for all students of all types and all levels of schools; therefore "informatics education for all" is needed. According to [6] „Informatics education for all" means informatics education also for students who will not necessarily become computer science experts. Many authors make a distinction between Information Technology (IT) - deals with the use of the computer and its applications and Computer Science (CS) - deals with the design of informatics products or more theoretical objects. This is often related to difference between users and designers with underlying idea that designers are „those who know" and users are „those who don´t need to know". I agree with the author of [6] that it is necessary to go beyond this dichotomy, which is undesirable from the education point of view.

The question is: How to prepare and then perform proper informatics education also for future non-professionals in CS, so that it is not just about "clicking the correct buttons" in the current packets of programs - and therefore not getting obsolete just after a few months/years?

I work at the Faculty of Economics Technical University of Košice at the Department of Applied Mathematics and Business Informatics. I and my colleagues educate future (finance, banking and investment) managers. At present their informatics education and their ICT competencies are very demanded. Although our students are future non-professionals in Informatics, it is desirable and very important for them to obtain timeless informatics knowledge. And that means, according to [17], [18], and according to our experience as well, to implement basic fundamental principles of Informatics into their education as much as possible.

The paper describes the way how Informatics II Course (in the first year at our faculty) is taught. This course primarily focuses on information systems (IS) - as a basis of digital business - their modeling, analysis, design of user requirements, and management, based on models and analysis of key business processes. The importance of this field is emphasized by many authors and sources, e.g. according to [5]: "nowadays digital business represents a more extreme revolution than previous technology-driven changes". Gartner predicts that a lack of digital business competence will cause 25 % of businesses to lose competitive ranking by 2017.

As we can read in [15], "managers are no longer able to afford the luxury of abdicating participation in information systems decisions. Managers who choose to do so risk having their business compromised. With the proliferation of web and e-business, information systems are the heart of virtually every business interaction, process and decision. Managers who let someone else make decisions about their information systems are letting someone else make decisions about foundation of their business".

II. SECONDARY SCHOOL GRADUATES KNOWLEDGE OF INFORMATICS FUNDAMENTAL IDEAS

On considerations of constructivism (applied in our way of education) teachers must take into account the previous learning processes and the existing knowledge of

the individual learner. It is important to discover what the students' knowledge base is, what are their models, representation, and information processing methods [2]. That is why we permanently survey students' previous knowledge base.

Our Informatics courses are prepared for students who have completed compulsory informatics education at secondary level of Slovak education system. To provide informatics education in accordance with real demand, we determine real state and level of secondary school graduates knowledge and skills also in the area of fundamental informatics ideas/principles. Since 2003 the key query has been follows: What are the real students´ informatics knowledge in the area of Informatics fundamental ideas and principles, gained during their previous education?

Mandatory educational documents [9], [11], and the prospective concept Digipedia [10], take into account the fundamental ideas of informatics and put forward next five thematic areas of the State educational program for compulsory teaching of Informatics at secondary schools:

- **Information around us -** it is focused on understanding the principles of collecting, processing, and presenting information. Students have to know the types of applications to process information, to be able to choose and effectively use the appropriate application depending on the type of the task.
- **Communication via ICT -** provides mastering the principles of computer networks and network elements, Internet services, non-interactive and interactive communication, web, browsers, websites, information retrieval, netiquette, security on the Internet and social networks, understanding different ways of searching for information, creation of websites, and compliance with the protection of networked computers.
- **Procedures, problem solving, algorithmic thinking -** provides mastering stages of solving the problem - the problem analysis, algorithm, program, and debug. The performance standards prescribes managing problem analysis, problem solving algorithm, comprehensible presentation and verifying the accuracy of the algorithm, programs understanding, determining the properties of inputs/outputs, testing, modifying and correcting the program code.
- **Principles of digital technology -** provides knowledge of software, its distribution by area of use of computer networks, its benefits, architecture, segmentation of networks by the wideness, and the connection method. Performance standard requires knowledge of the approximate capacity of various types of memory, and limits their use and basic characteristics computer accessories and equipments.
- **Information society -** introduces the use of Informatics in various areas - administration, electronic office, education, sports, arts, entertainment, e-society. Students should master the basic operation of the software business - concepts upgrade registration software, electronic documentation, information technology risks. Performance standard requires knowledge of current ICT trends, the basic features of the information society, its pros and cons, the characteristics of different types of software in terms of legal protection (freeware, shareware), explanation of the cultural, social and health aspects of the use of computers.

However, the reality may differ from the presented curriculum requirements. To determine the real situation the questionnaire entries based on current legislation, documents and standards [9], [10], [11], was created. Using the questionnaire administrated in the first semester at the first seminars of Informatics I Course, We searched for the extent and content of compulsory education and real students´ knowledge and skills. Except other things we focused on the following important areas:

- algorithms creation, working with program languages, coding, and web page creation, (Fig. 1)
- basic concepts of Informatics (data, information, knowledge, etc.), von Neumann architecture and basics of hardware, IS using, (Fig. 2)
- dealing with computer networks (topology), the Internet (protocols and rules) and internet services, browsers and search machine, (Fig. 3).

The survey sample consisted of 1921 secondary school graduates who entered first grade at the Faculty of Economics TUKE between 2003 and 2014th. 87 % were graduates from grammar schools and 13 % (mostly) from business secondary schools.

The evaluation of students' answers as a view on the real state of students´ knowledge in the mentioned fundamental obligatory areas is presented in Fig. 1, Fig. 2 and Fig. 3. in percentage.

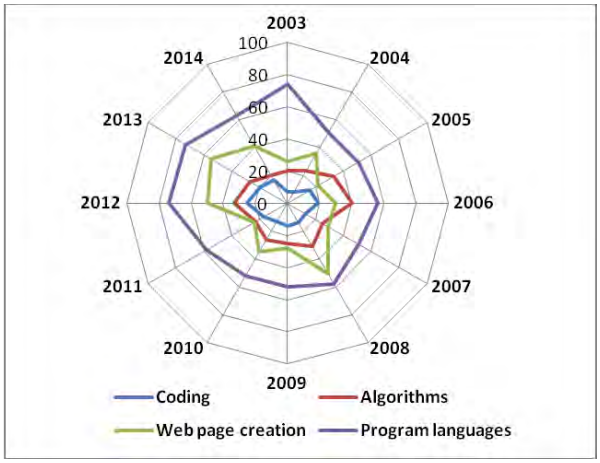


Figure 1. Dealing with program languages, algorithms, web page creation, and coding - evaluation of students´ answers "yes" in percentage.
Source own.

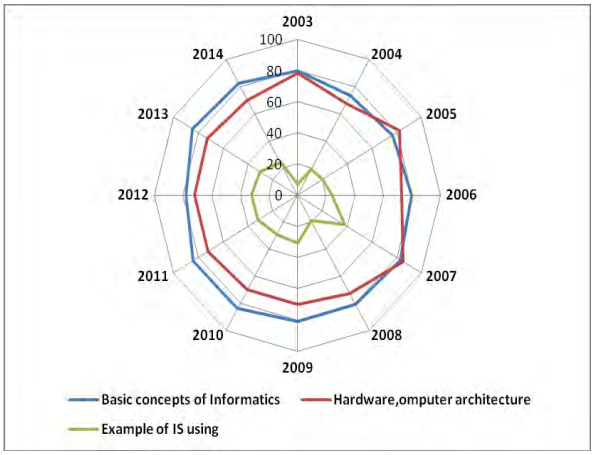


Figure 2. Dealing with basic concepts of Informatics, hardware and computer architecture, and working with information systems - evaluation of students´ answers "yes" in percentage. Source own.

As stated in the previous section, the obligatory curriculum of the school subject Informatics at the Slovak secondary schools, contains parts which cover all important fundamental ideas such as programming, algorithms, data processing, problem solving, etc. But findings, presented in previous graphs, show that a relatively high proportion of our students did not properly comprehend and manipulate with basic terms, algorithms, program languages, applications, patterns and methods based on fundamental principles of Informatics.

As we can see in Fig. 1, in the surveyed period the percentage of students who worked with the programming language does not exceed 80 %, in average only around 60 % of students had dealt with programming at secondary school. On average only 30 % of students were familiar with the concept "algorithm". Proportion of students who were familiar with web development oscillated around 40 % and only 20 % of students dealt with encoding.

Fig. 2 presents that only 80 % of students indicated education in the area of basic concepts of Informatics such as data, information, information processing. Basic knowledge of the computer architecture confirmed, on average 70 % of students. On average only one third of students experienced working with IS before.

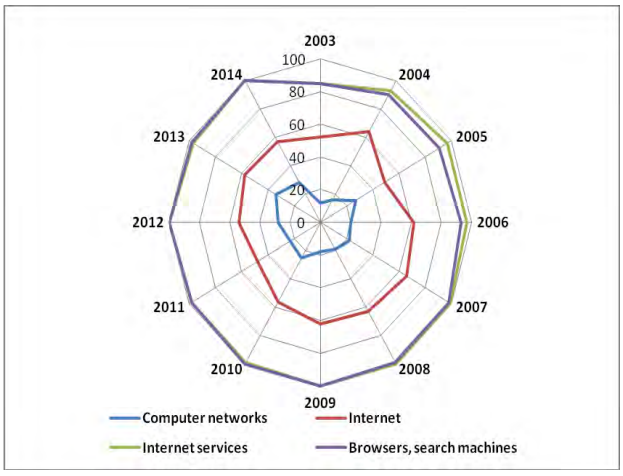


Figure 3. Dealing with computer networks, the Internet and internet services, browsers and search machine - evaluation of students´ answers "yes" in percentage. Source own.

Fig. 3 presents the results in the area of "working in the computer networks". On average 30 % of students had been briefed with basic concepts of network topology, with rules and protocols of the Internet on average of 60 %. Nearly 100 % of students experienced "working with browsers and the use of internet services - web, e-mail, Skype".

Unfortunately, the survey results didn´t meet expectations in all mentioned important area, in spite of the fact that Informatics is a compulsory subject at secondary schools in Slovakia.

III. FUNDAMENTALS OF OUR APPROACH

At our faculty we (I and my colleagues) have created and have constantly been improving learning and teaching framework for Informatics education. We are trying to increase students' interest in business informatics and in studying computer science as a discipline which can help to better understand today´s world. We are trying to explain that "knowledge and skills in the field of Informatics are powerful tools in solving problems which may occur in different application areas in professional and common lives" [20].

To answer the question what should be taught and how, Schwill in [18] and [17] suggests using the notion of fundamental ideas introduced by Bruner. There are three fundamental ideas that dominate all activities in computer science – algorithmization, structured dissection and language. Different point of view – the information-centred approach considers the process of software development to be a good guide for learning information structures using stages like constructs proper models, programming, and so on. We have tried to combine these points of view and have created a new manner of project education based on all mentioned fundamental ideas, processes models, analysis and design of functionality and data model for the IS software development.

Students find themselves in the position of a manager of a virtual company/firm. They have to create the project of specifying users' requirements of modern IS and therefore they have to think about the proper settings of main process and proper IS of "their" company, how to differ from the others/competitors, and which innovations they should bring.

Sequence of activities provided currently is presented in Table 1. The recommended structure of the students' projects consists of parts which correspond to ARIS methodology according to [16] and [3]. In these parts we focused on computer science fundamentals including creation proper models, algorithm i.e. process, abstraction, dividing problems into the parts, and systematic approach. We also consider social aspects, working in computer networks, students' experience, and possibility to develop creativity, algorithmic, computational, and critical thinking.

TABLE I. CURRENT SEQUENCE OF ACTIVITIES

	Content and activities
	Implemented fundamental principles and ideas
Reasoning, motivation	IS and processes around us - discussion, examples, pointing out their growing importance. Description of experience with IS. How does it work? Real IS used, pros, cons, benefits, risks, what is convenient, what does not work, what should have change …
	Abstraction, why the software development is needed – motivation, critical thinking, social aspects.
Get basic information	We created and used two studying materials: Information systems for economists and Basics of business process modeling and also provide three a face-to-face lectures in the business informatics field and IS and one lesson to familiarize with ARIS Expres modelling tools. The functionality of LMS Moodle environment is used.
	Language, modelling, social aspects, systematic approach.
Choose project area	What are you interested in? Think, observe, criticize, summarize the current state, consider how to do it better. Feedback via Moodle on the students choice of topic of the project.
	Process of software development as a quide for learning.
Selfstudy e-earning	Familiarize with basics in the field odf IS, analyse and modelling of business processes, best methods, project structures, and possibilities of modelling and description. Consultation the parts of the project.
	Algorithmization, language, proper models creation.
Process analysis and modelling	Analyze the selected process - describe in a text form its characteristics and express in graphical form sequence of activities. Use the ARIS standards or UML diagrams. Think about the model the functional and organizational structure of the business, the products and their parameters, and the structure of applications to support the business. Consider who, when, why, divide into pieces, select only relevant information, and model in understandable way. Communicate, ask, consider possibilities, incorporate thoughts, ideas, use modern ICTs.
	Abstraction, modelling, structured dissection – dividing proces into the parts, algorithmic thinking.
User requirements formulation	Think and formulate: name of IS, formulate the problem, briefly describe the essential objectives (what, why), define the system users (who, when, how), list the most important parts and functions of the IS, find relationship with the existing systems and the requirements for the infrastructure (hardware, software, networks) describe methodology of development, security and reliability of IS (data protection), set supposed deadlines, costs, and implementation process (prototyping , in parts, trial operation), describe requirements for documentation and training, perspectives of IS, next development and maintenance. Formulate your thoughts, ideas, expressed them in a clear and transparent way. Present your solutions and suggestions.
	Abstraction, modelling, structured dissection, division into the parts, social aspects, systematic approach.
Use Case diagram	Specify requirements for system functionality by use case diagrams: create a list and brief description of the actors, the use cases, and scenarios for the use cases.
	Abstraction, modelling of funcionality, algorithmic thinking, structured dissection, division into the parts.
Class diagram	Specify the basic classes of objects appearing in the description of IS. Describe the attributes of these classes of objects suitable for creating relational database tables. Deal with logical structure of the data base system.
	Abstraction, data modelling, structured dissection, division into the parts.

IV. DISCUSSION

In our new way of education we have implemented following fundamental ideas and principles:

- Constructivism.

The constructivism ideas appear to be an effective starting point to change transmissivity and passivity in education. In accordance with [7], we try to overcome the traditional school, which is seen as a definite transfer educational content to students who are in fact condemned to the role of passive recipients. According to [4], we change our function as teachers. We pass from the position of mentors, who shifted knowledge to individuals for the position of facilitators, which facilitates the construction of new knowledge. Our main task is to prepare for students appropriate situations, indirectly control and create situations to construct new facts and knowledge.

In application of constructivist approach we focus on students and consider learning to be an active process. Motivation is the key to learning, experience is essential in the learning process. Learning emphasizes the principle of connections between school and life [8]. As Papert in [14] says, education is important, but learning is more important. Our new education approach is based on preparing and giving students interesting tasks, themes, and issues on which they could work with interest enabling them to learn better.

- Three fundamental Informatics approaches.

As we can read in [6], there are three main/fundamental approaches to Informatics. The first and classical vision corresponds to the automatic processing of information by machines, with the classical vision flow: data, processing and results. The second one is associated with personal computer development. It is the field of interaction with a focus on human activity and a strong interest in human interpretation of machine state. The third one is associated with the development of communication networks, especially the Internet. It is the field of social informatics, based on cooperation between actors or agents and collective intelligence. The three approaches complement one another are incorporated in our process of development and elaboration of students projects.

Mainly the first approach is adopted in the Informatics courses. The four concepts (machine, algorithm, information, and language) correspond to the vision where machines process data and human beings are not a part of the process. They are involved before in preparing data, or after in interpreting the results but not during the process. Informatics education should take into account also the second and third views. Students should understand how interaction is part of the information process, and how informatics products integrate this [6].

- Development of algorithmic and computational thinking.

Algorithms (central informatics concepts) are tools for developing and expressing solutions to computational problems. The development, use, and analysis of algorithms are the most fundamental aspects of computing

[19]. In the project parts: "Process analysis and modeling" and "Use case diagrams" we try to support development of algorithmic and computational thinking as much as possible. In the implicit form students' work with algorithms - they analyze processes as algorithms empirically.

Algorithmic and computational thinking were rather considered by non-specialists as a problem solving approach inherently connected to computer science than a general approach tailored for all students to use computers and computing skills in solving problems coming from various scientific and applied areas [20].

We agree with [20] "computational thinking could be added to the traditional three Rs: reading, writing and arithmetic as an additional basic skill needed especially by university students – they will be better prepared to choose a future career not only as a computer specialist but also as a specialist in other disciplines, who professionally, according the needs of her/his profession, can use computing methods and tools professionally."

The importance of defining proper user requirements introduces to students the concepts and techniques used in writing programs and the ways in which programs are developed; programming is not the focus of our courses. But via designing, modelling and developing projects students gain insight into how systems and networks operate, to the principles that facilitate their design, and to the effects of systems and networks on people and society. Students become familiar with the many ways in which computing enables innovation in other fields [19].

Students also find out how the creation of software has changed our lives, how digital devices, systems, and the networks can interconnect people and enable them to solve problems.

- Development of creativity and abstraction.

This manner of education will emphasize the creative aspects of computing. Computing is a creative human activity that engenders innovation and promotes exploration. Creativity and computing are prominent forces in innovation. Students create projects - interesting and relevant artefacts with the tools and techniques of informatics. Their education also includes examples of abstractions used in modelling the world, in managing complexity, and in communicating with people as well as with machines. Students work with data by using a variety of tools and techniques to better understand the ways in which data is transformed into information and knowledge. These aspects are similar to [19] where can be found "seven big ideas of computer science". Informatics uses abstraction as a basic principle (central problem-solving technique) and people also use abstraction every day.

- Processing data and information.

Computing enables and empowers new methods of data and information processing that have led to monumental changes across disciplines, from art to business to science. A staggeringly large amount of raw data provides part of the foundation of information society and economy. Via elaborating projects parts students can see how people use computers and computation to translate, process, and visualize raw data, creating information, according to [19].

- Business informatics.

Business informatics, information management, and process modelling are at present very important for proper educated future managers, e.g., as can be found in [1], [12], [13], and [21]. Business organization is a system designed by people for the purpose of creating products and services for customers. Computers have become critical to the efficient functioning of all kinds of businesses; it is hard to find a workplace that is not computerized in some way. When we view a company as a system within an environment, each of the basic system concepts takes on a specific meaning. An IS is a set of interrelated parts that work together to produce, distribute, and use information products. One purpose of an IS is to provide information on the feedback and control function of a larger system. For example, an IS system for a business organization produces feedback about the performance of the firm and distributes that information to managers in such a way that is useful for decision making.

- Using computer networks.

Digital devices and the Internet have had a profound impact on society. The principles of systems and networks are also critical in the implementation of computational solutions. Computer networks support both communication and collaboration. Computation has changed the way people think, work, live, and play, as we can read in [19].

V. CONCLUSION

Informatics education of our students during the second semester is set as creating a logical concept of enterprise IS, which includes the structure of the enterprise information processes, the organizational structure, the basic structure of the application architecture, IS procedures, role of users, requirements of hardware and network infrastructure, security requirements, possibilities of IS/ICT further development, documentation and staff training requirements, operation and continuous improvement of processes. Students have to solve tasks involving fundamental ideas and concepts of informatics – algorithms as processes flow, modelling of processes and states, data structures as attributes in creation class diagrams, human computer interaction as a use cases, and so on.

Our manner of future managers´ informatics education tries to create the right mindset and share understanding by explanation that the "digital business is not just about expanding the use of technology. Digital business leaders must think about technology in a fundamentally different way than in the past. It is not an enabler to be applied to what the business wants to do but a source of innovation and opportunity for what the business could do [5]."

We try to assist and support students in becoming knowledgeable participants in information systems management, in doing proper plans and decisions in digital business. In our limited conditions by which we

mean number of lessons, level of students´ knowledge, material and human resources - becoming a knowledgeable participant means learning the basics and feeling comfortable enough to ask questions.

No text will provide managers with everything they need to know to make important IS decisions. Some texts instruct on the basic technical background of IS, some texts discuss application and their life cycle, some texts describe how technologies are designed, used, and integrated into business life [15]. That is why a different approach is taken. We try to start the dialogue/discussion with students of how Informatics and implementation of its fundamentals will help, hinder, and create opportunities for companies, organizations, and institutions.

"Our students will probably face several paradigm changes in their future careers and life. Their acquired school knowledge and skills might become obsolete within a short time. That is why they must be robust enough to meet the challenges of the latest fashion, and also enable the students to cope with changes. This competency is possibly achieved best if students and teachers are given a sketch of the fundamental ideas, principles, methods and ways of thinking of computer science, instead of trying to teach them an incoherent set of recent technology-driven developments. Only the fundamentals seem to remain valid in the long term and enable students and teachers to acquire new concepts successfully during their professional career [18]."

There are strong ties between education system and society. Schools prepare children, young people and adults not only to live and work at present, but also to live and work in the future. That is why it is necessary to anticipate changes in society, and the concept of education and training should be designed so as to be able to respond to changes. The changes that are taking place in the 21st century are likely to outperform the changes that took place throughout the last millennium, in regards to results, the speed, scope and significance. Thus, educators need new approaches, a new point of view, new technologies and applications that go beyond.

REFERENCES

[1] G. Beekman, G. R. Brent, E. J. Rathswohl, Computer confluence, exploring tomorrow´s technology, The Benjamin Cummings publishing Company, Addison Wesley Longman, Inc, USA, 1997, ISBN: 0-8053-2467-4

[2] Y. Bertrand, Soudobé teorie vzdělávání, Praha: Portál, 1998. p. 87, ISBN 80–7178–216–5 (in Czech)

[3] R. Davis, E. Brabander, ARIS design platform: Getting started with BPM, London, Springer-Verlag, 2007, ISBN-10: 1- 84628-612-3.

[4] P. Doulík, J. Škoda, "Tvorba a ověření nástrojů kvantitativní diagnostiky prekonceptů a možnosti jejího vyhodnocení" in Pedagogika, 2003, vol.53, issue 2, p.177-190, ISSN 3330-3815

[5] Gartner: Six Key Steps to Build a Successful Digital Business, STAMFORD, Conn., May 21, 2014 [online] Available at: http://www.gartner.com/newsroom/id/2745517

[6] P. Hubwieser,"The Darmstadt Model: A first step towards a research framework for computer science education in schools," in Informatics in Schools, sustainable Informatics education for Pupils of all ages, 6th International Conference on Informatics in schools: Situation, Evolution, and Perspectives, ISSEP 2013, Oldenburg, Germany, Februeary/March 2013, Proceedings, Springer, ISBN 978-3-642-36617-8 DOI 10.1007/978-3-642-36617-8

[7] Z. Kalhous, O. Obst, et al, Školní didaktika, Praha: Portál. 2002. ISBN 80-7178-253-X. (in Czech)

[8] Z. Macek, "Konstruktivizmus v pedagogické praxi – základní východiská, " In: Vzdělávat učitele. Příspěvky o inovatívní praxi. Bratislava, 2001, (in Czech)

[9] Ministry of Education, Science, Research and Sport of the Slovak Republic: Rámcový učebný plán pre gymnáziá so štvorročným a osemročným štúdiom s vyučovacím jazykom slovenským, 2011-7915/18752:1-922, 2011, [online] Avaliable at: http://www.statpedu.sk/files/documents/svp/gymnazia/rup3_sjog.pdf

[10] Ministry of Education, Science, Research and Sport of the Slovak Republic: Digipedia – Koncepcia informatizácie rezortu školstva s výhľadom do roku 2020, Bratislava 2013, [online] Avaliable at: https://www.iedu.sk/digipedia/Documents/4796.pdf

[11] National Institute for Education. (2012). The national education program for high school ISCED-3a (Štátny vzdelávací program pre gymnázia ISCED-3a), 2012 [online] Avaliable at: http://www.statpedu.sk/files/documents/svp/gymnazia/vzdelavacie_oblasti/informatika_isced3a.pdf

[12] D. Paľová, "ICT tools developed for support the accounting education " in Aplikácie informačných technológií. N°. 2, 2012, p. 44-59. ISSN 1338-6468

[13] D. Paľová, M. Vejačka, "On-line e-learning platform supporting education and practice of accountants in EU space" in Mipro 2013, 36th international convention : conference proceedings : May 20-24, 2013, Opatija, Croatia. Rijeka, p. 771-776. ISBN 978-953-233-074-8

[14] S. Papert, "Constructionism vs. instructionism: speech to an audience of educators in Japan", 1980, [online] Avaliable at: http://www.papert.org/articles/const_inst/const_inst1.html

[15] K. E. Pearlson, Managing and using information systems, a strategic approach, John Wiley and Sons, Inc., United States, 2001, ISBN 0-471-32001-3

[16] V. Řepa, Podnikové procesy. Procesní řízení a modelování, second edition, Praha 2007 Grada Publishing, ISBN 978-80-247-2252-8 (in Czech)

[17] A. Schwill, "Fundamental ideas of Computer Science" [online] Avaliable at: http://ddi.cs.uni-potsdam.de/Forschung/Schriften/EATCS.pdf

[18] A. Schwill, "Computer Science education based on fundamental ideas", 1998, [online] Avaliable at: http://ddi.uni-muenster.de/didaktik/Forschung/Israel97.pdf

[19] Seven big ideas of Computer Science [online] Avaliable at: https://csprinciples.cs.washington.edu/sevenbigideas.html

[20] M. M. Syslo, A. B. Kwiatkowska, "Informatics for all high school students – a computational thinking approach" in Informatics in Schools, sustainable Informatics education for Pupils of all ages, 6th International Conference on Informatics in schools: Situation, Evolution, and Perspectives, ISSEP 2013, Oldenburg, Germany, February/March 2013, Proceedings, Springer, ISBN 978-3-642-36617-8 DOI 10.1007/978-3-642-36617-8

[21] M. Vejačka, J. Bucko, "User knowledge development in the field of electronic services and their security in Slovakia " in MIPRO 2015, Rijeka, Croatian Society for Information and Communication Technology, Electronics and Microelectronics, p. 844-849. ISBN 978-953-233-083

Fostering Creativity in Technology-Enhanced Learning

Anisija Žižić*, Andrina Granić ** and Ivona Šitin***
* Primary school Petra Kružića Klis, Klis, Croatia
** University of Split, Faculty of Science, Department of Computer Science, Split, Croatia
*** Primary school prof. Filip Lukas, Kaštel Stari, Croatia
anisijas@yahoo.com

Abstract - Being "digital natives", contemporary students are already enabled to use almost any kind of technology in the classroom. Due to that fact, teachers should focus on choosing the right tools to help them develop their competences - basic skills, knowledge and attitudes necessary for personal fulfilment and employability in a knowledge society. Critical thinking, creativity, initiative and problem solving are playing big role in achieving eight key competences defined by European Reference Framework 2006. This work focuses on a research in the field of creativity; the definition of creativity is followed by various examples of using computers as a tool which helps us foster creativity in education. Thinking "outside the box" and generating new original ideas are what we have to encourage in students' minds and, in order to accomplish that, we have to make them realize that creativity can be thought and learned. New theories, studies and achievements in the field of creativity are presented with the emphasis on technology-enhanced learning.

Key words: creativity, education, computers in education, technology-enhanced learning

I. INTRODUCTION

Over the past few decades there has been a virtual explosion of topics, perspectives and methodologies in creativity literature. Understanding of the psychology of creativity has grown amazingly sophisticated, but the researchers in one subfield often seem entirely unaware of advances in another [1]. This led to overall fragmentation of the creativity field. Why is creativity so important? "In an era where information and knowledge are becoming shared commodities, available anywhere and anytime, the generation of new ideas is not anymore a luxury, but the very essence of our existence. It is, therefore, necessary to understand creativity in all of its many derivations and fields of applications, and create an educational and operative culture that gravitates around this unique human skill [2]." Among the themes applied in Recommendation of the European Parliament and of the Council of 18 December 2006 on key competences for lifelong learning (2006/962/EC), creativity also found its place. Seeing that creativity is an essential 21st century skill and that not enough efforts are taken to better understanding of creativity, in 2011 European Commission's Directorate - General Information Society and Media - Future and Emerging Technologies launched a consultation process with researchers on "Creativity and ICT" in order to identify new challenges and opportunities for the future [3]. In 2013 EU started financing projects that address creativity and the tools and environments in which it takes place. Goals were: (i) equipping different industries with more effective creative tools, (ii) expanding the potential of technology in the human creative processes and (iii) advancing the scientific understanding of creativity (providing the basis for future innovative technologies) [4]. Once one acknowledges the rich and growing set of possibilities for connecting technology, minds and action, research on intelligent computational environments stimulating and enhancing human creativity is of great interest to education professionals. The term Technology-Enhanced Learning (TEL) is used to describe the application of Information and Communication Technology (ICT) to teaching and learning [5]. Using technology to help other people learn is complex and needs careful planning, but used in a right way can fruitfully enhance all three kinds of learning (implicit, informal and formal) together with their combinations. Technology can be used to tighten or slacken the bonds between perceiving, learning, knowing and action; it can provide scaffolding for unsteady attempts at tackling new problems; it can be designed to shape our behaviours in ways that we only subsequently come to understand and endorse, or reject [6]. If we add creativity to TEL and get tools that support this combination for free, it would be a great opportunity for educators to add some value to their students' education with minimum effort and adaptation. Using technology is a great part of student's life, so it is natural to presume that developing creativity, among other skills, by using TEL will be easily adopted.

II. CREATIVITY: DEFINITION AND THEORETICAL BACKGROUND

According to Corazza, founder of Marconi Institute for Creativity (MIC), "a quick jump out of the box is more insightful than a lifetime of standard thinking" [7]. This thinking "outside the box" relates to creative thinking, and furthermore to creativity. Creativity is present since ever, but efforts to understand it and describe it scientifically are not so old. Modern creativity research began in the 1950's and 1960's [8]. Early attempts to define creativity go to Guilford who characterized the two phases of creative process as divergent thinking - the ability to produce a broad range of associations to a given stimulus or to arrive at many solutions to a problem, and convergent thinking - the capacity to quickly focus on the one best solution to a problem [9]. Torrance (1966)

recognized that cognitive variables such as ideational fluency, originality of ideas, and sensitivity to missing elements are in the basis of creativity phenomenon. He developed Torrance Tests of Creative Thinking (TTCT) that measure creative thinking in verbal and nonverbal ways [10]. Although there has been no consensus regarding the definition and assessment of creativity [11], the most widely used working definition of creativity is that "creativity is the ability to generate ideas and artefacts that are new, surprising and valuable [3]". From mathematician's point of view, a symbolic equation for creativity could be represented as $C = f(KIE)$, where creativity is generated by the interaction between Knowledge (K), Imagination (I), and Evaluation (E), while the most important factor in this formula is the individual's attitude (a) [12]. Additionally, Boden (2004) observed three ways in which creativity is realized:

1) *combinatorial creativity* produces unfamiliar combinations of ideas (e.g. analogy);
2) *exploratory creativity* is based on some well-defined space of thinking. The space is explored, looking both for previously unreached places and to understand its potential and limits;
3) *transformational creativity* is when the space itself is transformed by altering some of its dimensions; so, ideas are generated that could not be generated before the transformation.

Since Rhodes proposed 4P's conceptual model in 1961, thinking about creativity has been structured in four separate strands which represent the essential cornerstones for any kind of creativity research: Person, Process, Press and Product (see Table I) [14].

Recently, Glăveanu (2013) rewrote this model by using terms that explicitly endorse a systemic, contextual, and dynamic approach in trying to achieve a more comprehensive and unitary perspective on creativity (see Table II). Glăveanu's 5 A's framework - Actor, Action, Artefact, Audience, Affordances - is presented to express author's attitude of need for radically changing the way we theorize and study creative acts.

TABLE I. EXPLANATION OF 4P'S IN RHODES CONCEPTUAL MODEL OF CREATIVITY

Term	Explanation
Person	Covers information about personality, intellect, temperament, physique, traits, habits, attitudes, self-concept, value systems, defence mechanisms, and behaviour.
Process	Applies to motivation, perception, learning, thinking, and communication.
Press	Refers to the relationship between human beings and their environment. This notion and the word "press" are rather common in the field of education.
Product	When an idea becomes embodied into tangible form it is called a product. The word idea refers to a thought which has been communicated to other people in the form of words, paint, clay, metal, stone, fabric, or other material.

TABLE II. GLĂVEANU'S REWRITING OF 4P'S TO 5A'S OF CREATIVITY

The four P's of creativity		**The five A's of creativity**
Focus on:		Focus on:
Internal attributes of the person	**Person → Actor**	Personal attributes in relation to a societal context
Primarily cognitive mechanisms	**Process → Action**	Coordinated psychological and behavioural manifestation
Features of products or consensus around them	**Product → Artefact**	Cultural context of artefact production and evaluation
The social as an external set of variables conditioning creativity	**Press → Audience** **Affordances**	The interdependence between creators and a social and material world

According to Kaufman and Beghetto (2009), most investigations of creativity tend to take one of two directions: everyday creativity ("little-c"), which can be found in nearly all people, and eminent creativity ("Big-C"), which is reserved for the great. Consequently, they proposed a Four C model of creativity that expands this dichotomy adding the idea of "mini-c," creativity inherent in the learning process, and "Pro-c", the developmental and effortful progression beyond little-c that represents professional-level expertise in any creative area.

Furthermore, creativity can also be a collaborative process: groups of people can collaborate in creative work, becoming more productive together than any of them would be separately. Connecting groups of people together effectively can stimulate more creativity, especially when rituals of participation (such as networking, familiarity, group dynamics) are taken into consideration for building the trust necessary for creativity to blossom in group environments [3]. Walton (2013) suggests that groups of people contribute to innovation, but creativity is derived from a single idea generated by a single individual. It is a function of "trans-domain ideation" and requires looking at the world in a different way to others. So, we should try to understand creativity in the individual case before broadening to the social case and collaborative aspects.

Amabile and Hennessy published review of creativity in 2010, as they realized that there is a need for a systems view of creativity and believed that more progress will be made "when more researchers recognize that creativity arises through a system of interrelated forces operating at multiple levels, often requiring interdisciplinary investigation [17]". To visualize major levels at which these forces operate and to organize their review, simplified scheme was made (simplified because levels in it don't cross and in reality they do) (see Figure 1.)

Term "Creativity" could be also distinguished as everyday creativity, historical creativity, psychological creativity, group creativity, non-human creativity, technological creativity, artistic creativity, spontaneous

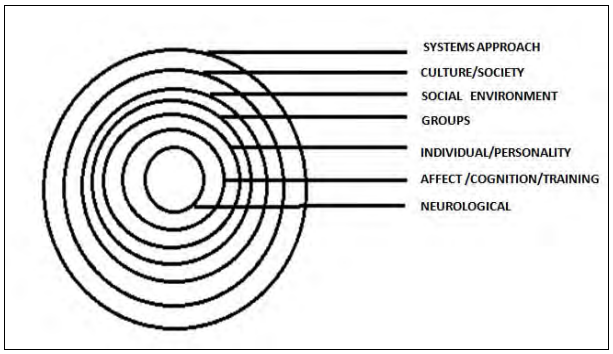


Figure 1. Concentric circles represent the major levels at which creativity forces operate [17].

creativity etc. Because of the multiplicity of definitions, many questions remain open in the field of creativity research: the priority of person-centred or systemic methods, quantitative or qualitative approach to the phenomenon of creativity, is creativity general or a context-specific phenomenon, can animals or machines also be creative and so on. Brodbeck (2006) advocates the notion that not the development of creativity itself should be the primary focus, but rather under which conditions creativity can originate.

III. PROMOTING CREATIVITY IN EDUCATION BY USING TECHNOLOGY

Children are born students, driven by their curiosity. But, most of them don't enjoy in schools. Education systems kill children's creativity. Results of longitudinal study conducted between 1968 and 1985 in USA [19] say in favour of this statement: 1600 five-year old children were given a test on their ability to think divergently and the same children were tested when they were 10 years old and again when they were 15 years old (see Table III). 98% of 5 year olds are "highly creative" and by the time they are 12, this percentage goes down to only 12% of "highly creative" children. The table presents also results of 280 000 adults tested with the test based on a NASA test that measures divergent thinking in engineers and scientists.

The change we need for the future of education is to stop teaching children for the tests. Margaliot and Kreitler (2013) came to conclusion that higher creativity might be achieved by encouraging the use of inner-world

TABLE III. TEST RESULTS FROM THE LAND AND JARMAN STUDY [19]

Age group tested	Number of subjects	Year of testing	"Highly creative"
5 years	1600 children	1968	98%
10 years	1600 children	1973	30%
15 years	1600 children	1978	12%
25+ years	280 000 adults	1985	2%

dimensions such as feelings, thoughts and emotions. The study that Subbotsky (2013) made with 5 to 6 years old children (it was repeated with 7 to 8 year olds) showed that exposing children to a movie with the impossible, fantastical creatures and events facilitates their creativity in realistic thinking to a significantly larger extent, than exposing the children to the movie with equally interesting and exciting, but ordinary creatures and events. This study implies that movies with fantastical content can be an important source of enhancing creativity and that thinking of the impossible can be viewed as a resource for boosting creativity in children. Using TRIZ and OTSM-TRIZ (Russian acronyms of, respectively, Theory of Inventive Problem Solving and General Theory of Powerful Thinking) methods and tools in course "Problem Solving and Inventive Design" held at Politecnico di Milano improved students' Self-Efficacy in solving Inventive Problems, which is crucial aspect for facing challenging and uncomfortable situations (e.g. learning of new concepts, creative application of knowledge, appropriately dealing with unexpected situations) [22]. So, the good thing is that research on applied creativity in the field of education shows us the potential of using different Creativity Support Tools (CSTs) and ways we can improve teaching and learning to promote creativity in the classroom.

Number of EU funded projects started in 2012 and 2013 aiming to make an impact on TEL by providing an open source service-set for social creativity able to mash-up with existing learning processes and solutions (e.g. COLLAGE [23]); to create a modern environment (interactive walls and specially developed applications) to assist creative minds to concentrate on the development and communication of new ideas and concepts (e.g. IdeaGarden [24]); to shed new light on, propose and test concrete ways in which our current understanding of creativity in education and creative thinking, on the one hand, and TEL tools and digital games, on the other hand, can be fruitfully combined to provide young learners and their teachers with innovative opportunities for creative learning (e.g. C2Learn [25]); to design and develop an intelligent computational environment to support the production of media content for educational purposes with the focus on social creativity to enhance creativity in mathematical thinking (e.g. MCSquared [26]); to explore the consequences of exploiting novel technologies in a deliberate attempt to improve creativity and develop domain specific creativity measurement and enhancement (e.g. CREAM [27]).

Research in creativity has led to development of Creativity Support Tools (CSTs), as means for the enhancement of creativity beyond the classic psychological methods, which can, in combination with learning models, provide creative learning environments. Some of the newest CSTs that are available for educational uses are presented in Table IV. Whether the teacher chooses brainstorming to motivate students to produce new, unordinary ideas (e.g. HatParty, Explore and Expand); problem solving from relevant person's perspective to provide students with new point of view

TABLE IV. DESCRIPTION OF CREATIVITY SUPPORT TOOLS (CSTS) AVAILABLE FOR EDUCATIONAL USES (CSTS DEVELOPED WITHIN EU FUNDED PROJECTS (2012-2016))

Creativity support tool	**Description**
Cruise ++	Creative User Centric Inspirational Search (Cruise ++) is a search tool that relates concepts, ideas and elements in order to provide user with surprising or inspirational search results related to a topic. It can be used to find unexpected information, visualize relevant information from the social chatter (e.g. Twitter) and use them as clues for exploring the Web or receive personalised recommendations for inspirational information and exploration actions (http://cruise.imuresearch.eu/ui/explore).
HatParty	HatParty is a real-time team based game for brainstorming which helps different teams to generate new ideas and evaluate them. Hat Party uses game mechanics such as timed sessions, goals, reputation points, rating, racing and social recommender such as inspirational search to help groups of 3 to 8 people generate large number of ideas and rank them in record time (http://hatparty.eu/).
BeCreative	BeCreative is a web-based creativity toolbox that contains advice on many different problem-solving techniques, from problem definition to idea generation, selection and implementation. Users of BeCreative (1-15 participants) can describe their problem and retrieve relevant techniques complete with descriptions, step-by-step guides and usage information to help organize and run workshops or team-working sessions (http://becreative.city.ac.uk/).
Bright Sparks	Bright Sparks is a tool that searches the web to provide problem solvers with support from famous and relevant characters; it provides a new focus on your problem. The tool is suitable for unstructured problem solving with open questions and answers, and can be used effectively in a just few minutes (http://brightsparks.city.ac.uk/).
4Scribes	4Scribes is a creative story-making game for groups of 4 players (20-30 minutes). To play it, group needs one Android tablet which will be shared among co-players. The story starts with a short starting statement or phrase. The objective is to develop the story together with co-players by playing in turns. At the start, players get some cards ("Creative Elements") that have a word or short phrase on it. When it is one's turn to play, he chooses one of his cards and this means he has to take the word or phrase appearing on the card into account when continuing the story. Words or phrases may seem "irrelevant" to the story, but that's exactly the point: the cards are there to "disrupt" story, to make one think differently (creative thinking seeds to spark imagination) (http://www.c2learn.eu/index.php/lets-be-creative/tell-the-story-4scribes/).
Creative Stories	Creative Stories is a free collaborative writing game for generating ideas for innovative scenarios of action (10-30 minutes). Creative Stories invites you to write creatively: the aim is to use as much of the creative advice the Wizard is giving as possible, so that you win more points while you are writing. It can be played in single-player or multiplayer mode. There is also a time limit. Wizard usually decides to offer words that you find rather difficult to incorporate in your text, unless you think really creatively (his intention is to help you think "out of the box") (http://www2.c2learn.eu/index.php/lets-be-creative/write-creatively-creative-stories/).
Iconoscope	Iconoscope is a game deployed on tablets and on the web that gets you to play with visual creativity and ambiguity (10-15 minutes). The inspiration behind Iconoscope comes primarily from non-digital construction and guessing games. Iconoscope motivates players to creatively interpret concepts (described linguistically) as icons (depicted visually) which convey the same message. You can play Iconoscope as part of a group of 4-12 players or alone. When you finish with your visual creation and submit it, it is time for guessing and voting. Once everyone has voted for all creations, you get your scores (http://www.c2learn.eu/index.php/lets-be-creative/draw-and-guess-iconoscope/).
Explore and Expand	Explore and Expand is a simple playful activity inviting you to take a break and think creatively; quickly exploring the conceptual space, in order to brainstorm new ideas (a few minutes). The Wizard is asking you to create a simple conceptual diagram - starting from a given concept in the centre, think of four other concepts that are linked to it from a specific perspective. Your score increases each time you get it right, but if you write the same concept that you have written before you will lose a point. To help you, the game gives you an article from Simple Wikipedia that is relevant to the theme you are playing (http://www.c2learn.eu/index.php/lets-be-creative/explore-the-concept-explore-and-expand/).
GuessWhat	GuessWhat is a multiplayer game (2 to 5 players) that invites you to guess what is shown in a picture and generally think through images, taking into account the perspectives of other players. During a game session, you are presented with a concept expressed in words, and an example of a pictorial representation of that concept. The game asks you to provide your own depiction of the concept using the drawing tools provided. The other players have to guess the concept that you have depicted by selecting the right one in a list of four concepts (you get points if at least one other player correctly guesses the concept you have visualized, but the points you are awarded are equal to the number of players that have NOT guessed correctly). The aim of the game is to have you create visualizations of the concept that are not trivial but are nevertheless coherent and relevant (http://www.c2learn.eu/index.php/lets-be-creative/guess-the-concept-guess-what/).
House of emotions	House of Emotions is a multiplayer game (2 to 5 players) which aims to teach us to "train" ourselves in emotive judgments through 30 minutes of expressing and interpreting emotions. There are four types of missions (Dramatize an emotion!, Draw an emotion!, Write to evoke an emotion!, Color an emotion!) that players need to carry out to turn on all the lights in the haunted house before ghost appears. Beside the missions, random events can influence the progress (fire, extra light on etc.) (http://www2.c2learn.eu/index.php/lets-be-creative/into-the-house-of-emotions/).
C2Space	C2Space is a gameful social digital space designed to foster co-creativity in learning. Gameful learning C2Experiences are structured in Creative Quests, Creative Missions, and Creative Challenges. Three of the previously described games (4Scribes, Iconoscope and Creative Stories) are fully integrated into this web-based environment to offer a unified experience (users' creations and creative learning experiences are stored for sharing with peers and reflecting or getting feedback on them). C2SpaceGames and activities for creative play with words, images and emotions, as well as a social space for co-creative classrooms and groups were developed (http://www.c2learn.eu/index.php/lets-be-creative/create-together-c2space/).
The C-book	The C-book Environment is available as a web app on the internet in several languages. The c-book software provides an authorable dynamic e-book infrastructure that empowers teachers, as creative designers, and students to engage in activities that ignite Creative Mathematical Thinking. Final version of the C-book Environment consists of two parts: C-book Author and C-book Player. C-book Author is an authoring environment, in which teachers and designers can collaboratively design c-books. C-book Player is the environment that students use to read and interact with the c-books and for teachers to access their students' work (http://www.mc2- project.eu/index.php/technology).

(e.g. Bright Sparks); making story with the words that are hard to embed, in collaboration with other students (e.g. 4Scribes, Creative Stories); using C-book to ignite creative mathematical thinking or to simply use inspirational search tool (Cruise++), it could be in favour of changing patterns which students use every day in the classroom. Providing students with new perspectives and presenting CSTs to them to spark their imagination, stimulate ideation, to show them different way of thinking when finding a solution to a problem is very valuable to their future – to be creative in every aspect of their lives. It might seem as a small step, but every journey begins with the first step.

Current research into the TEL in collaborative settings, primarily suggests tools that focus on either a specific aspect of the design process (e.g. support systematic idea generation or structured concept visualization) or a specific phase of it (e.g. brainstorming). To move beyond what presently may be achieved with these tools, their combined use needs to be explored (for example combining mind mapping [28] and story writing, brainstorming [29] with lateral thinking [30] and critiquing, etc.). Jahnke, Haertel and Wildt (2015) found that not creativity itself can be taught – but the culture, the boarder "environment", teaching methods, learning scripts and the "structures" can be arranged in such way that it fosters learners to become creative. An optimal creative learning culture includes six aspects developed from empirical findings [31]:

1) fostering independent, self-reflective learning;
2) Enabling the ability to work autonomously;
3) supporting a (research) curiosity, increasing motivation to learn;
4) fostering constructive learning, where users create something;
5) fostering a new culture of thinking (taking multi-perspectives);
6) enabling the opportunity to develop new original, entirely new ideas.

When the context is right, creativity techniques can play their role in raising the level and type of creativity within organization; creativity is 80% context and 20% techniques [32].

New generations of students ("digital natives") have higher expectations of digital collaboration and the quality of their digital experiences. They prefer 3D virtual world to "traditional" distance education system, because it offers them more emotionally involved experience and feeling of connectedness to learning; there is strong sense of "being there" for the learner [33]. Unfortunately, from the students' perspective, problems such as underpowered or misconfigured computers, poor internet connection speeds and low quality audio/video equipment tended to paint the whole experience in a poor light, and – for some – became a large barrier to the adoption of cutting-edge technology [33]. Even though being at early stage in the development of these sorts of tools, modern web browsers are already capable of rendering highly realistic images, even on tablet devices, so we can expect 3D collaborative environments to become part of future mainstream learning platforms and that virtual worlds could significantly enhance students' ability to engage in collaborative creativity, despite being separated in space and time.

IV. CONCLUSION

TEL uses Information and Communication Technology to secure advancements in learning. By taking advancements as the objective, it goes beyond the attempt to reproduce classical ways of teaching via technologies. Tools and technologies to support distributed learners are likely to become more sophisticated and more prevalent [34]. Apparently the appropriate use of the technologies should result in improvements in learning – making it more effective and more efficient.

The use of creativity techniques in education varies from using paper and pencil (e.g. brainstorming and mind mapping) to adopting ICT-based techniques, usually reflecting culture, teaching style and domain. As individual's creativity depends on deep knowledge in the relevant field, typically built up over many years of study and work and involves not just problem solving but problem seeking, this raises the question what can we do to democratize creativity so that it is available to everyone? Thinking "outside the box" is not enough; we have to build bridges in thinking so as to move people from their current position to a new position. Facilitating creativity works on two levels: on external influences that inhibit the rise of creativity and on internal factors in the form of self-imposed thinking barriers such as habits and predetermined opinions [35].

According to current models of brain research, the human brain is trained to use patterns and routines. Creative thinking means suspending the use of these patterns. Recommendation to educators who want to foster creativity in learning and teaching by using technology is to choose Creativity Support Tools or their combination appropriate to students' knowledge and needs. These tools are not available in Croatian, but open-source software and the source codes of some of them can be found on Web [36]. If we want our students to be ready for the future, we have to teach them how to be creative and innovate, so they can be flexible and quickly adapt to constant changes.

Creative thinking is not a talent; it is a skill that can be learnt. To motivate educators to foster creativity in their teaching and their students' learning, we conclude this paper by George Bernard Shaw's quote: "Imagination is the beginning of creation. You imagine what you desire, you will what you imagine and at last you create what you will."

REFERENCES

[1] B. A. Hennessey and M. W. Watson, "The Defragmentation of Creativity: Future Directions with an Emphasis on Educational Applications", in Multidisciplinary Contributions to the Science of Creative Thinking, G. E. Corazza and S. Agnoli, Eds., Creativity in the Twenty First Century, Springer Science+Business Media, Singapore, 2016.

[2] Vision of Marconi Institute for Creativity, Web page: http://mic.fgm.it/index.php?option=com_content&view=article&id=26&Itemid=25

[3] Report on FET Consultation Workshop on "Creativity and ICT", Brussels, 28 November 2011, http://cordis.europa.eu/fp7/ict/fet-proactive/docs/shapefetip-wp2013-01_en.pdf

[4] Technical background notes for FP7, Strategic Objective ICT-2013.8.1 "Technologies and scientific foundations in the field of creativity", http://cordis.europa.eu/fp7/ict/creativity/background-doc-call10.pdf

[5] A. Kirkwood and L. Price, Technology-enhanced learning and teaching in higher education: what is 'enhanced' and how do we know? A critical literature review", Learning, Media and Technology, 39(1), pp. 6–36, 2014.

[6] P. Goodyear and S. Retalis: "Learning, Technology and Design" in Technology-Enhanced Learning: Design Patterns and Pattern Languages, vol. 02, P. Goodyear and S. Retalis, Eds., Sense Publishers, Netherlands, 2010.

[7] G. Corazza, "Creative thinking - how to get out of the box and generate ideas", speech at TEDxRoma, 2014., https://www.youtube.com/watch?v=bEusrD8g-dM

[8] R. K. Sawyer, "Explaining Creativity: The Science of Human Innovation", 2nd ed., Oxford University Press, 2012.

[9] J. P. Guilford," Creativity", American Psychologist, vol. 5, pp. 444 – 454, 1950.

[10] K. H. Kim, "Can We Trust Creativity Tests? A Review of the Torrance Tests of Creative Thinking (TTCT)", Creativity Research Journal, vol. 18, no. 1, pp. 3–14, 2006., http://people.uncw.edu/caropresoe/GiftedFoundations/SocialEmotional/Creativity-articles/Kim_Can-we-trust-creativity-tests.pdf

[11] A. P. Walton, "Creativity and a Human Dichotomy: Individual or Part of a Team?", MIC Conference 2013, http://mic.fgm.it/index.php?option=com_content&view=article&id=98&Itemid=41

[12] "What Is Creativity?" , Official Web page of International Center for Studies in Creativity, Buffalo State, http://creativity.buffalostate.edu/what-creativity

[13] M. A. Boden, "The creative mind: Myths and mechanisms", 2nd ed., London: Routledge, 2004.

[14] M. Rhodes, "An Analysis of Creativity", in Phi Delta Kappan, vol. 42, no.7, pp. 307–309, 1961.

[15] V. P. Glăveanu, "Rewriting the Language of Creativity: The Five A's Framework", Review of General Psychology, vol. 17, no. 1, pp. 69–81, 2013.

[16] J. C. Kaufman and R. A. Beghetto, "Beyond big and little: The Four C Model of Creativity", Review of General Psychology, vol. 13, pp. 1-12, 2009.

[17] B. A. Hennessey and T. M. Amabile, "Creativity", Annual Review of Psychology, vol. 61, pp. 569-598, 2010.

[18] K. H. Brodbeck, "Neue Trends in der Kreativitätsforschung" (New trends in creativity research), Psychologie in Österreich, H. 4&5 (26), pp. 246-253, 2006.

[19] R. Steinbeck, "Building Creative Competence in Globally Distributed Courses through Design Thinking", Comunicar, Media Education Research Journal, 2011, http://www.revistacomunicar.com/pdf/preprint/37/02-PRE-11702-Steinbeck-EN.pdf

[20] A. Margaliot and S. Kreitler, "A new approach to promoting creativity in 11-14 year-old children", MIC Conference 2013, https://www.youtube.com/watch?v=J_JczUiWpnY

[21] E. Subbotsky, " „The Impossible as a Stimulator of Creativity in Children", MIC Conference 2013, http://www.docfoc.com/the-impossible-as-a-stimulator-of-creativity-in-children-eugene-subbotsky-reader

[22] N. Becattini and G. Cascini, "Improving Self-efficacy in Solving Inventive Problems with TRIZ", in Multidisciplinary Contributions to the Science of Creative Thinking, G. E. Corazza and S. Agnoli, Eds., Creativity in the Twenty First Century, Springer Science+Business Media, Singapore 2016.

[23] Official Web page of COLLAGE: http://projectcollage.eu/

[24] Official Web page of IdeaGarden: http://idea-garden.org/ideagarden/

[25] Official Web page of C2Learn: http://project.c2learn.eu/?q=node/84

[26] Official Web page of M C Squared: http://www.mc2-project.eu/

[27] Official Web page of CREAM: http://www.ict-cream.eu/objectives/

[28] T. Buzan, "What is a Mind Map?", official Web page: http://www.tonybuzan.com/about/mind-mapping/

[29] M. Feinberg and C. Nemeth, "The "Rules" of Brainstorming: An Impediment to Creativity?", IRLE Working Paper No. 167-08, 2008., http://www.irle.berkeley.edu/workingpapers/167-08.pdf

[30] E. de Bono, "Lateral thinking", official Web page: http://www.edwdebono.com/#!lateral-thinking/ge9l2

[31] I. Jahnke, T. Haertel and J. Wildt, "Teachers' conceptions of student creativity in higher education", in Innovations in Education and Teaching International (journal), 2015.

[32] P. Cook, "The Creativity advantage - is your organization the leader of the pack", Industrial and commercial training, vol. 30(5), pp. 179-184, 1998.

[33] C. Burnett, J. F. Cabra and A. Burnett, "Towards Frictionless Collaboration: Teaching Creativity in a 3D Virtual World", MIC Conference 2013, https://www.youtube.com/watch?v=J_JczUiWpnY

[34] A. Granić: "Intelligent Interfaces for Technology-Enhanced Learning", in Advances in Human-Computer Interaction, P. Shane, Ed.,Vienna, Austria: I-Tech Education and Publishing KG pp. 143-160, 2008.

[35] I. Jahnke: "How to Foster Creativity in Technology Enhanced Learning", in B. White, I. King and Ph. Tsang, Eds., Social Media Tools and Platforms in Learning Environments: Present and Future, Springer, pp. 95-116, 2011.

Teaching physics in primary schools with tablet computers: Key advantages

Vladimir Grubelnik
Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia
vlado.grubelnik@um.si

Lidija Grubelnik
Osnovna šola Sladki Vrh, Sladki Vrh 8/a, 2214 Sladki Vrh, Slovenia
lidija.grubelnik@guest.arnes.si

Abstract - In the past two years the project "introduction and application of e-learning and e-services, e-schoolbag" aimed to equip schools with ICT infrastructure and tablet computers in Slovenia. The article presents the key advantages of using tablet computers in primary school physics teaching and learning. Tablets promoted the internet information source and changed the role of teacher who becomes more like a coach who leads and steer learning process. Using the example of falling bodies we show how tablet computers with suitable software were able to substitute the experimental equipment and enables better individualization and differentiations of the experimental school work. Tablet computers and LMS is also perfect didactical tool for acquiring the didactical feedback on students' experimental performance. It proved to be good background for the synchronization of teaching content and schedule. The key advantages of the tablet computers were recognized in the possibility of studying the effects of every individual parameter in the dynamic experiment which is almost impossible in the real experiment.

I. INTRODUCTION

Teaching methods are today's priority in the education. Development in the area of social, technical and multimedia changes learning environment and implies the changes in the TLS (Teaching Learning Studying) processes with the student centered approach principles. Different means for information gathering; optimized learning; and problem solving strategies lead to the self-constructed learning processes [1–5]. Student is more active in the creation of the learning process and teacher becomes the mentor who guides the learning process.

Most changes were contributed to the increasing use of tablet computers in the education. Students individually access the web in order to acquire information from the different (non)educational environments and were not restricted to the computer classrooms to do that anymore. Learning in the different environments; new learning concepts; and use of different online services (web 2.0, web 3.0) enables full scale blended learning. Social learning platforms and blended learning on different levels [6,7] changed the TLS processes and learning is not limited anymore to the social coupling student-teacher in the traditional educational institutions. Creative learning and problem solving strategies which can be experienced and enhances in the home environment becomes increasingly important. Especially in learning contemporary physics topics, e.g. about liquid crystals [8–11] or biophysics [12], at different level in schools and university. Blended learning and flipped classroom slowly and steadily increase their presence in the traditional learning environments and gives increasingly good learning outcomes [7,13–16]. Inevitably teachers' roles change in the changed reality. Teachers are not the only source of knowledge anymore. Their knowledge is used for mentoring and steering the learning process to provide helping hand when needing and warn the learners of faulty conclusions and bad information sources with the aim of find the solution of the problems faced by the learners. Traditional teachers' knowledge should be enhance with the ICT didactics [17] and teachers ability to actively participate in the production of the e-learning materials [18]. Important role plays the high quality of visualization [19].

The introduction of the tablet computers in the schools is made through the project "introduction and the use of e-learning materials and e-services" under the project framework "e-school bag" [20] financed by Ministry of Education, Science and Sport; National education institute Slovenia; and European Social Funds.

The outcomes of the project have shown that motivation of both teachers and students to use tablet computers are high despite the concerns of the lack of e-learning materials supporting some learning areas. In the beginning the benefits of using tablets computers were shown in the natural science courses where e-learning materials are available. Later the use of tablet computers spread to the other courses from different scientific disciplines. Today the e-learning materials for use on tablet computers are available for almost all study courses and disciplines [21].

In our article we are going to present the key advantages discovered with the introduction of tablet computers in the physics courses in the primary schools. Data were acquired in the primary schools during the teaching practices. The focus of our research was the information acquisition over the web using tablet computers as experimental tool and use of applications for studying physics system dynamics.

II. RESEARCH METHODS

All our experimental data were gathered on one primary school which were equipped with tablet computers. We chose this school because it is the project partner of the project "introduction and the use of e-learning materials and e-services" (e-school bag project) [20]. School was equipped with 36 tablet computers (windows tablets) and ICT infrastructure from the ARNES (The Academic and Research Network of Slovenia).

Between 2013 and 2015 we observed the work of 46 students from 9th grade of primary school, we monitor the role of the teachers, and based on their interactions and acquired knowledge we extract the key benefits the tablet computer presents in the physics education.

We have focused on three areas of application of tablet computers in the classroom:

- data acquisition
- experimental work
- simulation of dynamics of physics systems.

A. Data acquisition

Tablet computers promote the use of web as primary information source. We wanted to find out what are the differences in the data acquisition to solve physical problems on different taxonomy levels. We were able to specify three areas of communication (search the web using the web browser, use of e-learning materials, and communication with fellow students). We have also focused to the assessing the credibility of acquired data.

B. Experimental work

In the education of physics, the experimental work is the key component and cannot be left out in the study of dynamic systems. Our aim was to discover the advantages the tablet computers offer in the processes of solving these types of physics problems. The focus of our experiment was the free fall and understanding the acceleration in the gravity phenomenon.

C. Simulation of physical systems dynamics

Gravity phenomenon is a nice physics system dynamics problem that can be solved in different environments. Students can repeat experiments even at home. The tablet computer can use programs that simulate this natural phenomenon and enables analytical study. Students can record the fall of bodies and watch video forward and backward, count the frames, measure time, produce graphs, and match their finding with the simulation [22].

III. KEY FINDINGS

The process of introduction of tablet computers to the school work, examples of use of tablet computers, and some key advantages discovered in the research are shown on introductory video of the web site of the project (http://url.sio.si/kWD). The most important advantage of using tablet computers is in its mobility, because we are not limited to work only in computer class-room. This gives us opportunity to execute different tasks in numerous environments simultaneously: experimental exercises, using tablet computers to search and collect information and to examine simulations regarding the experiment. Tablet computers can be implemented also as experimental tool to measure the time or to create video, this is of high importance in study of dynamical physical systems.As previously mentioned the reports are in the three topics where the discussion on key advantages of the use of tablet computers in physics education are expressed in details.

A. Data acquisition

Students receive different physics problems and during their work on solution we monitor their activities. Data was gathered for 46 students and are presented in table 1. Teacher monitor students work in the classroom. Students data when working at home solving assignments were gathered from students' activity reports.

Analyzing the process of **information acquisition** by students we discovered that students were quite independent. Most effective way to acquire information proves to be the search engines used also by most of the students (table 1). In this process it is good to separate all known never changing old data from the new data produced from the contemporary research. The first type of data (the old one) were easily discovered by students but the new data proves to be much more challenging task. Students were unable to separate digital garbage from the quality data they discovered. To name a case from the astronomy, students did not have problems finding data about mass of the planets in our solar system but finding how many planets were discovered outside our solar system proves to be much more challenging. Students trust data from Wikipedia the most even if they are false. Most of the students (80 %) did not question the data about the normal atmospheric pressure which is wrong written as the value "6533 kbar" on Wikipedia (https://sl.wikipedia.org/wiki/Zračni_tlak).

For **learning the vocabulary and understanding the natural processes** the search engines prove to be inadequate. The results of search engines were usually correct but on the comprehension levels exceeding the level of understanding of general public and specifically primary school students. In these issues the most successful were e-textbooks (e-learning materials) which are designed for comprehension level of school students. Especially useful were interactive content included in the e-learning materials. The major benefit is also the central point for access to all e-learning materials (https://eucbeniki.sio.si/) which does not only simply the search but also enables intercurricula connections and interdisciplinary learning and presentation of the new and already acquired knowledge from different perspective. We also need to know that all e-learning materials are accessible under the Creative Commons 2.5 Slovenian licence agreement and therefore free of charge. Interestingly students show interest for the some topcis that exceed the content of e-learning materials.



Figure 1. Interactive learning materials for physics [23].

Solving the complex problems is the topic that students perform at home as homework. Interpersonal communication was used even before they use e-communication. But with the availability of tablet computers students e-communication using various programs and the social networks explode (table 1). For the communication students use Skype because they can use speech for the data sharing students use ARNES OwnCloud service "Arnes mapa" (https://mapa.arnes.si). Students were actually become better prepared and the benefits were seen in the classroom as well.

TABLE I. INFORMATION ACQUISITION

		search engine	e – learning materials	Communication (Skype and Facebook)
in the classroom	data acquisition	96%	13%	not allowed
	learning the vocabulary and understanding the natural processes	48%	67%	not allowed
homework	Solving the complex problems	17%	52%	78%

B. Experimental work

To understand the quantities that defines the system dynamics the experimental work is crucial. It requires experimental equipment not available to every individual. The presented assignment of falling body (gravity) we present the way how to use the tablet computer to substitute some experimental equipment. If each student has its own tablet computer, he/she can work at its own pace and system allows individualization and differentiation in the experimental work – previously unimaginable option.

Learning topic was experimentally determining the gravity acceleration of a falling body. Tablet computer were used for time measuring and video analysis of falling body (Figure 2). Students record the video of a falling body and use the program "Tracker" (http://physlets.org/tracker/) to analyze the fall. This principle proves to be highly effective to analyze the dynamics of falling bodies. The complete description of the classroom lecture and work is available on the web site of National Education Institute Slovenia [24].



Figure 2. Using software "Tracker" for video analysis of a falling body.

C. Simulation of physics systems' dynamic

For the analysis of the dynamic systems a visual oriented programs are increasingly used. Programs like Madonna[1], Stella[2] and Dynasys[3] enables us the graphical development of the mathematical model of the processed system (Figure 3). The education grasp this approach in the area of analysis of mechanical systems subjected to the 2nd Newtonian rule [25,26].

We studied the specific case of falling bodies and analyze dynamic systems in elementary school. Our experiments did not neglect the air resistance which is generally omitted in physics for the 9th grade students of primary school because of its complexity. The omission of the air resistance can mislead the students to the irrational conclusions – speed of falling body can become infinite.

Students in our case first review the computer program Berkeley Madonna and then the problem of falling body with the air resistance. Later the students use the Madonna program and create the mathematical model of the falling body system and show the change of speed of falling body as a function of the time (Figure 3).

[1] http://www.berkeleymadonna.com/
[2] http://www.iseesystems.com/
[3] https://www.hupfeld-software.de

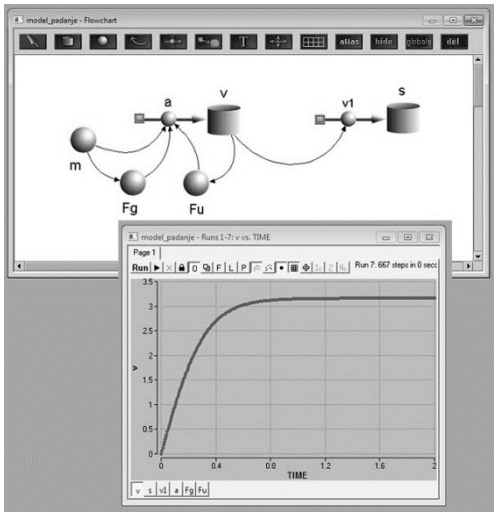


Figure 3. Example of simulation of the falling body with the air resistance (Berkely Madonna).

Teachers guide the students throught the required steps and log their success on the scale from 1 to 5 (1 = student failed to assimilate required knowledge; 5 = student assimilate required knowlede perfectly). Results shows that students have most problems in the writing the 2nd Newton law ($ma = F_{\mathrm{g}} - F_{\mathrm{ar}}$, where gravity force is $F_{\mathrm{g}} = mg$ and air resistance $F_{\mathrm{ar}} = k \cdot v^2$), despite teachers present the problem to them. This is actualy expected because students in the primary schools always have the problems with the matematical notations due to lack of experiences; especialy when describing movement of the bodies under the influence of outside force. Working with Madona software students achieve much better grades (average grade of each step was higher than 4 [24]) which shows this type of learning to be also much more efficient.

One of the advantage of the tablet computer is also the mobility and we are not limited in their usage to the classroom only. Students were in the physics classroom and experimenting with the falling paper cones and compare results with the mathematical model simulation. This enables them to study air resistance $F_{\mathrm{ar}} = k \cdot v^2$ on a falling body.

IV. Conclusion

Based on the practical examples of application of tablet computers in primary school we conclude that tablet computer enables students broader base for the personal knowledge construction. Tablet computer provide "always on" way of life with almost instant access to the information in almost any learning environment, creation and play multimedia content, learning with the interactive learning materials, enhanced connectivity with the classmates and faster communication with the teachers whose role changes rapidly. Teachers in this new environment need to be good source of knowledge, review specialist, guide, mentor, helping hand if needed, and also ICT specialist.

In the physics education and other natural science courses the tablet computer with the right software replaces many experimental equipment. We proved that software for analysis of movement of bodies under foreign forces are highly effective in real and in virtual environment of dynamic system studying. Virtual experiments do not replace the real ones but they enhance them and improve clearness.

Good practice and clearly seen advantages of the tablet computers increase their applicability to other courses and study disciplines in schools. Teachers who started with the concern and disbeliefs of the didactical abilities of tablet computers but managed to overcome initial reluctance accepted tablet computers as good educational tool. In our opinion the tablet computers will spread into the educational system like wildfire and only lack of suitable quality e-learning materials will hinder their inevitable conquest of learning space. New interactive e-learning materials suitable for tablet computers available for all study disciplines will change the special didactics. Therefore, we need to prepare quality interactive learning materials; prepare examples of the good practice to use tablet computers in lectures and students work; and encourage the use of tablet computers through the benefits they present. This must be accompanied with the necessary support for teachers who need additional training and should reflect on university educational study programs as well.

References

[1] T. Keeney, “Self-Constructed Learning: Situations and Technology for Enhancing Lifelong Learning,” *4th International Technology, Education and Development Conference (INTED) Location: Valencia, SPAIN Date: MAR 08-10, 2010* , pp. 5084-5087, 2010.

[2] R. Repnik, A. Sabeder and M. Krasna, “E-learning materials to study properties of liquid crystals at university level and in schools,” *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* pp. 649-652, 2014.

[3] R. Repnik and I. Gerlic, “Liquid Crystals and Development of Natural Science Competences,” *Molecular Crystals and Liquid Crystals,* pp. 168-174, 2012.

[4] R. Repnik, M. Cvetko and I. Gerlic, “Development of Some Natural Science Competences in Undergraduate Study by Training Visualization Skills on Subject Liquid Crystal Phases and Structures,” *Molecular Crystals and Liquid Crystals,* pp. 249-254, 2011.

[5] R. Repnik and V. Grubelnik, “Development of natural science and digital competences by using e-learning materials for third grade of primary-school physics,” *5th International Technology, Education and Development Conference (INTED) Location: Valencia, SPAIN Date: MAR 07-09, 2011,* pp. 1593-1600, 2011.

[6] A. Wallace, “Social learning platforms and the

flipped classroom," *International Journal of Information and Education Technology,* vol. 4, pp. 293-296, 2014.

[7] L. F. Cerdà and C. N. Planas, "Facebook's Potential for Collaborative e-Learning," *Revista de Universidad y Sociedad del Conocimiento (RUSC),* vol. 8, no. 2, pp. 197-210, 2011.

[8] A. V. Dubtsov, S. V. Pasechnik, D. V. Shmeliova, S. Kralj and R. Repnik, "Controlled Nanoparticle Targeting and Nanoparticle-Driven Nematic Structural Transition," *Advances in Condensed Matter Physics,* 2015.

[9] R. Repnik, A. Ranjkesh, V. Simonka, M. Ambrožič, Z. Bradač and S. Kralj, "Symmetry breaking in nematic liquid crystals: analogy with cosmology and magnetism," *Journal of Physics-Condensed Matter,* 2013.

[10] R. Repnik, V. P. Nita and S. Kralj, "Mixtures of Nanoparticles and Liquid Crystal Phases Exhibiting Topological Defects," *Molecular Crystals and Liquid Crystals,* pp. 115-122, 2012.

[11] M. Krasna, R. Repnik, Z. Bradac and S. Kralj, "Sudden isotropic-nematic phase transition within a plan-parallel cell," *Conference: 8th European Conference on Liquid Crystals (ECLC 2005) Location: Sesto, ITALY, FEB 27-MAR 04 2005,* pp. 127-135, 2006.

[12] R. Markovic, M. Gosak, R. Repnik, S. Kralj and M. Marhl, "Defects in Planar Cell Polarity of Epithelium: What Can We Learn from Liquid Crystals?," in *Advances in Planar Lipid Bilayers and Liposomes*, San Diego, CA, USA, Elsevier Academic Press Inc., 2014, pp. 197-217.

[13] N. Petrovic, V. Jeremic, M. Cirovic, Z. Radojicici and N. Milenkovic, "Facebook vs. Moodle: what do students Really think?," in *ICICTE 2013*, Crete, Greece, 2013.

[14] L. J. Bishop and A. M. Verleger, "The Flipped Classroom: A Survey of the Research. American Society for Engineering Education," in *120th ASEE Annual Conference & Exposition*, Atlanta, Georgia, USA, 2013.

[15] C. Peklaj and B. Vodopivec, "Effects of cooperative versus individualistic learning on cognitive, affective, metacognitive and social processesin students," *European Journal of Psychology of Education,* vol. 14, no. 3, pp. 359-373, 1999.

[16] C. S. Du, T. Z. Fu and Y. Wang, "The Flipped Classroom–Advantages and Challenges," in *International Conference on Economic Management and Trade Cooperation*, Xi`an, China, 2014.

[17] M. Krašna and B. Bedrač, "ICT didactics: The new study discipline is needed," in *MIPRO*, Opatija, Croatia, 2013.

[18] M. Krašna, M. Duh and T. Bratina, "E-learning next step - learning materials for students," in *MIPRO*, Opatia, Croatia, 2014.

[19] M. Milfelner, M. Ambrozic, M. Krasna, M. Cvetko, A. Zidanšek and R. Repnik, "Visualization of Nematic Director Field With the RGB Color System," *Molecular Crystals and Liquid Crystals,* pp. 50-57, 2012.

[20] Zavod Republike Slovenije za šolstvo, "E-šolska torba," SIO - Sovensko izobraževalno omrežje, n. d.. [Online]. Available: http://projekt.sio.si/e-solska-torba/. [Accessed 06 02 2016].

[21] L. Grubelnik and V. Grubelnik, "Key advantages of using the tablet in different school fields at elementary education stage," in *Kaj nam prinaša e-Šolska torba*, Kranjska Gora, Zavod RS za šolstvo, 2015, pp. 242-247.

[22] M. Krašna and T. Bratina, "Didactical aspects os using web and video in contemporary learning material," in *Computational methods in circuits and systems application*, Corfu, Greece, WSEAS, 2003, pp. 289-292.

[23] Zavod Republike Slovenije za šolstvo, "Interaktivni učbeniki za fiziko," Ministry of Education, Science, and Sport, 29 1 2016. [Online]. Available: https://eucbeniki.sio.si/.

[24] L. Grubelnik, "Pospešeno gibanje in drugi Newtonov zakon," in *Kaj nam prinaša e-Šolska torba II*, Ljubljana, Slovenia, Zavod RS za šolstvo, 2016, pp. 39-43.

[25] V. Grubelnik and L. Grubelnik, "The use of graphic computer programs for mathematical modelling of dynamical systems in primary school," in *SIRIKT 2011*, Kranjska Gora, Slovenia, 2011.

[26] V. Grubelnik, M. Marhl and R. Repnik, "Modelling of realistic dynamical systems and development of natural science competences in education," in *International Conference on New Horizons in Education*, Famagusta, Turkish Republic of Northern Cyprus, 2010.

Project based learning (PBL) in the teachers' education

Marjan Krašna
University of Maribor, Faculty of Arts, Department of Pedagogy Maribor, Slovenia
marjan.krasna@um.si

Abstract: For a long time, the education means gaining knowledge and skills but this have changed and todays education is more focused on gaining competences. The competences are much more aligned with the concept of learning for life. In the 2012 we started the course Information support for didactical strategies for the 2nd year of postgraduate study program of general pedagogy. We have gathered students reports – their didactical feedbacks; and interviews to assess the students' abilities to cope with real professional problems. All students' assignments have common goals: good design; development; and implementation of students' projects. Topics of collaboration, communication, project managements, and assessment strategies are studied and used in practical students' work. Students are required to provide their own view where and how specific topics can or could be applied in known didactical strategies. Four years' study provide the interesting insight into our students' characteristics. We discovered some:

- **known facts (good knowledge of their study discipline, campaign learning),**
- **surprising facts (preferences to the simple e-learning materials PDF or Word),**
- **alarming facts (no self-initiative, cannot comply with time constraints, misunderstanding of quality concept), and**
- **interesting facts (cannot work in school, hate to talk in front of the audience, they do not "know" their schoolmates).**

I. INTRODUCTION

The study program of general pedagogy, at the Faculty of Arts, University of Maribor, have many elective courses. Students on the 2nd year of postgraduate study can choose the course of Didactics strategies information support. The contents of the course are:

- Didactical strategies for different education levels.
- Didactical strategies influence to the educational concepts.
- ICT support for didactical strategies.
- Test cases of ICT support for didactical strategies (research teaching, project teaching, team teaching ...)

In 2012 the course started for the first time. The first group of students show that they have huge theoretical knowledge about the didactical strategies and they have already attended teachers training in schools. They were totally aware what their work are going to be. The general pedagogy is the interesting study discipline. It is the only non-teaching discipline in the area of education. Graduates from the general pedagogy are equally suitable for jobs in the educational institutions and in the non-educational institutions. Many corporations like banks, insurance companies, hi-tech companies, … have in-house training programs where general pedagogy graduates excel. Therefore, the knowledge they gain during the study should not be limited to the educational institutions only (primary, secondary, and vocational schools). Never the less the main stream of employment of graduates of general pedagogy are educational institutions and therefore it is appropriate that additional optional competences are available as elective course.

The course of Didactics strategies information support give emphasises on different types of narratives. The good structured story from the start to the end is the backbone of every enterprise. The narrative encapsulates all types of professional activities: education, construction, research, and development. We also acknowledged that the important part of the course is also the teamwork. A Project-based learning (PBL) was therefore applied as suitable for our needs.

> *Project-based learning (PBL) is a model that organizes learning around projects. According to the definitions found in PBL handbooks for teachers, projects are complex tasks, based on challenging questions or problems, that involve students in design, problem-solving, decision making, or investigative activities; give students the opportunity to work relatively autonomously over extended periods of time; and culminate in realistic products or presentations* [1]

The team formation is the fundamental process for the team work. We know from the experiences that students make teams around "natural leaders". But the real world constraints change natural events. The newcomer almost always come into the established structure and the leaders are not always those who are emphatic, giving, understanding, and rewarding persons. The team formation is not an easy task. Studying the algorithm for selection of the members for the students project teams [2] gives us the highly valuable insight into the team formation processes. Because we have educational studies that generally consist of two major academic disciplines, freely elective by each student upon the enrolment to the Faculty of Arts, we have students' activity scheduling problems. It is practically impossible to schedule the courses for each student not to overlap with some of their other activities. To partially solve these problem students are obliged to attend the laboratory and the seminar work but not the lectures. The rationale behind this decision is the fundamental meaning

of study: "*an act or effort made in the pursuit of knowledge*", which is possible to achieve studying of the course literature. Attendance on the lectures are therefore free choice of the students despite the fact that it significantly correlates with the grades.

We form the team on the first laboratory work where we explain the role of the team and the team leader, their requirements, course outcomes, and influence on grades. A student who is in the position of the team leader becomes aware that he/she is responsible for the outcome, presentation of the team work, and his/her report is a part of the grade of all team members. All students are aware that the team leader grades them in his/her report and these grades presents the 10% of their course grade. When all facts are discussed and all responsibilities are cleared students were given the access to the LMS Moodle where they formed the teams. The first fortnight assignments are then given to the students' teams. The teams can restructure after their first assignments on requests of majority of the team members but afterwards the teams stay firm till the end of course.

The course assignments are aimed to cover different students' competences from the EU key competences framework [3]. To promote competences, we need to design the education (learning activities) and show the benefits of the teamwork [4]. We also need to prepare assignments that suits our needs to fulfil the course objectives according to the contemporary findings [5] [6] [7]. We wanted to give our students the best choice to express their creativity [8] and the team-teaching using ICT [9].

Each year we fulfil temporal proximity according to the contemporary events in the World with the engineered education [10]. This is possible because we separate the methodologies and techniques apart from the content of the students' projects. Learning strategies have not changed much since they become the educational focus in the past centuries [11] [12]. What did changed are the learning tactics: peer learning [13], problem-based learning [14], media assisted learning [15], creativity in education [16] and learner-cantered education [17].

II. THE COURSE DIDACTICS STRATEGIES INFORMATION SUPPORT

The Didactics strategies information support course started in 2012 and was continuously elected by enough students to run continuously till now. The course worth 6 ECTS and consist of 30 hr lectures, 30 hr laboratory work, and 120 hr individual work. It is mainly selected by the students of Pedagogy but other students could also select it because of the intra-faculty courses elective process.

TABLE 1: NUMBER OF STUDENTS ATTENDING THE CURSE BY YEARS

Year	# of students
2012	15
2013	23
2014	22
2015	28

The course is engineered to address most of the key competences from the EU framework. To achieve these goals, we have to structure the lectures and the laboratory/individual work appropriately. In the table 3 and the table 4 we present the associations between competences and lectures and students home/lab work. The table 2 is for the reference purpose.

TABLE 2: THE KEY COMPETENCES FROM THE EU FRAMEWORK [3]

	Competence
KC1	communication in the mother tongue
KC2	communication in foreign languages
KC3	mathematical competence and basic competences in science and technology
KC4	digital competence
KC5	learning to learn
KC6	social and civic competences
KC7	sense of initiative and entrepreneurship
KC8	cultural awareness and expression

TABLE 3: THE LECTURES' TOPICS

	description	**EU key competence framework**	**focus**
E-communications:	It is most widely used ICT service in education.	KC1, KC4, KC6	formal/informal communication, text/audio/video, e-mail/bulletin board/chat.
Project management	Introduction to the project management task, assessment of human effort, costs, schedule, planning, and implementation	KC1, KC3, KC4, KC5, KC6. KC7	Using Microsoft Project and Excel. See the limitations of Excel in the project managements tasks.
Software tools for scheduling events	Searching for the time when event could occur, outside established schedule, proves to be the most problematic between students.	KC1, KC3, KC4	Using common tool for communication, using special scheduling online services (doodle, ARNES planer, group calendar google, and Microsoft OneNote)
Software tools for simultaneous work on documents	To understand the importance of version control. The rules to work with the documents on the web	KC1, KC2, KC4, KC5, KC7	Google docs, Microsoft OneDrive, Microsoft OneNote Wiki
Learning Management Systems (Moodle)	See the differences between roles in LMS, prepare for being teacher in LMS and prepare learning materials and quizzes	KC1, KC2, KC4, KC5,	Moodle (ver. 2.9) basics Quiz library HotPotatoe Choice
Interactive technologies	Using interactive learning materials Using interactive boards	KC1, KC2, KC4, KC5,	Smart notebook OneNote Interactive PowerPoint

TABLE 4: THE LABORATORY WORK AND STUDENTS INDIVIDUAL WORK TOPICS

	description	**EU key competence framework**	**focus**
Research team work	Work division, search and analyse available electronic sources, use software tools for simultaneous work on the report, presentation	KC1, KC2, KC4, KC5, KC6, KC7, KC8	Contemporary topics for ICT in education (i.e. mobile learning, cheating in the classroom, e-learning vs distance learning)
Project – educational trips	Project management tasks (schedule, planning, budget), presentation, dissemination	KC1, KC2, KC3, KC4, KC5, KC6, KC7, KC8	Walkthrough the home town with oral presentation
Electronic classroom	e-learning materials, quizzes, assignments, communications (forum/chat/videoconferences)	KC1, KC4, KC5	Individual arbitrary topics
Preparing promotional materials	Creative use of software tools for image and text processing	KC1, KC2, KC4, KC5, KC6, KC7, KC8	Prepare poster, flyer or leaflet for arbitrary event.
Micro teaching	Using available software tool to present arbitrary topic.	KC1, KC4, KC5, KC6, KC7, KC8	Individual arbitrary topics

III. DATA ACQUISITION AND RESULTS

A. Observation

It is interesting that we see repeated behavioural pattern in the use of the computers. We continuously failed to persuade, influence, or retrain users not to maximize each program they use. Even students (digital natives) could not provide the answer why they always maximize every software application they work with even if their computer screen is big (more than 22 inch full HD resolution or higher). Explaining them and show them that they could have two open windows side by side (word and web browser) and could work much more efficiently did not change their behaviour. It would be interesting to assess the lost hours because of such users' behaviour.

The students' report and students- digital products were used to assess further parameters. From the methodological standpoint it is not desired to address multiple attributes from the textual sources since we could have the problems with the validity. Therefore, we decide to assess (1) individual work, (2) schedule, and (3) student's impression. To study the correlation between parameters we use the students' course grades. The grades are presented for year 2012, 2013 and 2014. Data for year 2015 is only partially available therefore we omit them.

B. Individual work

Video recordings of the students' lectures are particularly important for their job because they have to know their voice and posture when they are going to address the audience. Reviewing their recording they get the feedback on how the other people see and hear them and they can improve their appearance. Students are required to have 10-minute lecture because this time is too long for them to successfully achieve the goal without the careful prepared narration and scenario.

Each student has to made his/her presentation on arbitrary topic but they have a choice to present it in the class where we record him/her; or he/she can prepare it at home and upload video of their presentation. Most of the students (around three quarter) choose the second option. Students often express that they do not want to speak in front of the class and being recorded. This reason is quite an oxymoron for the (future) teachers.

This year we have seen that most of the student who did not have classroom presentation just recorded their video in front of the computer and read the text from the screen (presumably). The so called digital natives lack the knowledge of video editing or they are just lazy.

A part of our study was the assessment of completeness of the students' reports. The completeness of students' reports can be a measure of the students' understanding of the **quality**. Students know what is considered the good report, great report, or just a text. To encourage their creativity, we did not provide the blueprint for the reports. Our opinion is that the forms can limit the creativity and unify thinking. This is already shown in the master theses where the structure of the theses become completely rigid and student who want to propose the new way to do something (like prepare e-learning materials for something new) must have the same structure as those who preform statistical analysis of the questionnaires.

TABLE 5: THE GRADING SYSTEM FOR COMPLETENESS OF STUDENTS' REPORTS

Grade	meaning
0	Incomplete report
1	All major parts are present in report
2	Complete report

TABLE 6: THE STUDENTS' REPORT COMPLETENESS.

	grade			
year	0	1	2	Total
2012	3	8	3	14
2013		4	18	22
2014	5	5	7	17
Total	8	17	28	53

The data from the table 6 show that students are generally unable to provide complete reports. This can be the problem of understanding of the minimal standards. Achieving the minimal standards means only positive grades. The year 2013 was very good year for the students' reports. The students were organized and some natural team leaders provide the examples how others should prepare report for them. Therefore, they just continue in this

fashion and prepare their individual reports the same. Despite the good results all reports look the same.

C. Schedule

A part of students' report was the timetable for their work (consists of two general parts: the course work and the students' project work). The team leader has to prepare the report for the project based on the individual students' reports. The time in hours is the measure for the students' effort and show the individual efforts inside the whole team effort. The team leaders could use this measure to grade their colleagues.

Students teams negotiate the way to record the timetable and consequently the students' effort. If the student provides the dates and the short description of the tasks accomplished, we grade him/her with 0. If the student provides dates with hours and short description, we grade him/her 1; and if he/she provides separate complete reports for course and project, we grade him/her 2.

TABLE 7: THE GRADES OF STUDENTS' TIMETABLE REPORTS

	grades			
year	0	1	2	Total
2012	13	1		14
2013		11	11	22
2014	1	8	8	17
Total	14	20	19	53

The data from the table 7 shows that despite there are some advancement in the students' report on the topic of timetable they are still unable to prepare the good report needed for work on the EU projects. The lack of this knowledge may lead to the contemporary political issue of grading the work of the teachers in Slovenia. Should the teacher work 8 hours a day in their schools or could they be trusted that they work the 8 hours a day (or even more) to finish their required daily tasks.

D. Students impression

The assessment of the student's impression of the course can be obtained from their reports. It is always the problem if we want the students to provide feedback to improve the curse to be more suitable to them. The students were told many times that this is required part of their reports but it does not affect their grades. We wanted to get the quality feedback and not gold-plated useless text. The grades of the students' analysis were not available to the students but were only used for the purpose of this research.

TABLE 8: THE GRADING SYSTEM FOR THE STUDENTS' ANALYSIS MEANS:

grade	description
0	No analysis
1	Superficial text just because its needed
2	Constructive criticism of the course topics and the proposed changes for the future

From the table 9 we see that in the beginning we did not receive the good students' feedback. In the year 2013 students provide much better feedback and we use their feedback for the year 2014. In the year 2014 student again provide less desired feedback than the previous year. Because the ICT is subjected to the constant change and it is a part of our life we expect that the students' feedback are going to correlate with the grades the students give to the course. And indeed when the students provide the good feedback a year later the grade of the course improves.

TABLE 9: STUDENTS' GRADE ON FEEDBACK

	grades			
year	0	1	2	Total
2012	4	7	3	14
2013	2	1	19	22
2014	5	3	9	17
Total	11	11	31	53

E. Students course grades

The students have to grade the courses and the end of the study year. From the year 2013 the grades are publicly available on the web. Each student can grade different aspect of the course with grades ranges between -2 and 2. From the table 10 it can be seen that the course is highly graded by the students. This correlates with the impression part of their report (table 9).

TABLE 10: THE STUDENTS' GRADES OF THE COURSE

year	2013/2014		2014/2015	
	Grade	Std. dev.	Grade	Std. dev.
Group 1	1.11	1.28	1.60	0.56
Group 2	1.41	0.64	1.92	0.28

IV. CONCLUSION

In the recent years we managed to successfully accomplished the course Didactics strategies information support for the four generations of the students. Engineering this course proves to be the success but there are still possibilities for the improvement. The concept of the competences was successfully proven before [18] and verified again. Observation and acquired data gives us valuable feedback to the evolution of the course through the years and this prove to be the winning concept.

It proves that the students acquire enough background knowledge about didactics strategies but they need more practical examples. Reports have shown that most of them have clear understanding how to use ICT in the different didactics strategies once they see ICT potentials in the courses' examples and teamwork.

The teamwork proves to be the good for some and the bad for the others. The required changes of the teams' members give us the clue that students discover the compatibility of team's members only after the first team assignment. Logic tell us they do not know themselves enough at the beginning. These are the 5th year students and their preferences for the individual work is highly expressed. It is not yet clear it we want this at the end of their study. They will probably be working in the team environment and team skills are something they will need to acquire for their success. We have already proven that the interdisciplinary teams provide much better results even

without extensive supervision [19]. One to the possible addition to the course would be the team assignments that would involve students from different disciplines. But this would require some restructuring of study programs which are highly intensive processes with many bureaucratic obstacles.

The students prove to be the campaign learners and the campaign workers. This can be seen in the inability to provide the good schedule of their activities. The failure to provide complete students' reports also show they do not fully understand the concept of quality. We are going to need additional effort in the future to address these issues. If they will follow the same principles in their employer environment they are going have huge problems with other team members.

LITERATURE

[1] T. John W., "A review of research on project-based learning," The Autodesk Fundation, San Rafael, California, 2000.

[2] D. Kermek, "Algorithm for selection of members for students project teams," in *MIPRO*, Opatia, Croatia, 2002.

[3] European Union, "Key competences for lifelong learning (Recommendation 2006/962/EC)," EUR-Lex, 2006.

[4] I. de los Rios-Carmenado, F. Rodriguez Lopez and C. Perez Garcia, "Promoting Professional Project Management Skills in Engineering Higher Education: Project-Based Learning (PBL) Strategy," *International journal of engineering education,* vol. 31, no. 1, pp. 184-198, 2015.

[5] I. Hartescu, "Providing technology support for project-based learning," in *10th International Scientific Conference on eLearning and Software for Education*, Bucharest, Romania, 2014.

[6] M. Balkevicius, A. Mazeikiene and S. Svediene, "The first steps of project-based education in Lithuanian high schools," in *2nd World Conference on Educational Technology Researches (WCETR)*, Nicosia, Cyprus, 2012.

[7] I. Fernandez-Plazaola, M. Pons, C. Llinares, A. Montanana and E. Navarro-Astor, "Implementation of project based learning methodologies to improve students affinity to a knowledge area. applying a case study," in *6th International Conference of Technology, Education and Development (INTED)*, Valencia, Spain, 2012.

[8] A. M. Daud, J. Omar, P. Turiman and K. Osman, "Creativity in Science Education," in *Universiti-Kebangsaan-Malaysia Teaching and Learning Congress*, Malaysia, 2012.

[9] L.-C. Chang and G. C. Lee, "A team-teaching model for practicing project-based learning in high school: Collaboration between computer and subject teachers," *Computer & Education,* vol. 55, no. 3, pp. 961-969, 2010.

[10] J. E. Mills and D. F. Treagust, "Engineering education - is problem-based or project-based learning the answer?," *Australasian journal of engineering education,* pp. 2-16, 2003.

[11] H. F. J. O'Neil, Learning Strategies, New York, London: Academic Press, Inc., 1978.

[12] C. E. Weinstein and E. T. Goetz, Learning and study stretegies: Issues in Assessment, Instruction, and Evaluation, San Diego, London: Academic Press, Inc. , 1988.

[13] D. Boud, R. Cohen and J. Sampson, Peer learning in higer education: learning from & with each other, London, Sterling: Kogan Page Limited, Stylus Publishing Inc., 2001.

[14] C. E. Hmelo-Silver, "Problem-Based Learning: What and How to Students Learn?," *Educational Psychology Review,* vol. 16, no. 3, pp. 235-266, 2004.

[15] I. Gerlič and N. Jaušovec, "Multimedia: differences in cognitive processes observed with EEG," *Educational technology research and development,* vol. 47, no. 3, pp. 1042-1629, 1999.

[16] N. Jaušovec and K. Jaušovec, "Brain, creativity and education," *The open education journal,* vol. 4, pp. 50-57, 2011.

[17] S. Aslan and C. M. Reigeluth, "Examining the challenges of learner-centered education," *Phi delta kappan,* vol. 97, no. 4, pp. 63-68, 2015.

[18] R. Repnik and V. Grubelnik, "Development of natural science and digital competences by using e-learning materials for third grade of primary-school physics," in *INTED 2011 : International Technology, Education and Development Conference*, Valencia, Spain, 2011.

[19] A. Wirth and R. Repnik, "Analysis of learning by doing in the creation of instructional video in teaching optics," in *MIPRO*, Opatija, Croatia, 2015.

Didactical suitability of e-generated drill tests for physics

R. Repnik* and M. Sovič*
*Faculty of Natural Sciences and Mathematics, University of Maribor, Maribor, Slovenia
robert.repnik@um.si, magdalena.sovic@student.um.si

Abstract - Contemporary education of physics in primary and secondary schools are very well organized but still not perfect. Learning materials, multimedia learning objects, knowledge testing and grading is driven to the details in Slovenia but the drill seems to be forgotten. Knowledge retention positively correspond with the amount and duration of the drills. Drill should not increase costs of education for students. In Slovenia a project E-va (electronic drill book) produced publically available e-drills for primary and secondary schools [1]. Article present the test of efficiency of e-drills on the population of 1st year students of undergraduate study program of elementary education. Those students are interdisciplinary in nature and knows physics from secondary schools. Two random groups were formed to be tested with the paper drill questions and e-drill questions. We do not expect the difference between groups since both groups are going to have three drills of each type. Differences are going to be influenced only by the type of drill (e-drill or paper drill). Results highlight the motivation factors of ICT and enables us to identify benefits and drawbacks of e-drills. The results will be the part of the requirement specification of the new development of e-drills.

I. INTRODUCTION

In the past years many projects for development of e-learning materials were funded by the European Social Fund and Ministry of Education. Faculties with pedagogical study programs at the University of Maribor participated in almost all of them [2–4]. A lot was learned and a lot discovered after the projects were finished and become operational [5]. Today a lot of electronic data are available (electronic magazines, books, educational computer programs, electronic textbooks etc.) [6–8]. Teachers' opinions on the use of these e-materials in schools differ, however, all agree that a sensible amount of all teaching methods should be used, traditional as well as e-learning [9].

E-materials can be included in the educational process in different ways: in frontal lessons, individual lessons, for homework, pair work, in supplementary and remedial lessons. With these teachers are given a chance to improve the quality of their work. The article focuses on the electronic drill book for physics E-VA.

II. THE ORIGIN OF THE PROJECT

E-VA was made as an expansion to the project E-um [1], which includes the area of mathematics. The project was prepared by high school "Srednja šola Slovenska Bistrica", as a leading partner, and the Faculty of Natural Sciences and Mathematics at the University in Maribor, as a consortium partner in the tender by the Ministry of Education, named Public tender for the selection of operations for the development and execution of consulting and support to schools, e-materials and training of teachers for the use of ICT in teaching and learning for the period 2008–2013. The project was funded by the European Social Fund and the Ministry of Education and Sport. The authors of the e-drill book for physics are teachers with many years of experience. They were divided into groups: the editorial board, the reviewers, the authors of the material, who were at the same time the programmers, and the technical team. The e-drill book was developed between January 1st, 2009 and September 1st, 2010.

Nowadays, there are numerous e-materials on the World Wide Web. Some are good and efficient but many more are not. The e-drill book for physics is an exercise training book, not a textbook. It is intended for teachers and students in elementary school and in grammar schools. Each chapter includes individually reviewed materials, which entirely cover the curriculum. The chapters of the primary school physics are: Introduction to Physics, About Forces, Pressure and Buoyancy, Work and Energy, Work, Internal Energy, Motion, Universe, Electric Current and Electric Charge, Tension and Work, Static Electric Charge, Magnetic Field and Optics. Each of these chapters includes some subchapters. There are 84 altogether. In the part for grammar schools there are 19 chapters; these are: Measurement, Physical quantities and Units, Linear and general motion, Force and Torque, Newton's laws and gravitation, Theory of Motion Quantity, Work and Energy, Fluids, Structure of Matter and Temperature, Internal Energy and Heat, Electric Circuit, Magnetic field, Induction, Oscillation, Waves, Light, Atom, Semiconductors, Atomic Nucleus, Theory of Relativity. Also these chapters have subchapters, namely, 83. At the beginning of each chapter there is a revision of the material, and then drill exercises on different taxonomic levels follow. The material is generally arranged in 5 screen shots with graded difficulty. In the first screenshot (Fig. 1), there is a summary of the learning part and a short motivation. In the second screenshot, there are drill exercises which check minimal standards of knowledge. The third and fourth screenshot include more advanced drill exercises with different animations, simulations, videos and audio recordings. There are more complex drill exercises in the last screenshot [10,11]. Different types of drill exercises

with carefully chosen questions have dynamically generated parameters. Drill exercises include pictures, animations, videos, sounds, applets and other interactive elements. The teacher can use the e-drills as a supplement to (e-)textbooks as well as a tool in active lesson methods.

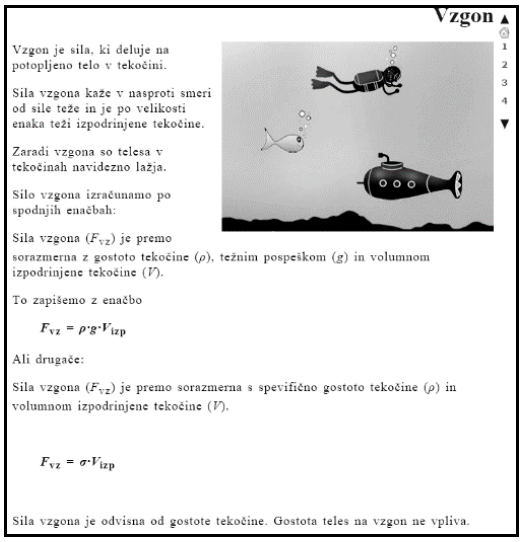


Figure 1. In the first screenshot is the summary of the learning part and a short motivation.

III. RESEARCH

In the research, we tested the first year students of elementary education in the Faculty of Education Maribor, who are generally a good population to test. To some extent they are sociologists, on the other hand, they are also naturalists. The sample of 70 students was divided into two groups. Each group did three topics electronically and 3 topics classically (Fig. 2). The first part included drill exercises from the chapters of Physical units and Quantities, from the chapter of Pressure, and drill exercises connected with the topic of Motion (Fig. 3). The second part included drill exercises from the chapters on Forces, from the chapter of Work and Energy, and Optics. We chose elementary drill exercises from the e- drill book that are static (without the use of animation and simulation), as otherwise we would have influenced the experiment already in principle.



Figure 2. Students solve the tests.

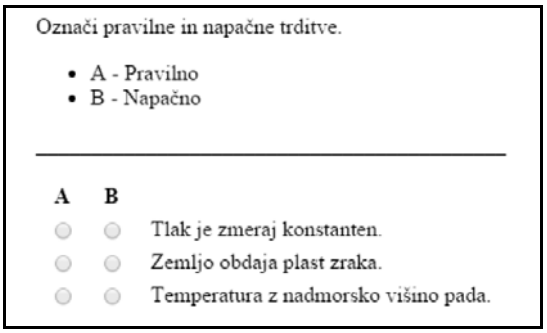


Figure 3. The example tasks in the test. The student should mark the correct answer.

IV. THE RESULTS OF THE RESEARCH

The analyses of the results inside the groups showed huge differences among individual achievements, which indicate considerable differences in the knowledge of students. We estimate that the reason is the fact that the students come from different high school programmes and they had not had Physics lessons in the Faculty till then. Consequently, we need to use the Mann-Whitney test to analyse the collected results.

We compared the times of solving the drill exercises of both parts regarding the method of solving. The results show the existence of a statistically characteristic difference in the time of solving drill exercises of the first part (Table 1), namely, students solving the drill exercises in a classical way used less time than the students solving the drill exercises electronically (U=466,5 in P=0,017).

Table 1. The time of solving drill exercises of the first part of the test.

Time of solving		$\overline{R}$	U	*P*
Method of solving	Classical	31,61	466,5	0,017
	Electronic	43,39		

The solving of the second part in students using the classical method was also faster (Table 2). There is a statistically characteristic difference in the time of solving drill exercises of the second part between the groups, namely, students using the classical method of solving were faster (U=77,5 in P=0,000).

Table 2. The time of solving drill exercises of the second part of the test.

Time of solving		$\overline{R}$	U	*P*
Method of solving	Classical	21,54	77,5	0,000
	Electronic	52,65		

Both examples indicate shorter time in solving drill exercises with a classical approach to drill exercise solving. The reason may be the fact that the e-drills for physics does not enable the teacher the direct insight in the correctness of the drill exercises, therefore, the students had to fill in these formalities on their own. It would be beneficial if teachers received the results immediately.

We compared the number of points reached in both parts, regarding the method of solving. We noticed there was not a statistically significant difference in the number of achieved points of the first part between both methods of solving. However, we can see from the Table 3 that students using the electronic method of solving were a bit more successful (U=537,0 in P=0,109)

Table 3. The achieved points at solving drill exercises of the first part of the test.

The number of point reached		$\overline{R}$	U	*P*
Method of solving	Classical	33,51	537,0	0,109
	Electronic	41,49		

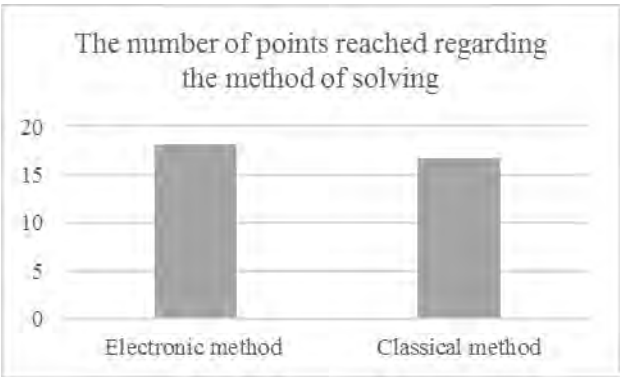


Figure 4. Rate between the achieved points in the first part of the drill exercises.

Solving of the second part of drill exercises was more correct in group of students solving drill exercises compared with a classical method (Table 4). There is not a statistically characteristic difference in the number of the achieved points of the second part between the groups, namely, students using the classical method of solving were more successful (U=494,5 in P=0,125).

Table 4. The achieved points at solving drill exercises of the second part of the test.

The number of point reached		$\overline{R}$	U	*P*
Method of solving	Classical	39,49	494,5	0,125
	Electronic	31,98		

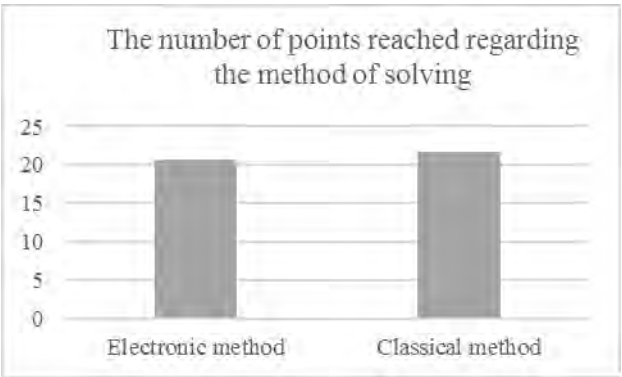


Figure 5: Rate between the achieved points in the second part of the drill exercises.

The biggest disadvantage of the e-drill book is the fact that it is not tolerant enough in accepting the correct results. The rounding of results causes the most problems, as it is usually not stated next to the drill exercise how precise the result should be. The word order of phrases is also important. As an example we can state the drill exercise where the answer "resistance of air (*slo: upor zraka*)" is accepted as the correct one, and the answer "air resistance (*slo: zračni upor*)" is marked as incorrect. In the future, therefore, more time should be devoted to forming the possible answers.

For the applets to work, E-VA requires the program JAVA. It would be more beneficial if the e-exercise book worked without additional programs, as the user would have less technical problems.

V. CONCLUSION

E-VA is a good supplement to e-books which do not include a lot of automatically generated drill exercises for reinforcing knowledge. Students can use it at home or in school. In the research, we established that it is more useful for independent practice at home. E-VA does not enable the direct insight in the correctness of the solved drill exercises of the whole class for the teacher, therefore, it is a bit less useful for checking the students' understanding of the material in class. It provides only the information of the correctness of solving for the individual user himself. However, it can be used as one of the possibilities to make the lessons more interesting, since the materials are well prepared, despite some problems regarding intolerance to differing answers. The results of our research provides us information, that the time difference in solving either in classical or Electronic way was statistically significant. However in the achieved points between this two groups there was no statistically characteristic difference.

REFERENCES

[1] "Project E-VA," 2010. [Online]. Available: http://e-va.e-um.eu. [Accessed 7 2 2016].

[2] R. Repnik, A. Sabeder and M. Krasna, "E-learning materials to study properties of liquid crystals at university level and in schools," *37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),* pp. 649-652, 2014.

[3] R. Repnik, M. Cvetko and I. Gerlic, "Development of Some Natural Science Competences in Undergraduate Study by Training Visualization Skills on Subject Liquid Crystal Phases and Structures," *Molecular Crystals and Liquid Crystals,* pp. 249-254, 2011.

[4] R. Repnik and V. Grubelnik, "Development of natural science and digital competences by using e-learning materials for third grade of primary-school physics," *5th International Technology, Education and Development Conference (INTED) Location: Valencia, SPAIN Date: MAR 07-09, 2011,* pp. 1593-1600, 2011.

[5] T. Bratina, M. Duh in M. Krašna, „E-learning controversy in practical application," v *MIPRO*, Opatija, Croatia, 2013.

[6] B. Kaučič and M. Krašna, "E-learning evolution in the educational institution," *M. Čičin-Šain (ur.), I Turčić Prstačić (ur.) and I. Sluganović (ur.), MIPRO 2008: 29th international covention, May 22-26, 2008, Opatija, Croatia: proceedings, MIPRO, Rijeka 2008,* pp. 225-229, 2008.

[7] R. Repnik, V. Vajngerl and M. Krašna, "Virtual learning environments," *M. Čičin-Šain (ur.), I. Turčić Prstačić (ur.) and I. Sluganović (ur.), MIPRO 2006: 29th international convention, May 22-26, 2006, Opatija, Croatia: proceedings, MIPRO, Rijeka, 2006,* pp. 92-94, 2006.

[8] T. Bratina and M. Krasna, "Trends of e-learning in the education of teachers," *The Fifth International Conference on Informatics, Educational Technology and New Media in Education, Sombor, 2008, Infomedia, Sombor: Faculty of Education,* pp. 65-69, 2008.

[9] R. Repnik, B. Kaučič and M. Krašna, "E-learning in the Modern Curriculum Development," in *eLearning - Theories, Design, Software and Applications*, Rijeka, Croatia, InTech Europe, 2012, pp. 225-238.

[10] R. Repnik, V. Grubelnik and M. Krašna, "The importance of different types of review of physics materials in electronic exercise book," *MIPRO 2011: 34th international covention, May 23-27, 2011, Opatija, Croatia,* 2011.

[11] B. Rebernak, "Pomen IKT in e-gradiv pri pouku v sodobni šoli // The Importance of ICT and E-materials at teaching in modern school," *Vzgoja in izobraževanje,* pp. 91-93, 2008.

.

Utilizing MOOCs in the development of education and training programs

Petri Linna, Timo Mäkinen and Harri Keto
Tampere University of Technology / Pori unit, Pori, Finland
e-mail: petri.linna@tut.fi

Abstract - Open education and distance learning are not new pedagogical innovations. However, through the introduction of Massive Open Online Courses (MOOC), they have recently attracted a great deal of attention among educational establishments. MOOCs can be considered a threat to small universities, but, on the other hand, they can also be a means of providing opportunities to develop their core activities. The challenge is how universities will perceive this phenomenon and take advantage of the new chances it brings. This paper examines the utilization of MOOCs from several points of view. The focus is on degree courses and continuing education offered by universities, but in-house personnel training in companies is also discussed. The issue is how to find proper ways to utilize third-party MOOCs in these three domains. Based on our investigations, the paper introduces a preliminary model for exploiting MOOCs in the development of education and training programs.

I. INTRODUCTION

Massive Open Online Courses (MOOC) refers to courses where there can be a lot of students simultaneously, which are open to everybody, and performed totally on a network.[1] The term open no longer necessarily means that the courses will be totally free of charge. For example, there may be a certificate fee, but the course itself is free. The term open also refers to the fact that anybody can go on the courses, university students, as well as anybody else who is interested in the subject. The supply of MOOCs is already huge in the U.S.A, but the supply in Europe is also increasing [2]. The MOOC concept has a short history but it can be related to areas that have been developed over a longer period like distance learning, virtual learning environments, and blended learning. There is also a weaker link in MOOC to the newer flipped learning term, in which the student becomes acquainted with the subject before the lessons take place.

The University Consortium of Pori is a consortium formed by units of several separate universities in Finland. One of the units is Tampere University of Technology (TUT), Pori Unit, which is the research focus area of this study. The financial crisis and government cuts regarding educational organizations, and changes in educational trends (see examples of trends [3-11]) were some of the reasons that forced the TUT Pori unit to make some radical structural changes. At the same time, student numbers had fallen to alarming levels. The TUT Pori unit started to offer blended learning educational programs in the autumn of 2013. This was a brave but necessary change to survive the competition between universities for students. By the beginning of the first year it had proved successful. [12] About the same time, the open data research project, Avaras, was started. During the project, a very interesting open education concept, MOOC, came to our notice. As a result, new research projects focus areas are to develop open education from different angles. One of these, the Kasku project, will focus on degree education and continuing education and the Ajatar project will focus on companies.

The aim of the Kasku project is to develop teaching in a versatile way in the Pori unit. The project consists of four work packages. In work package one, methods will be developed to utilize MOOCs in the degree programs as well as in continuing education. The aim of the Ajatar project is to develop digitalization extensively in small and medium-sized enterprises. One of the four work packages will examine the use of MOOCs in companies. The project will examine and test some MOOCs and make a MOOC list for the ICT theme. It will be researched whether companies can take courses independently or do they need some sort of guidance services? In addition, course verification requirements will be studied. For example, are the MOOC providers' own certificates sufficient for companies or do they need an examination organized by the university?

In the second chapter of this paper, MOOCs are examined from the degree program point of view. In the following chapter and its subsections MOOCs will be studied from the viewpoint of continuing education and companies. In the fourth chapter, a preliminary model is introduced for utilizing MOOCs in these three domains. The concluding chapter contains a summary and discussion of the presented solutions and some thoughts on the future.

II. THIRD PARTY MOOC SUPPORTING COURSE DEVELOPMENT AND IMPLEMENTATION

At the beginning of last year, TUT Pori unit introduced a three-year project to develop its educational processes. One of the work packages (WP) of the project focuses on MOOCs. The aim of the WP is not to produce its own large-scale online courses but to utilize MOOCs of other universities as part of the department's curriculum. Four ways to apply third party MOOCs have been identified: 1) A MOOC is implemented directly as a standard course; 2) A MOOC replaces a standard course,

usually offered in the local language, for international students; 3) A MOOC or a group of MOOCs is implemented on a course whose content is not completely fixed; 4) MOOCs or their modules are used as parts of a standard course. [12]

This section discusses an analysis of a case concerning activities found in the education process. The case falls into the first category of the MOOC application model cited above. It utilizes a MOOC by the University of Helsinki, which is the pioneer of MOOCs in Finland. [13] At present, their offerings consist of 12 courses, most of which are related to programming [14]. The courses are mainly in Finnish, but three of them are offered in English. One of the courses has been developed by Aalto University, rather than the University of Helsinki.

The aim of the case study is to illustrate which activities in an education process are affected when a third party MOOC is utilized. The structure of the process here is based on the elements identified in the MOOC applied: the Server Side Web Development course by the University of Helsinki [15]. The course can be studied freely without a link to any education establishment, but the course can also be tailored by an authorized organization for its own implementation. The second mode was adopted in this case.

From a student's point of view, the course consists of web-based reading material including an extensive set of programming exercises, which are intended to be resolved with an integrated development environment (IDE). The exercises are associated with project templates, which can be downloaded from the course server using the IDE. A typical template contains a skeleton program and a set of runnable test cases. When a student is satisfied with his/her solution, he/she can submit it using the IDE to the course server, which immediately accepts or rejects the submission based on the tests it runs. If the submission is acceptable, points are registered to the student's account, and he/she can view the suggested solution. If the student encounters problems during the development, the IDE offers a possibility to ask support from a teacher by sending a request for code review.

Based on the short description above, the activities and tasks shown in Table 1 at the very least should be done to develop and implement such a course. A task is considered part of an activity.

TABLE I. Activities and tasks related to course development and implementation.

Course development		Course implementation
Prepare readings - Review texts and technologies - Select topics and technologies - Study selected topics and tech. - Write course readings	Develop exercises - Write exercise specifications - Develop example solutions - Prepare program templates - Develop test cases	Support students - Review code Assess study attainments - Verify submissions

The elements in Table 1 are classified into two categories: Course development and Course implementation. The first category consists of activities that can be performed before the course implementation, while the activities of the second category have to wait until the implementation begins. In the Course development category, there are two activities, Prepare readings and Develop exercises, both of which have a direct counterpart among the deliverables of the MOOC. The Course delivery category also has two activities, Support students and Assess study attainments. The activities are at a more general level than their MOOC counterparts. The tasks of the activities, Review code and Verify submissions, have been derived from the services offered by the course server and associated IDE.

The Prepare readings activity of the Course development category has four tasks. The last of them, Write course readings, is self-explanatory. The other tasks of the activity can be considered implicit predecessors of the first task. In the second activity, Develop exercises, there are also four tasks, all of which are obvious counterparts for the elements found in the MOOC.

When a course is developed and implemented with the aim of having the same properties as the MOOC described shortly above, all ten Course development and implementation tasks listed in Table 1 have to be performed. But, when a suitable MOOC is available, only three of the tasks need to be carried out. The tasks, Study selected topics and technologies and Develop example solutions prepare a teacher with the skills required to support students through code reviews during the course implementation.

III. EXPANDING THE SCOPE

Next we expand the utilization of MOOCs to a larger context than degree education. First, the continuing education point of view is discussed. Because the development of a MOOC is too big an effort to discuss here, we concentrate on the utilization of existing external offerings only. Then, the company viewpoint is discussed. This can also be seen as an opening discussion to the customer point of view, because the success of their business is one important indicator of the eventual outcome.

A. Continuing education aspects

Continuing education and training are characterized by their high quality. For example, MBA programs include managers/executives from organizations, demanding the best possible education. Therefore, the utilization of MOOC courses in continuing education programs should be considered extremely carefully. In the Ajatar project study, it was observed that the quality level of MOOC courses can vary significantly and there are considerable differences in their implementation methods. In order to make the most of a MOOC, it should be evaluated carefully. In other words, the teacher needs to test the content, the usability, and the extent of the MOOC before the implementation. At the same time, this probably reduces the possibility of smaller practical risks.

Continuing education poses challenges due to intensifying price competition. This includes especially the training programs, short courses, and customized

training for organizations. Besides this, there is also public competitive bidding in education. At the same time, companies have reduced investment in training, partly due to the financial crisis and its prolonged effects. MOOCs can be seen to possess the same potential as that found on the side of degree programs. For the attendee, they can include a wider training specialization as an alternative. MOOCs offer the opportunity to carry out special courses that could otherwise cost quite a lot and include too great a financial risk. There is also the potentiality for a new kind of training services, but innovative ideas and open-minded attitudes are needed in order to test and pilot these new services with companies.

B. Companies' point of view

The idea of the MOOC work package of the Ajatar project was to look for new ways for companies to reduce education costs. The education offered to companies in the Satakunta region is intended to be diverse, international, and flexible. It has been identified that companies in Satakunta lack knowledge about MOOC concepts and its opportunities. If so, this means that the digitalization of education is not treated as a strategic issue or possibility for personnel development. A further survey providing deeper knowledge of companies' interest in MOOCs is required.

The university units in UCP are not able to offer a very broad range of international courses in their degree programs. Continuing education cannot offer very special courses to very small groups at a reasonable price. For a company, often the only possibility to train their experts in special skills means travelling to where the training takes place. This causes additional costs, such as travel expenses, accommodation expenses, daily allowances, and any overtime pay as well as the purchase of additional labor. In addition, one of the challenges in training is that the company may encounter a sudden, unexpected educational need for one or two employees. Degree education is too slow for that situation and continuing education services are too expensive to plan and implement for a couple of students. At best, an advantage is gained from the fact that the problem faced and the arrangement of the education meet each other. When supply and demand match, this kind of problem-based learning makes the investment more profitable.

IV. PRELIMINARY MODEL OF MOOC UTILIZATION

The MOOC utilization preliminary model (MUPM) has been developed during the Kasku and Ajatar projects. Both projects include their own development segments.

The starting point of MUPM development was that MOOCs could be utilized in as versatile a way as possible, not only to replace an existing course. This model is in its first development phase, and the next will begin with piloting and further development with the various organizations involved. The development of the model will study how open education resources (OER) [16] and its facilitator models will be developed in general (examples of facilitator experiences [17-18]). The MUPM course list might be expanded in the future to include other OER materials such as Microsoft Virtual Academy [19] or Code Academy training services [20].

MUPM consists of a MOOC utilization table (MUT) and the Mentor model. The MUT contains marked areas where MOOCs can be utilized in different ways. The table is separated into the main columns: degree program, continuing education, and company. These columns are divided into smaller parts, specifying potential utilization targets. All the columns have been given a unique number. There are descriptions of the utilization methods in the rows, which have been given a capital letter. Below is a more detailed explanation of the columns and rows.

A) A third party MOOC directly implements a standard course. B) A MOOC replaces a standard course for international students. C) A MOOC implements a course the content of which is not completely fixed. D) MOOCs or their modules are used as parts of a standard course. E) A MOOC can be used to provide compensatory duties. F) A MOOC can be used for additional tasks or as "nice to know". G) A MOOC can be used for coaching. H) A MOOC can be utilized in development work.

1) Course. 2) Master of Thesis work. 3) Self-study. 4) Ph.D. program and Post-doc. 5) Long training programs. 6) Short training courses. 7) Customized training. 8) Competitive bidding. 9) Project. 10) Undertaken outside the company. 11) Taken within the company. 12) Self-study.

TABLE II. Utilization table of MOOCs.

	Degree - program				Continuing education					Company		
	Course	Master of Thesis work	Self-study	(Ph. D. Program and Post-doc)	Long training programs	Short training	Customized training	Competitive bidding	Project	Undertaken outside the company	Taken within the company	Self-study
	1	2	3	4	5	6	7	8	9	10	11	12
A	P		X	X		X	X	X	X	X	X	X
B	P		X									
C	P			X								
D	P			X	X	X	X	X	X	X	X	X
E	X				X	X		X	X	X		
F	X				X	X	X	X	X	X	X	
G	X	X	X	X	X	X		X	X	X	X	X
H					X	X	X	X	X	X	X	X
M	Y		Y, I	Y		Y	Y	Y	Y	Y	Y	Y

Each entry of the table is described in brief here. "X" means that it has not yet been piloted but the potentiality to utilize MOOCs has been found. "P" means that it has been piloted, and has already been described in chapter one or in a previous paper [12]. "Y" means that it is possible to use the Mentor model in that specific column. "I" means that it might be possible to do a special international and cross-cultural course for international students utilizing MOOC courses. This course can also be offered to TUT Pori unit's own engineering students to achieve better awareness of cross-cultural issues. The first row "A" is perhaps the most important, because it mainly enables the Mentor model.

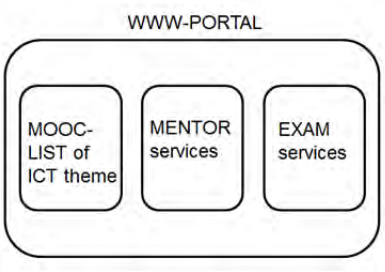


Figure 1. The Mentor model.

The Mentor model (Figure 1) can be used when the MUT table row M is "Y". The Mentor model has not yet been tested but the tests are scheduled to begin this fall. The Mentor model consists of three parts. The MOOC ICT list includes a list of MOOCs on the ICT theme. The Mentor section includes alternative mentoring services, which the company or any other organization can buy to manage the training, for example, a MOOC starting lesson, intermediate check-up lessons, and final lesson. The last Exam section is a verification service.

V. CONCLUSION

This paper has presented a review of the utilization of MOOCs in degree, continuing education, and company sectors. MOOC courses were piloted for two years, and now the research and pilots have also been extended to other sectors. Based on the study, a MOOC utilization preliminary model (MUPM) has been formed, which consists of a MOOC utilization table and the Mentor model. The benefit of the model is to present the various opportunities of MOOCs for teachers, education organizers, as well as companies that are not yet taking advantage of MOOCs.

The aim of the projects is to continue the pilot project during the spring with companies and look for a new product and access to continuing education. Suitable MOOCs will be selected on the ICT theme. The main objective is to start to expand the use of MOOCs at the University Consortium of Pori and its various university units in the spring. First, by adding this model uses as well as possibly to expand the MOOC themes in the fall. The aim is to publish a beta version of the Mentor model website in the spring. In addition, a MOOC survey will be sent to companies, to determine for example the awareness and knowledge level of MOOCs and utilization alternatives. The survey will enable the further development of the Mentor model.

It is quite obvious that the MOOC concepts will develop much further. The problem is how the cost of MOOCs, permission, or the limitations on number of participants will change in the future. The challenge is to know who will win the competition for the platform and course supply and quality. Who will own and handle all the students' data stored on the platform? Will there be a future for new models such as SPOC, i,e., small private online courses. [21]

Overall, the greatest chance a small university center has is to take advantage of MOOCs and utilize them in a blended learning concept. This is a key element in providing students with new forms of flexibility in the implementation of the means of study particularly while in working life. In addition, in the problem based learning can see a lot of educational development and business opportunities. It seems that companies have not yet MOOC or OER training strategy, so extensive development projects is needed and the dissemination of knowledge about MOOCs possibilities.

REFERENCES

[1] Massive open online course. https://en.wikipedia.org/wiki/Massive_open_online_course (Accessed 22-Feb-16)

[2] Open education europe, Scoreboard, http://openeducationeuropa.eu/en/open_education_scoreboard. (Accessed 22-Feb-16)

[3] Andy, H., Ann, L., Michael, F., & David, H. (Eds.). (2009). Second International Handbook of Educational Change. Second International Handbook of Educational Change. Springer.

[4] Dijk, J. (2006). The network society: social aspects of new media. SAGE.

[5] Castells, M., (1996). The rise of network society 1st, Blackwell publishers, Inc. Cambridge, MA, USA.

[6] H. Jaakkola, P. Linna, J. Henno, and J. Mäkelä, (2011). (Social) networking is coming Are we ready? MIPRO, 2011 Proceedings of the 34th International Convention. IEEE, pp. 1133–1139.

[7] H. Jaakkola, B. Brumen, J. Henno, and J. Mäkelä, (2013). Are we trendy? MIPRO, 2013 Proceedings of the 36th International Convention. IEEE, pp. 529–537.

[8] Johnson, L., Adams Becker, S., Estrada, V., and Freeman, A. (2015). NMC Horizon Report: 2015 Higher Education Edition. Austin, Texas: The New Media Consortium.

[9] Horizon Report Europe: 2014 Schools Edition. Luxembourg: Publications Office of the European Union, & Austin, Texas: The New Media Consortium.

[10] Davies, A., Fidler, D., & Gorbis, M. (2011). Future work skills 2020. Institute for the Future for University of Phoenix Research Institute.

[11] Senge, P. M. (2014). The Fifth Discipline Fieldbook: Strategies and Tools for Building a Learning Organization. Doubleday Religious Publishing Group.

[12] Linna P, Keto H, and Mäkinen T, (2015). Towards Blended Learning: A Case Study. MIPRO, 2015 Proceedings of the 38th International Convention. IEEE, pp. 685-689

[13] Hiidenmaa, P. (2013). Jos vastaus on MOOC, mikä on kysymys? (If the answer is MOOC, what is the question?). Report. University of Helsinki. In Finnish

[14] MOOC.fi. University of Helsinki, Department of Computer Science. (Accessed 22-Feb-16)

[15] Web-palvelinohjelmointi. (Server Side Web Development). University of Helsinki, Department of Computer Science. MOOC in Finnish. http://web-palvelinohjelmointi.github.io. mooc.fi/kurssit.html, (Accessed 22-Feb-16)

[16] Open education https://en.m.wikipedia.org/wiki/Open_education, (Accessed 22-Feb-16)

[17] Facilitator. http://education.okfn.org/facilitator/ (Accessed 22-Feb-16)

[18] Open facilitator handbook. http://booktype.okfn.org/open-facilitator-handbook/_draft/_v/1.0/ (Accessed 22-Feb-16)

[19] Microsoft virtual academy https://mva.microsoft.com/ (Accessed 22-Feb-16)

[20] Code academy https://www.codecademy.com/ (Accessed 22-Feb-16)

[21] Fischer G, (2014). Beyond hype and underestimation: identifying research challenges for the future of MOOCs. Distance Education. Vol. 35, Iss. 2, 2014

Distance Delivery and Technology-Enhanced Learning in Information Technology and Programming Courses at RIT Croatia

K. Marasović*[1], B. Mihaljević*[2] and I. Bačić*[3]
* Rochester Institute of Technology Croatia, Zagreb, Croatia
[1] kristina.marasovic@croatia.rit.edu,
[2] branko.mihaljevic@croatia.rit.edu,
[3] iva.bacic@croatia.rit.edu

Abstract - In today's modern education, distance learning methods and tools provide an effective way to offer instructions remotely, particularly when instructions in person cannot be provided. Delivering technology-rich courses that engage students in different group activities such as working on group projects, as well as laboratory and in-class exercises, create new challenges that need to be addressed. This paper discusses the authors' experiences with distance delivery and technology-enhanced learning in several information technology and programming related courses in the Information Technology program of the Rochester Institute of Technology (RIT), at the Croatian campuses in Dubrovnik and Zagreb. The learning environment and the remote course delivery platform include a video conferencing system, a course management system, a remote desktop, and different online services and tools. Various usage scenarios and their impact on classroom learning experience are described. A report on students' feedback and perceptions of real time interactive distance learning is also provided and forms a basis for improving the remote course delivery.

I. INTRODUCTION

Rochester Institute of Technology (RIT) is one of the earliest adopters of distance learning. RIT offered its first distance course in the spring quarter of academic year 1979-80. Courses were offered in the form of telecourses where video materials were prepared and delivered to distance learners over cable television [1][2].

In the mid-1980's, telecourses were enriched with email and online discussion boards that enabled the faculty to apply various instructional techniques such as current topic discussion, case study, reflection on assigned readings, sharing of links, interviews with guest experts.

By 1991, distance courses had completely turned online. To improve the online experience, in 2002, RIT implemented the myCourses course management system [3]. Aside of online access to course materials, myCourses is a rich web-based environment that provides building and handling of many course related activities such as attendance, calendar, chat, dropbox, email and grading.

Past and current technological and global changes have driven RIT to continuously explore new methods of accessibility and delivery to learners. In addition to online learning [4], RIT offers study abroad opportunities at RIT's global campuses in Dubrovnik & Zagreb (Croatia), Dubai (United Arab Emirates) and Pristina (Kosovo) [5].

Delivering courses in Croatia, in particular, RIT's undergraduate IT program, presents distinct challenges. First, there is the difficulty of recruiting faculty (even adjunct faculty) with the necessary expertise to cover distinct sections in both Zagreb and Dubrovnik. Second, while total enrollment may be sufficient, the number of students in one of the locations may be too small to justify the cost of a separate course section. To address these issues while presenting students with a set of offerings that cover both required and elective courses necessitated the use of remote instruction.

Delivering courses remotely, in real-time, works fine when the focus is on information transmission, but it presents many challenges in laboratory settings, where students are expected to apply what they've learned to new, and open-ended problems. Adding to this the necessary interactive guidance required in team-based projects, the desired environment may be difficult to provide [6].

The following sections describe first the institutional learning environment and the distance learning courses offered at RIT Croatia. Next, challenges instructors faced while conducting the remote lectures have been addressed. This is followed by a survey based report on how students perceive real-time interactive distance education.

Much was learned from this experience, and the paper concludes with observations and suggestions for those who may engage in such a project in the future.

II. INSTITUTIONAL LEARNING ENVIRONMENT

The four-year Information Technology program offered at RIT Croatia, in Dubrovnik and Zagreb, prepares students for perspective professions in following essential areas: Programming and application development, Web and mobile development, Database management systems and applications, Networking and system administration, User-centered design and deployment, and Human-computer interaction. Studies at RIT Croatia are also

accredited by the Croatian Ministry of Science, Education and Sports and meet the requirements of the Bologna Agreement. Upon successful completion, all students are awarded with two degrees: an American degree from RIT, and a Croatian degree from RIT Croatia.

The main turning point in the activities of RIT Croatia in terms of distance learning took place in the fall of 2013, when RIT completed a three-year transition from a quarter-based to a semester-based academic calendar. Many programs used this as an opportunity to revise and update their course portfolio. Among other changes, the IT faculty determined that its graduates needed a firmer foundation in basic software design principles and their application. The result was a new course, SWEN-383 Software Design Principles and Patterns. Complicating the development of this new course was the need to use remote instructions to accommodate students at the RIT Croatia's Dubrovnik campus [7].

Driven by mainly positive experiences of that time period, the number of courses offered remotely were increasing. Distance delivery works well in courses focusing mainly on information transmission as in:

- *ISTE-190 Foundation of Modern Information Processing*. The course provides an overview of modern information processing technologies, applications, practices and trends with an emphasis on how these technologies shape information environments and how participants in these environments are able to access, process, and use data and information.

On the other side, courses requiring laboratory settings, where students are engaged in in-class exercises and group projects present distinct challenges and require special attention, as in:

- *ISTE-140 Web and Mobile I*. The course provides students with an introduction to Internet and Web technologies. Topics include, among others, HTML and CSS, Web page design and Web site publishing.
- *ISTE-340 Client Programming*. The course explores the analysis, design, development, and implementation of client-side programming in the context of Internet technologies, mobile devices, Web-based client systems and desktop applications.
- *ISTE-430 Information Requirements Modeling*. The course master contemporary techniques used in information requirements collection and analysis, software design, and process and data modeling.
- *ISTE-432 Database Application Development*. The course examines various aspects of designing and developing larger-scale applications and information systems which use relational databases.
- *ISTE-436 Database Management and Access*. The course introduces relational database server implementation, administration, and management, as well as client-database server communication.
- *SWEN-383 Software Design Principles and Patterns*. The course explains, compares, and contrasts common software design and architecture patterns through analyzing an existing solution, as well as through implementing a solution from the scratch.

Depending on the student enrollments, these courses are usually offered in two sections: one in Zagreb and one in Dubrovnik. Usually sections are being merged into one class, making students enrolled in one of the sections distance learners.

To accommodate new course settings, both campuses provide a special purpose classroom - a computer laboratory equipped with Polycom videoconferencing system. The Polycom system consists of a large video display allowing everyone to easily see and be seen. In addition, cameras are automatically locating and zooming-in on the person speaking [8]. An array of ceiling microphones picks up voices in a 360-degree area, giving focus to the active speaker without physical intervention. To ensure quality audio experience, the Polycom sound system that provides echo cancellation, and noise reduction is set in place.

Whereas this setting creates an atmosphere that is close enough to the experience of being in the same room with the other distant students, additional equipment was required to support screening of instructor's slides and notes. To mirror the instructor's presentation, a projector is being used in both distant and local classrooms.

As for the experiences and challenges instructors faced while conducting the remote lectures, the following section provides the details.

III. Challenges In Remote Course Delivery

Providing a real time, face-to-face experience in an online learning environment presents a difficult challenge that includes several aspects that need to be taken into account. When it comes to course delivery, RIT Croatia takes advantage of the myCourses course management system and a web-based environment that provides building and handling course related activities [3]. This system provides online access to course materials, dropbox folders for assignment submissions, email access, grading rubrics, student-to-student and student-to-instructor chat, and attendance records.

Besides theory based lectures, many class periods center on a theoretical topic introduction, followed with in-class exercises and assignments. Thus, a fully functional computer lab is required at both local and remote sites, with an additional setup that will enable interactive delivery of the course, as well as supervision and guidance of student activities. The additional setup includes using the Polycom video conferencing system [8] to provide a synchronous real-time course delivery offering an interactive two-way communication.

While use of the Polycom can approach the feel of a traditional classroom, it fails when direct oversight and

TABLE I. OVERVIEW OF CHALLENGES IN REMOTE COURSE DELIVERY

Challenge	Choice	
	System/Tools	***Note***
Synchronous real-time course delivery offering interactive two-way communication.	Polycom	Simple to operate; proved useful for small class size - up to 15 students on the remote site
Instructor supervised and assisted hands-on lab experience	Apple Remote Desktop	Useful to control the remote computers. Proper internet bandwidth is needed to avoid latency.
Instructor supervised and assisted work on group projects	Git or Subversion	Great in cases were group assignments are programming based.
Instructor supervised and assisted work on software design in group assignments	Lucidchart	Good choice to support and share the work on visualizing a software design.
Building and handling course related activities: materials, attendance, calendar, chat, dropbox, email, grading	myCourses - RIT's course management system	Requires internet connection – interaction and access to course materials provided online
Providing office hours to remote students	Google Hangouts or Skype	Proved to be a good choice in providing real-time course - related discussion and help

guidance of student activities is required. This is one of the major challenges we faced while delivering a distance learning course. Thus, in addition to the usage of the Polycom system, we decided to use remote desktop technology. Since teaching labs at RIT Croatia are equipped with iMac computers, Apple Remote Desktop [9] was a natural choice that enables full supervision, assistance and control of the remote computers. Since the internet connection is critical in order to avoid latency, a high bandwidth connection has to be ensured.

Although the Polycom and the Apple Remote Desktop compensate well for the physical absence of the instructor, special attention is needed to conduct and supervise group assignments. Depending on the course and on the type of the group assignments, a variety of other online available tools and services are being used. Version control systems are being used for programming based group assignments. On the other side, Lucidchart [10] proved to be a good choice to support and share the work on the software design.

In addition, an assistant at the remote site, familiar with the hardware and software setup, is assigned to troubleshoot problems and provide local help for any problems that occur during the class. Last but not least, to provide an office hours experience, Google Hangouts or Skype are being used.

For an overview of all the challenges we have faced, as well as choices we have made to deal with them, please refer to Table I. While addressing those challenges, prior hands-on experience with distance learning and web-based collaboration tools provided an interesting insight and proved to be helpful in shaping our decisions [7][11][12].

IV. STUDENTS PERCEPTIONS OF REAL-TIME INTERACTIVE DISTANCE EDUCATION

Feedback surveys in higher education have become common in countries all over the word. These surveys are often offered not only to students enrolled in a face-to-face course, but also to instructors and students engaged in an online distant course.

The purpose of our survey was to explore how students perceive distance learning courses that are conducted online, in real-time. To explore different aspects of student's perception, the survey was designed to cover 5 categories, each explained in more details in the following subsections. The sample population consisted of 84 students enrolled in courses listed in Section II. The answers are measured using the five-point Likert scale. To simplify the visual representation of the results, the tables that follow show the percentage distribution of student responses, combining the 'Strongly Agree' and 'Agree' responses into an 'Agree' category, and combining the 'Disagree' and 'Strongly Disagree' responses into a 'Disagree' category.

A. What Type of Students Enroll in Distance Learning?

One key purpose of this survey was to determine students' profiles as distance learning courses require a lot of self-discipline and self-motivation to be successful. The student should come to classes prepared, with completing readings and other assignments. If the student has any questions or issues during a class session, he or she should feel comfortable asking for help. Participation in class discussions and collaboration on group projects when required are also welcoming characteristics.

Table II lists statements students were asked to agree or disagree with, and shows the percentage distribution of their responses. The majority of the students responded that they come to classes prepared, but what is interesting to note is that around 50% of the students do not feel comfortable participating in class discussions or asking for help, either during the class session or afterwards, through the usage of internet services such as e-mail, videoconference and chat. Even more, those students are not comfortable working with and communicating with people they have never seen face-to-face.

Although students have been made aware of the factors that influence successful learning, this opens up new questions about how to change their perception and make them feel comfortable outside their comfort zone.

B. How Does Distance Learning Impact Student Engagement?

In comparison to the traditional face-to-face instructions, students were asked how distance learning affects different areas of their engagements as presented in Table III. Our preliminary results show that, in general, more than 60% of the students reported decrease in class engagement, motivation and the quality of their learning

TABLE II. STUDENT PROFILE

Q#	Statement	Agree	Neutral	Disagree
Q1	I come to class prepared, with completing readings and other assignments.	80.95%	16.67%	2.38%
Q2	I check ahead that the technology I use in the class is working, up and running.	66.67%	19.05%	14.29%
Q3	During the lecture, I speak up if I have any questions or issues.	53.57%	28.57%	17.86%
Q4	I am a self-motivated learner.	73.81%	22.62%	3.57%
Q5	I like participating in class discussions.	45.24%	29.76%	25.00%
Q6	I like collaborating with my classmates.	71.43%	17.86%	10.71%
Q7	I feel comfortable asking for help through usage of internet services (e.g. e-mail, videoconference, chat …)	54.76%	27.38%	17.86%
Q8	I am comfortable working with and communicating with people I have never seen face-to-face.	45.24%	27.38%	27.38%

TABLE III. STUDET ENGAGEMENT

Q#	Statement	Decreased	No Change	Increased
In comparison to the traditional face-to-face instruction, in the remote course …				
Q9	… the quality of my learning activities …	75.00%	19.05%	5.95%
Q10	… my motivation to participate in class activities …	69.05%	27.38%	3.57%
Q11	… the amount of interaction with classmates …	40.48%	30.95%	28.57%
Q12	… the quantity / frequency of interaction with the instructor …	69.05%	26.19%	4.76%
Q13	… the quality of interaction with the instructor …	67.86%	27.38%	4.76%
Q14	… overall, my engagement …	65.48%	28.57%	5.95%
		Agree	**Neutral**	**Disagree**
Q15	… I have less sense of instructor control – it is easier for me to ignore the instructor.	78.57%	13.10%	8.33%
Q16	… I feel the distant students less as my own classmates.	86.90%	4.76%	8.33%

activities. This is important because it shows how crucial self-discipline and self-motivation are in a distance learning environment as discussed in section A. Even more, statement Q15 ('I have less sense of instructor control – it is easier for me to ignore the instructor.') shows that 78.57% of the students may still require a conventional, strict discipline in class.

On the other side, 59.52% of the students reported that the distance learning had no negative impact on the amount of interaction with their classmates. Even more, 28.57% of them reported an increase.

C. How Do Students Percieve Their Distant Classmates?

Courses with combined distant and local sections give an opportunity to meet people from different cultural backgrounds. We were interested in how this aspect of distance learning may have an effect on the student social network. As reported in Table IV, 22.62% of the students already made friends with students from other course sections, whereby 19.05% of the students manage to stay in touch. As expected, many of the students, 41.67%, would like to meet their distant peers in person.

D. How Helpful is Usage of Web-Based Systems & Tools in Distance Learning?

Table V, shows reports on how helpful students find the usage of different web-based systems and tools that are used in distance learning. 64.29% of the students stated that the usage of any of the provided web-based tools fits with their learning style, whereby 25.00% of them remained neutral.

More than 80% of the students find the usage of the MyCourses system useful, especially for accessing course materials, posting assignments and following up on instructor's feedback.

In cases were a version control system were applicable, around 20% of the students disagree with the statement that it makes it easy to collaborate with classmates on group projects, and to do programming group assignments. This could be due to the fact that prior the distance learning course, students have no experience in using any of the version control systems. Some students found it difficult to adapt to this new type of collaboration, especially when they are used to do the assignments individually, and all by them selves.

TABLE IV. SOCIAL NETWORKING

Q#	Statement	Agree	Neutral	Disagree
Q17	Courses with combined distant and local sections give me more opportunity to expand my social network.	17.86%	32.14%	50.00%
Q18	I would like to meet in person the students from other course sections.	41.67%	38.10%	20.24%
Q19	I stay in touch with students from other course sections.	19.05%	30.95%	50.00%
Q20	I make friends with students from other course sections.	22.62%	26.19%	51.19%

TABLE V. USAGE OF WEB-BASED SYSTEMS AND TOOLS

Q#	Statement	Agree	Neutral	Disagree	Not Applicable
Q21	Using any of the provided web-based tools fits with my learning style.	64.29%	25.00%	8.33%	2,38%
Q22	Using any of the provided web-based tools makes our class activities easier.	57.14%	27.38%	11.90%	3,57%
Q23	MyCourses modules assist me in managing and handling my course related activities.	78.57%	14.29%	4.76%	2,38%
Q24	Polycom features (e.g. tracking and auto zooming, voice and face detection) give me a sense of presence of my distant peers.	8.33%	29.76%	53.57%	8,33%
Q25	It is helpful to be able to communicate with my instructor via Google Hangouts or Skype for any help and support I need.	41.67%	21.43%	22.62%	14,29%
Q26	It is helpful to be able to collaborate with my classmates on group projects using a version control system.	32.14%	26.19%	23.81%	17,86%
Q27	Version control system makes it easy to do programming group assignments.	33.33%	26.19%	21.43%	19,05%
Q28	Online diagramming tool enables easy interactive work and sharing of diagrams among peers and instructor.	32.14%	29.76%	15.48%	22,62%
Q29	MyCourses enables easy access to course materials.	88.10%	8.33%	2.38%	1,19%
Q30	MyCourses makes it easy to post assignments and follow up on instructor's feedback.	89.29%	8.33%	2.38%	0,00%
Q31	It suits me better when I am examined via online quiz in MyCourses.	55.95%	35.71%	5.95%	2,38%
Q32	Overall, I am satisfied with online services and tools we use in this course to support and ease the process of remote lecture.	58.33%	27.38%	14.29%	0,00%
Q33	In situations when no face-to-face instructions are possible, I find the utilization of the Polycom and the Web-based tools we use a good alternative.	30.95%	25.00%	40.48%	3,57%

E. Miscellaneous

The final survey category consisted of open-ended questions on what students believe could help them and the instructors to conduct a remote lecture, and might contribute to an enhanced sense of presence.

Several students suggested how they would like to have instructor's slides and notes screened to their own computer screen instead to the projector screen. Even more, they would like to have the distant classroom and the instructor screened to their computer instead to the Polycom video display.

Their comments show preference for face-to-face lectures, but they also find Polycom a good alternative in situations were traditional lectures are not possible.

V. CONCLUSION

Delivering courses remotely, in real time, has its advantages and limitations. It works fine when the focus is on information transmission, but it presents many challenges in laboratory settings, where students are expected to apply what they've learned to new, and open-ended problems. Adding to this the necessary interactive guidance required in team-based projects, the desired environment becomes more difficult to provide.

To accommodate distance learning courses, RIT Croatia provides a special purpose classroom - a computer laboratory equipped with Polycom videoconferencing system. The usage of the Polycom system provides an excellent synchronous, two-way interactive communication between the instructor and the remote class, and is most suitable for small class environments (up to 15 students).

When it comes to distributing course materials, RIT Croatia takes advantage of the myCourses course management system that provides support for many class related activities: dropbox folders for assignment submissions, email access, grading rubrics, student-to-student and student-to-instructor chat, and attendance records.

The Apple Remote Desktop is used in courses that require in-class exercises, providing the instructor with means to supervise and assist the students in the learning process. In addition, collaboration and work on group projects have been assured through the usage of version control systems.

Providing a real time, face-to-face experience in an online learning environment does not guarantee student success in a course. In comparison to the traditional face-to-face instruction, distance learning courses require a lot of self-discipline and self-motivation to be successful.

Our preliminary survey results show that the majority of the students consider themselves responsible enough in terms of coming to classes prepared, with completing readings and other assignments. On the other hand, around 50% of the students reported not feeling comfortable participating in class discussions or asking for help, either during the class session or afterwards, which reflected negatively on their motivation and the engagement in different course activities.

The preliminary results also opened additional questions that we plane to investigate in the future. It would be interesting to discover why students report a decrease in their engagement and the quality of learning activities. In addition, in this era of social networks, why do some students report feeling discomfort while using internet services to communicate and collaborate with distant classmates and the instructor.

Student comments that were made available as part of the final survey category show preference for face-to-face lectures, but overall, they are satisfied with online services and tools we use in courses to support and ease the process of remote lectures.

REFERENCES

[1] J. Humbert and K. Vignare. (2004). *Teaching and Learning with Technology at the Rochester Institute of Technology* [Online]. Available: https://digitalarchive.rit.edu/xmlui/handle/1850/278

[2] C. Geith and K. Vignare, "Online Degree Programs: Service and Cost", *Online Education*, vol. 2

[3] The Wallace Center and the Innovative Learning Institute at RIT (2015). *myCourses Help and Resources*. [Online]. Available: https://wiki.rit.edu/display/myCoursesHR/Home

[4] Rochester Institute of Technology (2016). *RIT Online*. [Online]. Available: http://www.rit.edu/ritonline

[5] Rochester Institute of Technology (2016). *Study Abroad*. [Online]. Available: https://www.rit.edu/academicaffairs/global/global-campuses

[6] T. Volery and D. Lord, "Critical success factors in online education," *International Journal of Educational Management*, vol. 14 no 5, pp.216 - 223. 2000.

[7] K. Marasović and M. Lutz, "Remote Development and Distance Delivery of Innovative Courses: Challenges and Opportunities," in *Software Engineering (ICSE), 2015 IEEE/ACM 37th IEEE International Conference on*, vol.2, pp.299-302, 16-24 May 2015

[8] Polycom Inc. (2016). *Video Conferencing Solution*. [Online]. Available: http://www.polycom.com/content/www/en/hd-video-conferencing.html

[9] Apple Inc. (2016). *Apple Remote Desktop*. [Online]. Available: https://www.apple.com/remotedesktop

[10] Lucid Software Inc. (2015). *Diagrams Done Right*. [Online]. Available: http://www.lucidchart.com

[11] I. Bosnić, I. Čavrak, M. Žagar, R. Land, and I. Crnković, "Customers' role in teaching distributed software development," in Software Engineering Education and Training (CSEE&T), 2010 23rd IEEE Conference on. IEEE, 2010, pp. 73–802

[12] I.Bosnić, B. Mihaljević, M.Orlić, and M. Žagar, "Source Code Validation and Plagiarism Detections – Technology rich Course Experiences", in *4th International Conference of Computer Supported Education - CSEDU 2012*, pp. 149-154, 16-18 April 2012

Overview of IT solutions for career services and quality assurance at higher education

Emilija Gjorgjevska, Petar Tonkovikj and Marjan Gusev
Ss. Cyril and Methodius University, Faculty of Computer Science and Engineering, 1000 Skopje, Macedonia
{gjorgjevska.emilija, tonkovikj.petar}@students.finki.ukim.mk, marjan.gushev@finki.ukim.mk

***Abstract*—Education is a key component of society and is directly tied to its prosperity. In recent years we have seen changes in the labor market as a consequence of the development of information technology and globalization. This implied a major transformation in the structures of career and work. Therefore, the educational and careers services faced the challenge of responding to a wide range of different institutional agendas.**

In this paper we will give an overview of solutions that fall into several different categories by analyzing their advantages and the benefits they provide, but also state their shortcomings and highlight the areas of possible improvement. We would like to test the hypothesis that these kinds of solutions are mutually beneficial for students, higher education institutions and employers. However, we will also show that there is still a lot of space for improvements and innovation. The research shows that these solutions offer value to all the parties concerned: the students get higher quality education and training, the companies get high quality staff and the academic institutions get more tools to assess their quality, keep track of the student satisfaction with their work and discover new ways to improve.

***Index Terms*—Education, Faculty, Students, Evaluation, Information**

I. Introduction

With the rapid development of information technology, finding information has never been easier. We have witnessed the development of online knowledge bases, e-communities, forums and many other means of getting informed. The goal of these new tools is clear: easy and efficient access to information that is centralized and can be found on one place. Even though this technological revolution is spreading its roots into different spheres of our lives, there are still sectors that have not yet fully experienced the potential of IT. One such sector is education, where there are endless possibilities of sharing information and knowledge.

Prospective students are still having a hard time finding useful information that would help them choose their future faculty and department and undergraduate students struggle when they need to decide which course to pick or which professor to choose. In the end, they resort to the opinions of other more experienced people (older colleagues, friends, family, people on forums and social networks etc.). This indicates that even though the information is indeed somewhere out there, it is scattered on various locations and requires a lot of effort to find.

That is why we aim to design a system that will offer complex and structured information about career services and quality assurance at higher education. This system is intended to be an e-community consisted of students, organizations, instructors and companies, including student organizations, foundations and NGOs useful for students, such as Girls in Tech. It will offer exchange of student opinions regarding departments, courses and instructors in the form of ratings and in the same time it will calculate the weight and relevance of the opinion by calculating the authors competence to rate a certain entity. More experienced students will give ratings about these topics. Also, all students will have the opportunity to express their opinion about courses they have taken, instructors they were taught by, as well as their department as a whole. In the same time, the system will allow users to give suggestions for improvements. Inappropriate and abusive input will be discarded. The system will display internship, volunteering, job offers and company descriptions, as well as opportunities for traveling. In addition it will inform about workshops, academies, competitions, internships abroad etc. offered by the student organizations. This system will centralize all this information on one place, which will save the users time they would otherwise spend searching for this information on various places.

To conclude with this definition, the main goal of this system called CHub is to improve the quality of education provided by the faculty, as well as help produce graduates as quality professionals in their field. This can be achieved by conducting high quality evaluations that will improve the overall educational process, by offering different possibilities for informal education, as well as offer the opportunity to follow market trends.

The main hypothesis that we will try to prove in this document is that these possibilities are not utilized as much as they should be and that there is still room for major improvements in the education sector.

In this paper we give and overview of modern trends and current IT solutions for analyzed issues and identified problems. It is an inspiration for us to design a concept of a system that would centralize information on one location about all these issues, as well as many others.

The rest of the paper has the following structure. Related work is presented in section II along with several case studies that showcase the topic at hand. Section III explains what kind of methods will be used to analyze relevant information and Section IV represents the evaluation. Section V presents a comparison with similar systems, along with advantages, benefits and disadvantages. The results of the comparative

analysis are discussed in Section VI. Finally, the conclusions and recommendation for future work are given in Section VII.

II. Related work

Through the years, various professors, psychologists and analysts discussed the possibilities for improvements in education and now this topic has become even more interesting. IT offers new opportunities for creating solutions such as building and managing smart communities, as well as automating internal processes.

Somewhere around 2009, a group of Stanford students and professors created the CourseRank system [1]. It was a web platform that provided relevant information for the students, as well as interesting tools they can use to plan and recommend courses, track requirements etc. In the same time it was used as a tool which gave feedback to the university. The system was eventually implemented on 20 colleges across the U.S., including Berkeley, Stanford, Princeton, Cornell, Duke and many others. The system reached a point when 85% of Stanford students were registered in it [1]. In 2010, the system was acquired by Chegg [2], after which it stopped functioning.

The Gates Foundation funded the Measures of Effective Teaching (MET) project [3], a three-year study which included dozens of independent research teams and nearly 3,000 teacher volunteers from seven U.S. public school districts. The results proved that there is room for improvements and that it is possible to develop reliable measures that identify great teaching. The report describes the trade-offs involved when school systems combine different measures (student achievement gains, classroom observations, and student surveys) and provides guidance on the best ways to achieve reliable classroom observations.

To sum up, there is still a lot of work waiting to be done regarding this issue. However, it is obvious that things will develop in such a way in the future. Our focus is to contribute to this development in the educational sector.

III. Methodology

We applied the qualitative research methodology using observations in order to get relevant and up to date information for our research.

The following key indicators were identified to analyze the overview of functionalities that consist the CHub system:

- Quality control
- User reviews and ratings
- University seeker tools
- Job seeker tools
- Flexible recommendations
- Segmenting data by sections
- Additional useful information
- Encouraging user activity

A comparative analysis based on these key indicators was performed for different existing approaches.

Several case studies were analyzed for systems trying to solve the problem at hand, searched for scientifically proven facts regarding the behavior of users and looked for the identified key factors. A lot of relevant companies and universities that offered similar services were analyzed and compared. For this purpose we have created accounts on various systems and analyzed their features, as well as the way they work.

Even though the CHub system aims to provide all the services analyzed as key indicators, it is a fact that there are existing systems that offer some of the functionalities in a similar manner. In the next section, we will compare and match these functionalities into several groups, take a look at the existing alternative solutions, as well as explain what we think their disadvantages are and what they are lacking. It should be noted that none of the systems that we will discuss uses our approach in solving the problem. In all cases, all the relevant information for the students is dispersed and scattered across different social networks which cover different issues, instead of being centralized.

IV. Evaluation

In this section we will look for the key indicators we have identified previously in existing systems.

1) Quality control: ResearchGate [4] allows creating an account only by using the faculty institutional email address. Besides this, they additionally calculate and show the user RG score which measures the user scientific reputation, as well as the user's impact factor. The calculation is based on how other researchers interact with certain user content, how often they interact with that content, as well as who the viewers are. The score of the viewer affects the score of the author - the higher the user's score, the more the author's score will increase. Users can add publications, ask questions and add contributors. All of these parts are shown publicly. The number of reads that a certain user did, as well as the number of reads on his/hers publications are also shown separately. Users can add tags to everything they share.

2) User reviews and ratings: MyEdu [5], RateMyTeachers [6] and RateMyProfessors [7] allow users to rate professors and see the ratings of other users. These ratings include grading professors according to different parameters, as well as writing comments and reviews about their way of teaching. Another example of a system offering user ratings is InternsGoPro [8], which allows users to rate an internship experience in a certain company and leave comments about their satisfaction. This is a good way of allowing users to exchange experiences in an indirect manner. However, they moderate all posts that are submitted before they are posted to deal with inappropriate and abusive comments. This practice is also done by MyEdu and it has been recently adopted by RateMyProfessors as well.

Some of these systems additionally implemented up-voting and report buttons, so that users can say whether the given opinion is useful and report inappropriate comments.

3) University seeker tools: LinkedIn [9] has parts about prospective students, students, alumni, universities and employers. Their tool Youniversity provides insights in university rankings and contains tools like university finder, as well as

field of study explorer. Besides these features, the users are able to create a decision board where they can post their future decisions. LinkedIn uses these posts to give better recommendations and possible matches for the users. University LinkedIn profiles show information about their ranking, as well as information about their top alumni including their job position, workplace, study program, skills, residence location, as well as how a certain user is connected with them (first, second, third degree connections). Users can search the alumni by year, read recommendation from other users and search for notables (most popular users at the moment that have graduated at that university).

ResearchGate members can use the search bar in order to find the information they need. For example, if the user searches the word 'Oxford', the results will be displayed into several groups which contain the entered word in their content: researchers (names), topics, questions, publications, open reviews, jobs, institutions and departments.

4) Job seeker tools: ResearchGate has its own tool for job seeking called "Scientific Recruitment Solutions". They allow the clients to recruit employees by describing their ideal candidate and then match this description with the descriptions ResearchGate members provided about themselves. Additionally, clients can hire candidates based on their research history and statistics, as well as their skills, expertise, science discipline, career level and location. ResearchGate promotes job offers on the homepage of users, even if the they are not actively searching for a new position. Their analysis showed that it is good to target passive candidates, taking into consideration that 58% of people who applied for a job through ResearchGate were not actively looking for one.

When it comes to LinkedIn, they allow their clients to purchase different packages and the opportunity to select the candidates based on their level: executive, director, mid-senior level, associate, entry level, internship or not applicable. Users can search jobs by different, strictly defined parameters (e.g. location, company size, industry), save the jobs they applied for, create job alerts (notify users when new results match their criteria.), as well as review their past job applications in the "Applied jobs" section.

5) Flexible recommendations: The rich data stored in the system offers searching by different input parameters [10]. For instance, searching course ratings can be done according to the year of studies, department, student's GPA etc. Therefore, any recommendation work flow could be executed using the same engine. This differs from other recommendation systems, because their searching parameters are fixed (users are unable to choose searching criteria or are offered very limited choices). The result is a more personalized search, as the users can choose the search parameters according to their personal needs.

6) Segmenting data by sections: Milkground [11] defined various options in order to refine the search process. Visitors can search for jobs and internships by job type, sector, start date, recruiter, location, recruiter type, job title, distance etc. They can also search for events by keyword, date, location and event type and search for career advises by tags. Monster [12] also offers customized searching such that users can search for jobs by salary, time periods (choose when to start working) and search for popular job positions. Registered users are allowed to use additional useful features.

7) Additional useful information: Mladiinfo [13] displays information about scholarships (Bachelor, Master, PhD, Fellowship), grants, training, conferences, competitions, student events etc. on a local, national and international level. In addition it offers tips and tricks about different things, as well as useful articles for the students. Another example is Monster, which displays information about the institution giving the advertisement and its employees (e.g. it may show for a faculty that 86% of the professors have a Ph.D. and that 85% of the alumni recommend it).

8) Encouraging user activity: Some systems implemented additional tools in order to encourage user involvement with the system. For example CourseRank [1] offered various tools such as tools for checking requirements (GERS), tools for exchanging books, a forum for questions and answers, weekly course schedule planner called Scheduler, as well as a tool for planing future courses called Planner. MyEdu also offers tools such as a GPA Calculator that determines a student's course, semester and overall GPA, a tool called "Profiles" that students can use to show classmates and employers what kind of experiences they have had (projects, volunteering, internships, work experiences) through a visual representation with pictures and a tool called "Degree Timeline" that students can use to plan their future semesters and make sure that they are taking the right classes at the right time on their college journey.

The analyzed systems cover one or more key indicators, however, none of them covers all discussed topics leading to the conclusion that there is still room for improvement. A brief summary of matching the key indicators is presented in Table I.

V. Comparative analysis of existing solutions

Next, we will take a look at the solutions offered by the faculties and other proprietary solutions offered on the market by other companies and organizations. To this end, we will separately discuss the quality evaluation process, as well as the career services offered to the students.

A. Quality evaluation of Universities

Our previous paper, "Quality Evaluation in Higher Education Institutions in Macedonia, EU countries and U.S." [14], analyzes how different educational institutions in different regions perform quality evaluation. The conclusions show that even though quality evaluation reports are necessary for monitoring the faculty's progress, most of them are not done properly or face some kind of problem. Namely, most of the evaluation reports are done for the sake of following trends, not including the students in an appropriate way by forming committees where the student representatives are substantially less than the professors. Another problem is that most of

TABLE I: Matching key indicators identified

System	Quality control	User reviews and ratings	University seeker tools	Job seeker tools	Flexible recommendations	Segmenting data by sections	Additional useful information	Encouraging user activity
ResearchGate	Yes	Poor	Partialy	Partialy	No	No	No	Yes
MyEdu	No	Yes	No	Yes	No	Poor	No	Yes
RateMyProfessors	No	Yes	Yes	No	No	Yes	No	No
RateMyTeachers	No	Poor	Yes	No	No	No	No	No
Facebook	No	Poor	Yes	No	No	No	No	No
LinkedIn and Monster	No	Poor	Yes	Yes	No	Yes	No	No
Milkground	No	Poor	No	Yes	No	Yes	Yes	Yes
Graduateland	No	Poor	No	Yes	No	Yes	Yes	Yes
Mladiinfo	No	No	No	Poor	No	No	No	No

the faculties, in Europe in particular, refuse to share negative results of evaluation reports. In some cases, even if they do publish them, they do not inform anyone of the report release, or try to hide the fact that the report has been released. Furthermore, the overall results are sometimes not adequate due to low response rates, biased or inappropriate answers, resulting in professors losing confidence in the relevance of the results and having them fear the career consequences they might face as a result of such answers. An additional problem is that students do not trust faculty surveys that they are forced to take, which results with them giving more objective answers when using third party systems.

B. University Career services

Due to the fact that there is no legitimate law document on the topic of university career services, we had to research the practices of universities ourselves. In our previous paper, "University Career and Advising Services in Macedonia, EU countries and U.S." [15], we have reported a lot of issues concerning the current state of career and advisory services in educational institutions. This research concludes that there is still major room for improvements of career and advisory services for both the sophisticated and the less sophisticated universities, no matter whether it means implementing such services or improving existing ones.

C. Analysis of Macedonian Universities

None of universities in Macedonia have (professional) advising boards and a lot of them do not offer career services. Those who have career centers do not work properly, because of the following reasons:

- The website interface is not built according to the web design standards;
- There are no active updates regarding opportunities, which results in links to outdated information or broken links;
- The services are not upgraded with new features.

. The Universities in Macedonia are still far behind the practices of the developed countries regarding these issues.

D. Analysis of European and U.S. Universities

Almost all of researched universities offer advising, as well as career services. However, it is obvious that even in this case, institutions somehow fail to reach their goal regarding this topic. This can be seen from the fact that there are individual websites where students can post reviews and comments about the faculty (instructors, courses, the institution etc.), as well as search for career opportunities. These websites are popular and widely used by students everyday (posting and reading) because of the following reasons:

- Users are not identified by the faculty authorities;
- Users feel that their voice matters and can impact others;
- Users can search for more options and opportunities (e.g. companies that do not have an agreement with their faculty etc.).

E. Analysis of similar solutions

In this part we will take a look at the weaknesses of existing systems trying to solve the problems we are addressing. The majority of systems do not offer user authentication (one person may create multiple accounts), which means that the users can provide false information about themselves. As a result, the users can lie about their identity, grades and GPA to falsely boost the relevance of their opinions. What is even more worrying is the fact that the systems are open to everyone and anyone can create an account. This can be abused in many cases, for example multiple accounts can be created and used to flame random teachers on websites that offer ratings of professors without any repercussions. An additional problem is the fact that the users can rate any teacher or university (where such a feature is available), even if they have no experience with them. In most cases, systems provide no way to determine information accuracy and no mechanism for eliminating biased opinions, as well as abusive comments before they are posted. Another thing that should be noted is the fact that none of these systems offer informal education, or if they do, the user needs to search for it manually. Additional individual disadvantages are listed below.

1) ResearchGate: [4]

The system does not display internship and volunteering opportunities, nor useful information and guidelines for stu-

dents. Users cannot give ratings, however, they can upvote topics, questions etc.

2) MyEdu: [5]

The grading scale is imprecise, so there is a space for improving the feedback quality. When it comes to the connection with companies, it lacks a section for internships and volunteering and it does not display partnership agreements between companies and faculties.

3) RateMyProfessors: [7]

This system is limited to professor ratings. The connection with the companies is not provided in any way, as well as information and opportunities useful for the students.

4) RateMyTeachers: [6]

The system asks for users' name and surname, which means that it cannot guarantee anonymity. The amount of ratings is small, which leads to overall ratings often being consisted of one or two ratings (statistically not sound). When it comes to the connection with the companies, it is not provided in any way, as well as information and opportunities useful for the students.

5) Facebook: [16]

People use groups in order to get opinions from as many people as possible (paid adverts and pages are excluded). However, in this case it is hard to keep track of all the information (new comments and posts overshadow old ones quickly) and there is no provided information about source relevance. That is why groups represent a "makeshift" solution to the problem. Even though they exist somewhere out there, you need to manually search for them and group admins have the power to manipulate with the groups according to their will and personal belief (delete posts, invite / kick people etc.). When it comes to the user ratings and reviews, Facebook offers poor services, since users can only review pages by using the 5-star feature. In many cases, this grading scale can be imprecise and the opinions can be biased. On top of that, report buttons are missing as well.

6) LinkedIn and Monster: [9], [12]

LinkedIn and Monster lack a section for internships and volunteering and they do not offer a tool / option for graduating students to find a job and connect with employers that do not require experience. Students have to manually search companies and check if they have such offers. In the case of LinkedIn, this is happening because the advert process is based on the custom viewers profiles. The potential candidates are reached whenever they view the profiles of company employees by showing company open jobs positions.

7) Milkground: [11]

There is no information (ratings) about instructors, courses or departments. When it comes to the connection with companies, the system does not display partnership agreements between companies and faculties.

8) Graduateland: [17]

The system does not take the faculty of the user into consideration when searching. When in comes to the connection with the companies, the partnership agreements between companies and faculties are not shown.

9) Mladiinfo: [13]

The job offers are not targeting the students of a specific faculty (companies are unable to advertise directly to students) and the system does not display partnership agreements between companies and faculties. The fact that the information is dispersed suggests that the students should put additional effort to manually search for groups, pages or direct contact to a person in order to obtain it.

VI. Discussion

In this section we discuss the advantages, benefits and disadvantages of realizing the CHub system and comment the overall comparison results.

A. Advantages

There are many advantages that arise from our proposed system. It provides course and instructor evaluation, while filtering out biased ratings in the process and providing user anonymity. In addition, it centralizes all information that is useful to students regarding their professional development (employment opportunities and events), which is either absent in existing systems, or poorly implemented.

B. Disadvantages

One possible problem is the fact that in order to get a clear picture and a more accurate statistical analysis of the ratings, a large amount of users is required. In addition, there is the problem of maintaining the interest of users throughout the whole year.

C. Benefits

The features of our system yield benefits for all parties concerned. Universities save time and money intended for the career services and external evaluation, which adds to the quality control of the educational process. The formal and informal education, as well as other useful information is united and offered to students on one place. As a result, the faculties will produce higher quality staff prepared to fulfill the market needs. The companies will target students easier, allowing them to easily find employees. Organizations will be able to promote the opportunities they offer. All things considered, the system will improve the communication between all the above mentioned entities, as well as their business intelligence.

D. Comparison results

The results of this research yield interesting information. There are tools that students can use to get informed about different aspects of their studies. This kind of tools exist even in regions such as the U.S. and Europe, where faculties provide career centers and advising boards of their own. The very existence of these tools proves the fact that faculties are not providing satisfactory information to their students, both in terms of quality and quantity. The downside of these tools is that they focus on single problems and have their disadvantages, which proves our hypothesis.

Our research showed that most common mistakes these systems have is that they do not include the following:

- proper authentication of users,
- delivery of the information to the concerned people,
- protection against abusive comments,
- calculation of user credibility to comment certain topics,
- objectiveness in ratings,
- decentralized and dispersed information,
- features specifically meant for targeting students.

The benefits of the designed CHub system solves many of these problems the observed current systems face.

VII. Conclusion and future work

The conclusion, after thoroughly researching the topic, is that even though the existing tools prove to be useful in some regards, they are simply not good enough for today's standards. The way they display and manage information is inefficient, leading to increased time and effort to get to the desired information. All of these findings prove our hypothesis that educational facilities are not using IT to its full extent and that there is still room for major improvements.

One can conclude that there is a lack of systems that would gather all the information that students need on one place and encourage interaction between students, companies and faculties (including the instructors). This kind of system would be greatly beneficial to all concerned parties and would improve the overall educational process. This is the reason why we have proposed the CHub system.

By allowing students to exchange experiences, information and opinions between themselves, while simultaneously making sure that nobody misuses it in any way, the system will make life easier for students when they need to make informed decisions regarding their educational path. It will display information as objectively as possible by assuring user anonymity and eliminating user bias. Anonymity will also increase the motivation for students to express their opinions without fearing the consequences, which leads to a larger data set and more accurate data analysis. Instructors can use the feedback from the students to discover problematic segments in their teaching, so that they can improve their lectures and with the user credibility feature, they will not have to worry about biased negative comments having negative consequences on their career. Faculties can use this feature as external evaluation that they would otherwise have to pay for and see what students are dissatisfied with the most.

The proposed system is still in its design stage and in its early development. We will continue to work on improving the system by implementing additional functionalities that, according to our research, will prove to be useful, or to be lacking in current systems.

References

[1] G. Koutrika, B. Bercovitz, F. Kaliszan, H. Liou, and H. Garcia-Molina, "Courserank: A closed-community social system through the magnifying glass." in *3rd Int'l AAAI Conference on Weblogs and Social Media (ICWSM 2009)*. Stanford InfoLab, 2009. [Online]. Available: http://ilpubs.stanford.edu:8090/908

[2] Chegg. (2015) Web. [Online]. Available: www.chegg.com

[3] T. Kane, "Measures of effective teaching project releases final research report," *press release for the Bill and Melinda Gates Foundation, accessed online at http://www. gatesfoundation. org/media-center/press-releases/2013/01/measures-of-effective-teaching-project-releases-final-research-report*, 2013.

[4] ResearchGate. (2015) Social network. [Online]. Available: www.researchgate.net

[5] Blackboard. (2015) MyEdu Web. [Online]. Available: www.myedu.com

[6] RateMyTeachers. (2015) Web. [Online]. Available: http://www.ratemyteachers.com/

[7] RateMyProfessors. (2015) Web. [Online]. Available: www.ratemyprofessors.com

[8] InternsGoPro. (2015) Boosting Youth Employment Web. [Online]. Available: www.internsgopro.com

[9] LinkedIn. (2015) Web. [Online]. Available: www.linkedin.com

[10] G. Koutrika, B. Bercovitz, and H. Garcia-Molina, "Flexrecs: expressing and combining flexible recommendations," in *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*. ACM, 2009, pp. 745–758.

[11] Milkround Online Ltd. (2015) Web. [Online]. Available: www.milkround.com

[12] Monster Worldwide. (2015) Web. [Online]. Available: www.monster.com

[13] Mladiinfo. (2015) Web. [Online]. Available: www.mladiinfo.eu

[14] E. Gjogjevska and P. Tonkovikj, "Quality evaluation in higher education institutions Macedonia, EU countries and U.S," University Ss Cyril and Methodius, Faculty of Computer Science and Engineering, Tech. Rep. LiiT:67/2015, 2015. [Online]. Available: http://liit.finki.ukim.mk/?p=834

[15] P. Tonkovikj and E. Gjogjevska, "University career and advising services in Macedonia, EU countries and U.S," University Ss Cyril and Methodius, Faculty of Computer Science and Engineering, Tech. Rep. LiiT:61/2015, 2015. [Online]. Available: http://liit.finki.ukim.mk/?p=834

[16] Facebook. (2015) Social Network. [Online]. Available: www.facebook.com

[17] Graduateland. (2015) Web. [Online]. Available: graduateland.com

Selecting the Most Appropriate Web IDE for Learning Programming Using AHP

Igor Škorić, Boris Pein and Tihomir Orehovački
Juraj Dobrila University of Pula, Department of Information and Communication Technologies, Pula, Croatia
{igor.skoric, boris.pein, tihomir.orehovacki}@unipu.hr

Abstract – The focus in computing today has shifted from the desktop environment to the Web. Tasks that were previously performed exclusively on desktop can now be done in Web environment. Web tools have the same features as their desktop counterparts and offer additional specific benefits. These tools do not need to be installed or updated nor are restricted to individual computer or operating system yet available anytime and from any location. Work can be done in collaboration and result can be easily shared. The same applies to programming which indicates that complete process of software development can now be done without the use of traditional desktop tools. The basic tool used in software development is an IDE (Integrated Development Environment) which allows writing, translating and testing programs. It is also commonly used in a learning programming process. Replacing desktop IDE with its Web equivalent in learning programming has potential to improve that process and bring some advantages related to new environment. This paper presents the use of AHP (Analytic Hierarchy Process) method for multi-criteria decision-making related to choosing the most appropriate Web IDE for learning programming and reports experience of its use in introductory programming course.

I. INTRODUCTION

Programming is one of the fundamental skills in computer science. Mastering that skill is essential for anyone who sees his/her professional future in the field of information and communication technologies (ICT). The set forth is a reason why significant part of every computer science curriculum is dedicated to programming. Through a variety of courses, students adopt a number of concepts that enable them to master programming skills. Learning to program is very challenging [1] and failure rates for introductory programming courses are high [2]. In fact, teaching programming was recognized as one of seven great challenges in computing education [3]. To make the students as successful as possible in learning programming, many different tools have been proposed [4]. They can be classified into two main groups. The first one are mini-languages (like Guido von Robot [5]) that enable students to learn programming concepts by guiding an avatar in a micro world. The second group refers to visualization tools (such as ViSA [6] and Alice [7]) that combine multimedia elements and thus encourage students to achieve programming skills. Apart from mentioned tools, Radošević et al. [8] introduced specialized interface for learning programming named as Verificator that represents a mashup of editor, debugger, and tools (like tutor [9]) for analyzing a syntax of a source code written in C++ programming language. The findings of an empirical study [10] on the acceptance of Verificator revealed that it has been perceived by students as a useful and easy to use educational tool that significantly reduces their programming anxiety. More recently, educational games (such as Light-Bot 2.0 [11] and CodeCombat [12]) are gaining popularity in the context of learning programming [13]. This is because features of educational games meet both pragmatic (like helpfulness and feedback) and hedonic (such as playfulness) facets of quality perceived by students [14].

The process of learning programming concepts follows advances taking place in software industry and programming related courses are regularly enhanced in order to be up to date. As focus in computing today has shifted from the desktop environment to the Web, learning programming as well should follow these changes. Tasks that were previously performed exclusively on desktop are now easily done in Web environment. In that respect, the complete process of software design and development can now be carried out without the use of traditional desktop tools. Web tools offer some advantages over their desktop counterparts [15]. These tools are available anytime, from any location and can be employed on any operating system. Furthermore, Web tools offer much easier maintenance because there is no need for manual installation or updates. Due to nature of their environment, Web tools often contain additional functionalities for communication, collaborative work and content sharing, which are useful for educational environment and can improve learning effectiveness [16].

With an objective to improve the process of learning programming at university beginners' level, we initiated a study on employment of a Web tool that would encourage collaboration and sharing among students thus helping them to gain programming skills in a novel and easier fashion. Taking the aforementioned advantages of Web tools into account, the barriers related to practicing problem assignments solving by writing a source code will be minimized. This paper presents first step toward that process – the selection of a Web tool meant for peer programming that will replace Dev-C++ [17] IDE which is currently used in introductory programming course because of its simplicity and features that are sufficient for completing students' programs. Selection of a tool is a complex decision-making process that includes many different parameters. In that respect, we employed a formal multiple criteria method widely known as Analytic Hierarchy Process (AHP) [18] which helped us to make an

efficient decision on the most appropriate Web-based tool aimed for pair programming.

The remainder of the paper is organized as follows. The AHP method is briefly described in the next section. Details on employed research methodology are provided in the third section. The outcomes of the AHP use together with the findings of the empirical pilot study on the employment of selected Web tool in introductory programming course are presented in the fourth section. The last section offers an overview of major contributions, limitations, and future research directions.

II. Analytic Hierarchy Process

Analytic hierarchy process (AHP), designed by Saaty in 1970's, is a practical and effective method used for solving a wide variety of multi-criteria decision-making problems [18, 19]. AHP is based on structuring the elements of the problem into a hierarchy which consists of the goal, set of criteria, sub-criteria, and the set of alternatives. Structuring the problem in this way enables both qualitative and quantitative criteria to be compared. The aim of decision making is to determine the best alternative for reaching the goal, with respect to the criteria and their importance or preferences in the decision-making process. Saaty proposed a 4 step decision decomposition [18]:

1. Defining the problem.
2. Structuring the decision hierarchy from top to bottom, with goal on top level, criteria and sub-criteria in the middle, and list of alternatives on the lowest level of hierarchy.
3. Performing pair-wise comparisons of elements at a peer level with respect to the element from the upper level of hierarchy thus obtaining the relative priorities (weights) of one criterion or alternative over another in form of a priority vector.
4. Synthesizing obtained relative priorities through the hierarchy from top to bottom in order to obtain the overall priority vector of the alternatives.

In performing pair-wise comparison, the fundamental scale of absolute numbers, ranging from 1 to 9, is used.

Let n be the number of elements x_i on a current hierarchy level to be pair-wise compared. Decision maker must do $n(n-2)/2$ comparisons, results of which are collected into a n-dimensional square matrix $A=[a_{ij}]$, where a_{ij} expresses the degree of preference of element x_i to x_j, and the relation $a_{ji}=1/a_{ij}$ holds [20]. In Saaty's eigenvalue method, the priority vector w is calculated as the eigenvector of A associated with λ_{max}, the maximal eigenvalue of A, that is, by solving the system $Aw=\lambda_{max} w$, $\sum w_i=1$.

Synthesizing step is performed by weighting (multiplying) calculated priorities by priorities in the node immediately above, doing so for the whole hierarchy, and adding weighted priorities for the lowest level elements. This yields the overall priority vector of the alternatives.

AHP also includes checking for inconsistency in pair-wise comparisons. A numerical value, consistency index, CI is used to measure the consistency of the matrix A by using the formula $CI(A)=(\lambda_{max}-n)/(n-1)$, where n is the number of compared elements. Consistency ratio, CR is then calculated as ratio of CI and random index RI, an estimation of the average CI obtained from a set of randomly generated matrices of size n [8]. If the CR is larger than 0.1, the re-evaluation of comparisons is required.

In case of group decisions, the aggregation of overall priority vectors can be calculated using (weighted) geometric mean of individual priorities [18].

III. Research Methodology

First step in the study was the development of a hierarchy of problem in accordance with AHP method. Overall goal, criteria and attributes were arranged in different layers. The overall goal was selection of the most appropriate Web IDE whereas four main criteria were:

- minimal workload,
- usefulness of a tool,
- appropriateness of a tool for educational purposes, and
- additional features.

Web IDE was intended for use in lab-based exercises in courses related to programming. It should therefore meet needs of both students and teachers. The first two mentioned criteria are related to the acceptance of tool by users. According to Technology Acceptance Model (TAM) [21], motivation of user to employ a tool is affected by his/her attitude toward the use of this tool. The attitude is affected by the perceived ease of use of a tool and its perceived usefulness. This means that users will accept a tool that is considered simple and useful. In that respect, the first two criteria should measure the extent to which the use of proposed Web tools requires minimal workload and the degree to which they are beneficial in the context of learning programming. The role of tools that were analysed in this study is twofold. On the one hand, they have to replace a desktop IDE and become fundamental programming tool in which students will write their programs. On the other hand, they need to take advantage of the new environment (Web) and improve the quality of learning. Characteristics that can contribute to the quality of learning programming, and are naturally associated with Web environment, are collaborative work and the ability to share content. While the third criterion evaluates characteristics of tools that contribute to its implementation in educational settings, the fourth criteria is meant for measuring some additional features that are not essential but can increase applicability of a tool (e.g. access control, availability of a source code, the option of downloading and installing a tool, number of programming languages that a tool supports, etc.), and other functionalities that or are unique to particular tool. Therefore, the fourth criterion facilitates the assessment of two different tools that have similar capabilities. Each of the four main criteria is further decomposed to between three and five sub-criteria, as depicted in Figure 1.

The authors of the paper who are teachers in several programming related courses evaluated the weight of each criterion in the hierarchy. The analysis of data collected

from three domain experts was carried out by means of BPMSG AHP Online System [22].

Drawing of the outcomes of the AHP method, the framework for decision making was set up. The next step was selection of Web tools suitable for pair programming.

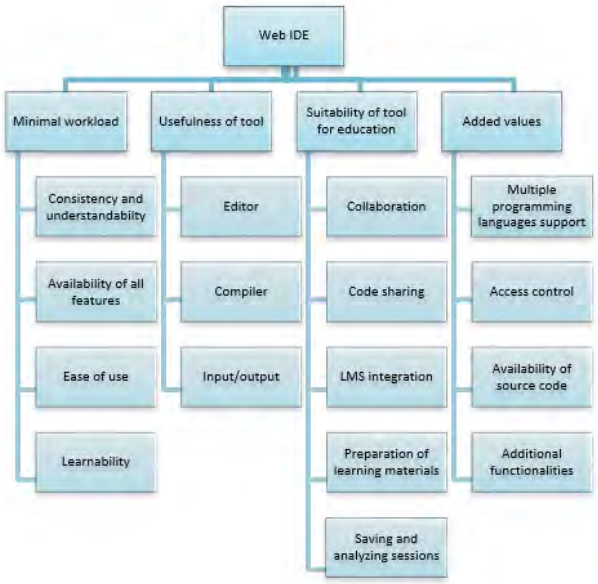


Figure 1. Criteria hierarchy

The list of possible candidates was created by searching terms 'web IDE' and 'online IDE' on Google. Initial set of candidates consisted of 35 Web tools. Such a large number of candidates was inappropriate due to a number of pair-wise comparisons. In order to reduce the list of candidates, following four elimination criteria were defined:

- Tool supports C ++ programming language
- Tool is free to use
- Tool is an open source
- Tool allows sharing of source code

Only four tools have met required criteria and were therefore selected for the process of comparing their characteristics:

- CodeBoard [23]
- Rextester [24]
- Ideone [25]
- Codepad [26]

Through the process of group decision making, by comparing properties of each pair of candidates and consolidating results, the most appropriate tool was chosen.

The last part of the study was focused on exploring the perceived quality of a selected Web tool when used for educational purposes. The selected Web tool was examined in the introductory programming course (CS1) which is designed to introduce programming fundamentals to students and help them to develop skills necessary to individually write simple programs.

At the beginning of the session, the Web tool was briefly introduced to students. The lab-based exercise was divided into three parts. In the first one, examples of programming problems and their solutions were presented to students and discussed. During the second part, partially solved problems were distributed to students who were required to write solutions and upload them to LMS Moodle. In third part of the session, assignments were administered to students who were supposed to solve them by themselves. All activities (presentation of a code, distribution of tasks, editing, compiling, testing, and uploading the solutions) were carried out with the selected Web tool.

At the end of the session, students were asked to complete a post-use questionnaire associated with their perception of the Web learning tool. The questionnaire consisted of nine items related to students' demography, 35 items related to the perceived quality scored on 4-point Likert-scale (1 – strongly agree, 4 – strongly disagree), and two open-ended items. In the context of this study, the perceived quality refers to a combination of constructs which originate from TAM model [27] (attitude toward use, perceived ease of use, perceived usefulness, continuance intention), expectation-confirmation theory [28] (confirmation of expectations and satisfaction), theory of flow [29] (internal motivation) and a constructs meant for measuring the attitude toward programming proposed by authors of this paper.

IV. Results and Discussion

Through the process group decision making, weight (importance) for each element in the hierarchy (goal, criteria, sub-criteria and alternatives) has been determined. As presented in Table 1, estimated eigenvector yielded the relative ranking of proposed criteria. It appeared that two most important criteria for selecting alternatives are usefulness and suitability because they most closely reflect the purpose of the tool. Within the criteria on usefulness of a tool (which has the highest weight among all criteria), the sub-criteria that evaluates properties of an editor has greatest importance. Consistency ratios (CRs) of matrices formed in pair-wise comparisons ranged from 0.9% to 4.3%. Taking into account threshold value proposed in [18], the reported results are acceptable. Overall consistency ratio for group results was 0.1%, while overall group consensus was 69.2%.

Process of group decision making was afterwards employed to compare alternatives, that is, four Web tools. Table 2 shows final ranking (group result) of alternatives, as well as each teacher's individual results. Two of the three teachers ranked CodeBoard (depicted in Figure 2) as their first choice whereas the third teacher ranked it as the second best. Group consensus of the reported result was 96.7%.

Second part of the study was conducted among students with an objective to explore their perception of employed Web tool. The sample was composed of 58 participants (68.97% male, 31.03% female). At the time study took place, majority of students (91.4%) had between 18 and 20 years. Most of them (89.66%) had no programming experience before enrolling the course. Even though 43.1% of students are perceiving their programming skills as bad or very bad, most of them are enjoying programming, and are willing to accept programming related job in the future.

The Cronbach's alpha [29] values ranged from .735 to .919, thus indicating sufficient reliability of scales [30].

TABLE 1. HIERARCHY WITH CONSOLIDATED PRIORITIES

Level 0	Level 1	Level 2	Global priorities
Web IDE	Minimal workload **0.179**	Consistency and understandabilty **0.1236**	2.2%
		Availability of all features **0.0977**	1.7%
		Ease of use **0.4275**	7.7%
		Learnability **0.3512**	6.3%
	Usefulness of tool **0.4127**	Editor **0.6675**	27.5%
		Compiler **0.1333**	5.5%
		Input/output **0.1992**	8.2%
	Suitability of tool for education **0.3262**	Collaboration **0.3048**	9.9%
		Code sharing **0.2499**	8.2%
		LMS integration **0.1183**	3.9%
		Preparation of learning materials **0.1836**	6%
		Saving and analyzing sessions **0.1433**	4.7%
	Added values **0.0821**	Multiple programming languages support **0.1524**	1.3%
		Access control **0.3111**	2.6%
		Availability of source code **0.4455**	3.7%
		Additional functionalities **0.0911**	0.7%

TABLE 2 INDIVIDUAL RESULTS

Participants	Ideone	Rextester	CodeBoard	Codepad
Group result	23.4%	27.7%	33.9%	15%
Participant 1	12.8%	26.8%	41.6%	18.9%
Participant 2	27.5%	28.9%	32.4%	11.2%
Participant 3	32.5%	24.8%	27%	15.7%

As shown in Table 3 of Appendix, students rated CodeBoard as simple and useful tool that has met their expectations in the context of learning programming. Regarding the data collected with two open-ended items, students commended its simplicity and the code sharing feature, especially in interaction with the teacher. On the other hand, students criticized low visibility of compiler's error messages and slow compiling of code.

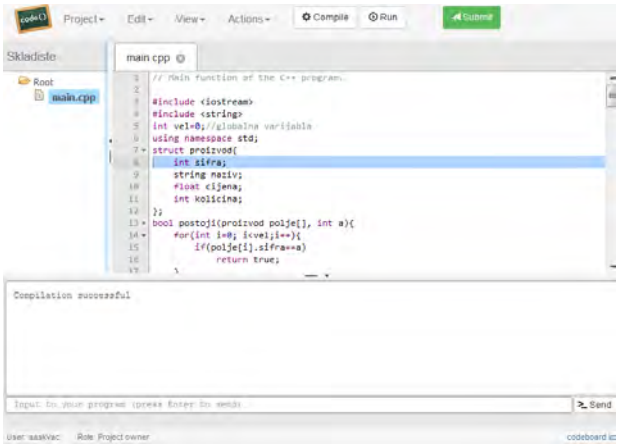


Figure 2. CodeBoard interface

V. CONLUSION

The objective of this paper was two-fold. First, we wanted to determine which criteria are the most important when selecting a Web IDE for pair programming. The analysis of data carried out by means of the AHP method revealed that usefulness of a tool was perceived as the most relevant in that respect, followed by suitability of tool for educational purposes, its effortlessness of use, and eventually additional features a tool provides. Among four tools that were included in the pair-wise comparisons, CodeBoard has met the aforementioned criteria to the greatest extent. Second, we wanted to examine perceived quality of selected tool when employed in real settings. The analysis of data collected from students by means of the post-use questionnaire uncovered that they have mixed feelings about programming. Nevertheless, students have positive attitude toward the use of Web IDE and are willing to employ it for the purpose of learning programming, especially since CodeBoard appeared to be useful, playful, and ease for use tool that has met their expectations.

As with all other empirical studies, this one also has limitations. The first one is related to the small and homogenous sample of experts involved in determining the relevance of criteria. In that respect, the robustness of AHP findings should be tested on a larger sample composed of teachers from different educational institutions as well as programming professionals. The second limitation deals with the homogenous sample of students that evaluated quality of selected tool. Although involved students are representative sample of Web IDE users, more heterogeneous sample of students in terms of study they enrolled (information sciences, information and communication technologies, computing, etc.) could provide different answers. Therefore, the reported findings have to be interpreted with caution.

Results of the work presented in this paper are encouraging because they support the decision to use a Web tool for the purpose of teaching programming. Reactions of students were positive and indicate that the tool has been well accepted. However, to draw some sound conclusions additional studies on measuring its effectiveness in educational process have to be carried out. Accordingly, our future work will be focused on exploring the interplay of quality dimensions which constituted the post-use questionnaire.

REFERENCES

[1] B. Simon, R. Lister, and S. Fincher, "Multi-Institutional Computer Science Education Research: A Review of Recent Studies of Novice Understanding," In Proceedings of the 36th Annual Frontiers in Education Conference, FIE '06, pages 12–17. IEEE, 2006.

[2] J. Bennedsen and Caspersen, M.E., "Failure Rates in Introductory Programming", SIGCSE Bull, vol. 39, 2, 2007.

[3] A. McGettrick, R. Boyle, R. Ibbett, J. Lloyd, L. Lovegrove, and K. Mander, "Grand challenges in computing education – A summary," The Computer Journal, 48(1), pp. 42-48, 2005.

[4] D. Radošević, T. Orehovački, and A. Lovrenčić, "New Approaches and Tools in Teaching Programming", In Proceedings of the 20th Central European Conference on Information and Intelligent Systems (CECIIS), Varaždin: FOI, 2009. pp. 49-57.

[5] Guido von Robot, http://gvr.sourceforge.net, Accessed: February 18th, 2016.

[6] I. Reif, and T. Orehovački, "ViSA: Visualization of Sorting Algorithms", In Proceedings of the 35th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija: IEEE, 2012. pp. 1146-1151.

[7] M. Conway, S. Audia, T. Burnette, D. Cosgrove, and K. Christiansen, "Alice: lessons learned from building a 3D system for novices", In Proceedings of the SIGCHI conference on Human factors in computing systems, The Hague: ACM, 2000, pp. 486 – 493.

[8] D. Radošević, T. Orehovački, and A. Lovrenčić, "Verificator: Educational Tool for Learning Programming", Informatics in Education, vol. 8, no. 2, 2009, pp. 261-280.

[9] D. Radošević and T. Orehovački, "An Analysis of Novice Compilation Behavior using Verificator", In Proceedings of the 33rd International Conference on Information Technology Interfaces (ITI), Cavtat: IEEE, 2011. pp. 325–330.

[10] T. Orehovački, D. Radošević, and M. Konecki, "Acceptance of Verificator by Information Science Students", In Proceedings of the 34th International Conference on Information Technology Interfaces (ITI), Cavtat: IEEE, 2012. pp. 223-230.

[11] Light-Bot 2.0, http://armorgames.com/play/6061/light-bot-20, Accessed: February 18th, 2016.

[12] CodeCombat, https://codecombat.com, Accessed: February 18th, 2016.

[13] T. Orehovački and S. Babić, "Inspecting Quality of Games Designed for Learning Programming", In P. Zaphiris, and A. Ioannou (eds.) Learning and Collaboration Technologies, Lecture Notes in Computer Science, vol. 9192, 2015, pp. 620-631.

[14] T. Orehovački, and S. Babić, "Evaluating the Quality of Games Designed for Learning Programming by Students with Different Educational Background: An Empirical Study", In Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija: IEEE, 2015, pp. 963-968.

[15] D. G. Synodinos, "Web-based IDEs to become mainstream?," InfoQ.com, 2009, http://www.infoq.com/news/2009/02/web-based-ide, Accessed: 10th January 2016.

[16] S. Cho and C. Liu, "Learning effectiveness in web-based technology-mediated virtual learning environment," 38th Hawaii Int. Conf. on System Sciences, 2005.

[17] Dev-C++, https://sourceforge.net/projects/orwelldevcpp, Accessed: February 18th, 2016.

[18] T. L. Saaty, "Decision making with the analytic hierarchy process", Int. J. Services Sciences, vol. 1, no. 1, 2008, pp. 83-98.

[19] O. S. Vaidya, S. Kumar, "Analytic hierarchy process: An overview of applications," European Journal of Operational Research, vol. 169, no. 1, 2006, pp. 1-29.

[20] M. Brunelli, "Introduction to the Analytic Hierarchy Process", Springer Briefs in Operations Research, Springer, 2015.

[21] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology", MIS Quarterly, vol. 13, no. 3, 1989, pp. 319-340.

[22] BPMSG AHP Online System, http://bpmsg.com/academic/ahp.php, Accessed: February 18th, 2016.

[23] Codeboard, https://codeboard.io, Accessed: February 18th, 2016.

[24] Rextester, http://rextester.com, Accessed: February 18th, 2016.

[25] Ideone, https://ideone.com, Accessed: February 18th, 2016.

[26] Codepad, http://codepad.org, Accessed: February 18th, 2016.

[27] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology", MIS Quarterly, vol. 13, no. 3, 1989, pp. 319-340.

[28] R. L. Oliver, "A cognitive model for the antecedents and consequences of satisfaction", Journal of Marketing Research, vol. 17, no. 4, 1980, pp. 460–469.

[29] M. Csikszentmihalyi, "Beyond Boredom and Anxiety", Jossey-Bass, San Francisco, 1975.

[30] D. Straub, M. Boudreau, D. Gefen, "Validation Guidelines for IS Positivist Research", Communications of the Association for Information Systems, vol. 13, no. 1, 2004, pp. 380-427.

APPENDIX

TABLE 3. SUMMARY OF MEASUREMENT SCALES

Items	**Mean**	**SD**
Internal Motivation (Cronbach's α = 0.873)		
Learning programming with this tool is interesting.	1.84	0.616
It is nice to work with this tool.	1.84	0.616
Using this tool is fun.	1.97	0.591
I feel good while using this tool.	2.12	0.727
Attitude Toward Use (Cronbach's α = 0.919)		
Using this tool for learning programming is a good idea.	1.66	0.664
Using this tool for learning programming is a wise idea.	1.69	0.754
This type of tool should be used in programming-related courses.	1.64	0.667
I have a positive attitude on using this type of tool in programming related courses.	1.71	0.676

TABLE 3. SUMMARY OF MEASUREMENT SCALES (CONTINUED)

Items	Mean	SD
Perceived Usefulness (Cronbach's α = 0.882)		
This tool has all the features necessary for writing programs.	2.03	0.725
This tool encourages me to actively participate in lab-based exercises.	2.28	0.812
This tool is useful for learning programming.	1.67	0.632
I learn better with this tool.	2.19	0.805
This tool is good for learning programming.	1.88	0.751
I learn more when I use this tool.	2.36	0.873
I try harder when I use this tool.	2.41	0.817
I understand this tool's compiler's error messages.	2.09	0.884
This tool's editor offers enough features.	2.09	0.571
This tool makes learning programming easier.	1.95	0.736
Perceived Ease of Use (Cronbach's α = 0.847)		
It is easy for me to become proficient in using this tool.	1.71	0.622
This tool is simple to use.	1.67	0.604
Interacting with this tool does not require a lot of effort.	1.81	0.712
This tool has a simple interface.	1.64	0.667
Continuance Intention (Cronbach's α = 0.735)		
I intend to continue to use this tool.	2.24	0.709
I would recommend this tool to programming novices.	2.10	0.788
I intend to use this tool often.	2.28	0.696
Attitude Toward Programming (Cronbach's α = 0.832)		
I enjoy programming.	2.10	0.765
I think I am good at programming.	2.59	0.795
I believe I will pass this course easily.	2.76	0.844
I am ready to accept a job that requires programming skills.	1.93	0.645
Confirmation of Expectations (Cronbach's α = 0.843)		
The experience of using this tool was better than I expected.	1.95	0.544
This tool's features are better than I expected.	2.12	0.623
This tool met my expectations.	2.07	0.525
Satisfaction (Cronbach's α = 0.910)		
I am satisfied with this tool's features.	1.93	0.525
I am pleased with the use of this tool.	1.90	0.519
I am impressed with this tool.	1.95	0.605

Using Real Projects as Motivators in Programming Education

Mario Konecki, Sandra Lovrenčić and Matija Kaniški
Faculty of Organization and Informatics, Varaždin, Croatia
mario.konecki@foi.hr; sandra.lovrencic@foi.hr; matija.kaniski@foi.hr

Abstract - It is very hard to properly motivate students to learn abstract programming concepts and programming is generally considered as hard. Students tend to fear programming courses and they develop skeptic attitude towards this kind of courses. One of the key factors that contribute to this kind of attitude is abstract nature of programming and the fact that students do not realize why exactly they should gain such knowledge. In this paper an approach that incorporates using real projects to show students the practical outcomes of their learning process in order to motivate them to learn programming is discussed and initial results of its usage are presented and elaborated.

Keywords **- programming, courses, students, motivation, real projects**

I. INTRODUCTION

Problems in programming courses, especially introductory ones, are well known and despite of many different attempts and efforts to overcome these problems, they are still persistent and they affect both students and their teachers. As long as there is no clear solution to the problems of programming courses, new attempts and ideas are needed in order to address the problems in programming education, which include many factors that can contribute either to the problem itself or to its solution.

Programming is abstract for students. It presents them with matter which is all but intuitive to them and which is something that they have not dealt with before [1, 2]. It can be said that the general attitude of teachers and students is that programming is hard to teach and learn. The fact that failure rates in introductory programming courses are relatively high is another proof that the mentioned attitude is not without its reasons [3]. Since most of students have no prior experience with programming, they start to develop fear and skepticism towards programming. This fear is also caused by many skeptic comments which are forwarded to students by their peers who have taken these courses. In this way students' motivation to learn decreases, which makes teachers' task even harder. Another problem is a variety of students' profiles, which makes it almost impossible to meet all students' requirements.

Although there are existing tools and methods that are aimed at helping students in their first programming steps, their usage greatly depends on willingness of the teachers to adopt such means, on their institutions' policies and on the support that is or is not provided for these kind of tools.

Problems in programming education require new joint efforts in order to create a framework that can serve as a basis for programming education and as such provide a viable and usable solution for teachers of programming courses and for their students. In this paper, an approach that includes using real projects as motivators in programming education is presented, elaborated and discussed as a potentially usable solution for programming novices. Presented approach is researched as one of many other possible approaches that can be used alone or in combination with other approaches to help students in being more motivated and to achieve better results and greater quality of their programming knowledge and skills.

II. CHALLENGES OF PROGRAMMING EDUCATION

There are many challenges that are present when talking about programming education and these challenges are influencing both students and teachers. Many students experience problems in dealing with the most basic programming concepts and elements, such as control structures, and they are unable to use these elements to create a proper solution for the given problem. Many possible reasons for this kind of state can be given [4; 5; 6]:

- Programming demands a high abstraction level.
- Programming needs a good level of both knowledge and practical problem solving techniques.
- Programming requires a very practical and intensive study, which is quite different from what is required in many other courses (more based in theoretical knowledge, implying extensive reading and some memorization).
- Usually teaching cannot be individualized, due to common classes' size.
- Programming is mostly dynamic, but usually taught using static materials.
- Teachers' methodologies many times don't take into consideration the student's learning styles. Different students have different learning styles and can have several preferences in the way they learn.

- Programming languages have a very complex syntax with characteristics defined for professional use and not with pedagogical motivations in mind.

When talking about programming education, several aspects should be considered [7]:

- Curricula
- Pedagogy
- Language choice
- Tools for teaching

Curricula defines elements that are to be taught and pedagogy considers the methods and means that are to be used in order to teach students in a proper and suitable manner that will be understandable and that will result in desired level of students' knowledge and skill. When talking about the programming language choice for usage in introductory programming courses C, Java and C++ are the top three languages that can be commonly found [7]. However, there has been much debate about the suitability of these languages [8; 9; 10; 11] since the fact is that they have not been designed with education in mind, compared to languages such as Logo, Pascal, Eiffel and Python. Although many attempts at developing various tools that would aid both students and teachers have been made, most of these tools have not been used on a larger scale [7]. There are only some tools that have achieved wider usage, for example BlueJ [12]. When considering various tools designed for usage in programming education, they can be divided into several groups [7]:

- Visualization tools
- Automated assessment tools
- Programming environments
    - Programming support tools
    - Microworlds
- Other tools

Visualization tools enable students to visually perceive various abstract programming concepts and structures as well as to see the dynamics of code execution. There are many tools that can be found which provide such features [13; 14; 15; 16; 17].

Automated assessment tools enable teachers to automate assessment and students are in this way provided with instant results of their work. There are some tools that can be found which provide automated assessment and most of them are based on checking of proper code execution and on checking for proper execution results [18; 19].

Programming environments can be divided into programming support tools and microworlds. Programming support tools are designed to provide programming novices with a subset of full programming environments so that it would be easier for them to cope with their tasks by reducing the overall tool complexity. Such tools are, for example, The NetBeans IDE BlueJ Plugin [20] or GILD [21].

Specific microworlds, such as Alice, Greenfoot and Scratch [22], are based on their own worlds and metaphors and in many cases on their own programming languages. The aim of these tools is to make the first steps in programming more intuitive for programming novices.

Some other tools also exist, such as the tools for plagiarism detection, which include JPlag, MOSS and YAP [23]. There are authors who report about beneficial effects of pair-programming [24], which promotes motivation and increases the time students are spending in learning of programming by doing their programming tasks. Some authors also argue that the problem of introductory programming courses is not only that students are not able to program at the end of their programming courses but that this problem has three components [25]:

- All students should be able to program
- Weaker students cannot program
- Weaker students still pass the introductory courses because too low passing rates are not acceptable by teachers' institutions

To address this three-component problem, one of possible approaches would be to create a clear level-based and goal-oriented teaching strategy in which the difference between the lowest passing grade and the highest grade would not be vague or very small, but in which every grade would have its own requirements, which are appropriate for that very grade. In this way, weaker students would be able to focus on just a subset of the course and they would be able to pass the course, while better students would have a chance to show greater knowledge and skills and in such a way get a higher grade, which would clearly state their level of knowledge after the introductory programming course.

In order to help students in understanding programming, an approach that includes using mini-languages has been proposed [26]. Mini-languages present a subset of full programming languages, designed with simplicity in mind. This kind of approach offers students much less overwhelming mean of learning programming compared to traditional full language approach, which has several disadvantages [26]:

- General-purpose languages are too big and too idiosyncratic
- General-purpose languages provide little leverage for understanding their basic actions and control structures
- Since general-purpose programming languages are orientated on number and symbol processing, the first possible problems used in teaching the language are far from the students' everyday experiences and are not attractive for them

In order to understand programming, students have to adopt two components [27]:

- Factual Knowledge
- Programming skills

Both of these require specific teaching and learning strategies. The part that students have the most problems with is adoption of programming skills since their past education has been mostly based on memorization of various facts. To learn a skill students are required to learn in a prolonged manner, which they rarely do.

III. REAL PROJECTS AS MOTIVATORS

One of the more significant problems that teachers of programming courses are struggling with is the lack of students' motivation to deal with programming. Without proper motivation it is very unlikely that students will spend much time learning programming, so addressing the issue of motivation is important step towards solving the problems in programming education.

The idea of inclusion of projects into education of programming can be found in scientific literature. The approach that is universally used for various programming paradigms, and on different study levels, is to give students a practical and possibly interesting project application to develop as an assignment during their programming course [28]. During the project students can have assistance from other faculties for whose domain the project is developed [29], they can choose the topic themselves or the topic can be given to them by their teachers [30]. A development of Internet application as a motivator can be included into project assignment [31] and, if needed, smaller parts of real projects can be used and projects can be developed in teams [32]. All those proposals are aimed at improving programming skills and grading results at the end of the semester.

In order to increase students' motivation to deal with programming, a teaching methodology which includes usage of real projects as motivators, but based on a different approach, has been developed and implemented. The proposed teaching methodology was used during one semester and 52 students have participated in the research. The proposed methodology consists of two motivational aspects that have been used during the programming course. Firstly, to increase students' motivation to deal with programming four demonstrations of real projects that are in practical use were introduced during the semester. The aim of these demonstrations was to show students that all abstract programming concepts have in the end very perceivable and practical outcome and that their education will enable them to do such projects as a part of their future professional work, although they are not able to produce such results at the moment.

The projects were shown from the user's point of view, but the overview of used technology and program code was also given. Two presented projects were desktop applications and the other two were web applications where the design aspect was also more obvious and more interesting for students. Secondly, as a particular lesson was introduced to students (e.g. Selections), an appropriate example from the real project that includes the concept/construct that is being taught (e.g. selection) was presented. In this way students learn while keeping in mind the practical and perceivable use of the concept/construct, which increases their motivation to learn the particular concept/construct.

Students were also able to ask questions about presented projects and they were also asked during and after the lecture to try to connect presented concepts from the lesson with the project and to conclude where these concepts can be used in practice. In this way students have been trained in perceiving the real use of abstract programming concepts/structures, which has been aimed at increasing their motivation to learn programming since it in this way stopped being so abstract and started to be perceived as a useful and practical tool which can be utilized in real projects on real business systems. Although this approach is aimed at university students who can comprehend the practical value of presented projects, the same approach could also be used in high schools although to conclude about its effectiveness in this case additional research should be conducted.

In order to conclude about the effectiveness of proposed teaching methodology on students' motivation, a research has been conducted among 52 students at the end of their programming course. Students were given five questions based on a Likert scale (5 - I strongly agree, 1 - I strongly disagree). Obtained research results are shown in Table I.

TABLE I. RESULTS OF CONDUCTED RESEARCH

Questionnaire item	Mean	Std. dev.
Presentations of real projects have been very interesting to me.	4,49	0,21
Presentations of real projects have made me aware of practical use of programming.	4,27	0,48
Presentations of real projects have made me able to understand abstract programming concepts better.	4,09	0,73
Presentations of real projects have made me more motivated to learn programming.	4,24	0,41
I would like to have this concept of teaching as a part of my other courses.	4,35	0,64

The results of conducted research have shown that proposed teaching methodology can be used as a valid motivational mean and that students have been motivated by its usage. Students' attention has been captured in greater amount compared to traditional approach and they were able to see practical results of presented lessons which made them more willing to spend time in learning about programming.

IV. CONCLUSION

Programming courses have experienced various problems in trying to teach students programming for years and this trend has not yet diminished. Students tend to fear programming and they perceive it as something abstract and hard. Teachers agree that programming is hard to learn and they use various aiding means in a greater or lesser capacity. However, the problems of programming courses have remained present. As a result, students are frequently unmotivated and unwilling to deal with programming and to spend time in learning.

In order to try to give a usable solution that can be used alone or in combination with other tools and

approaches, a new teaching methodology has been proposed, which uses real projects to motivate students to deal with programming. By introducing real projects into programming education, students are made aware of practical use of programming which in the end increases their motivation to learn how to program.

Research results have confirmed usability of proposed teaching methodology and its further testing and further development will be a part of future research. After further elaboration and implementation for all enrolled students, validation through comparison of students' results for specific assignments with results from previous years should give general guidelines for inclusion of the methodology into teaching process of programming courses.

REFERENCES

[1] M. Konecki, "Problems in programming education and means of their improvement", DAAAM International Scientific Book, pp. 459-470, 2014.

[2] M. Konecki, and V. Mrkela, "Algorithmic thinking and animated interactive presentation of sorting algorithms in education of students", In Proceedings of the 17th International Multiconference Information Society-Education in Information Society, pp. 105-112, 2014.

[3] J. Bennedsen, and M. E. Caspersen, "Failure rates in introductory programming", ACM SIGCSE Bulletin, vol. 39, no. 2, 2007, pp. 32-36.

[4] E. Soloway, and J. Spohrer, "Studying the Novice Programmer", Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1989.

[5] T. Jenkins, "On the difficulty of learning to program", In Proceedings of the 3rd Annual LTSN-ICS Conference, The Higher Education Academy, Loughborough University, United Kingdom, pp. 53-58, 2002.

[6] E. Lahtinen, K. Ala-Mutka, and H-M. Järvinen, "A study of difficulties of novice programmers", In Proceedings of the 10th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education, Monte de Caparica, Portugal, pp. 14-18, 2005.

[7] A. Pears, S. Seidman, L. Malmi, L. Mannila, E. Adams, J. Bennedsen, M. Devlin, and J. Paterson, "A survey of literature on the teaching of introductory programming", ACM SIGCSE Bulletin, vol. 39, no. 4, 2007, pp. 204-223.

[8] R. Biddle and E. Tempero, "Java pitfalls for beginners", SIGCSE Bulletin, vol. 30, no. 2, 1998, pp. 48-52.

[9] R. Close, D. Kopec, and J. Aman, "CS1: perspectives on programming languages and the breadth-first approach", In Proceedings of the 5th annual CCSC Northeastern Conference on Computing in Small Colleges, Consortium for Computing Sciences in Colleges, pp. 228-234, 2000.

[10] S. Hadjerrouit, "Java as first programming language: a critical evaluation", SIGCSE Bulletin, vol. 30, no. 2, 1998, pp. 43-47.

[11] R. P. Mody, "C in education and software engineering", SIGCSE Bulletin, vol. 23, no. 3, 1991, pp. 45-56.

[12] M. Kölling, B. Quig, A. Patterson, and J. Rosenberg, "The BlueJ system and its pedagogy", Computer Science Education, vol. 13, no. 4, 2003, pp. 249-268.

[13] V. Karavirta, A. Korhonen, L. Malmi, and K. St?lnacke, "MatrixPro - A tool for on-the-fly demonstration of data structures and algorithms", In Proceedings of the 3rd Program Visualization Workshop, The University of Warwick, UK, pp. 26-33, 2004.

[14] J. Jain, J. H. Cross II, T. D. Hendrix, and L. A. Barowski, "Experimental evaluation of animated-verifying object viewers for Java". In Proceedings of the 2006 ACM Symposium on Software Visualization, ACM Press, pp. 27-36, 2006.

[15] A. Moreno, N. Myller, E. Sutinen, and M. Ben-Ari, "Visualizing programs with Jeliot 3", In Proceedings of the Working Conference on Advanced Visual Interfaces, ACM Press, pp. 373-376, 2004.

[16] L. Mannila, M. Peltomäki, and T. Salakoski, "What about a simple language? Analyzing the difficulties in learning to program." Computer Science Education, vol. 16, no. 3, 2006, pp. 211-227.

[17] M. Konecki, and V. Mrkela, "Students' acceptance of animated interactive presentation of sorting algorithms", In Proceedings of the 17th International Multiconference Information Society-Human-Computer Interaction in Information Society, pp. 18-21, 2014.

[18] M. Joy, N. Griffiths, and R. Boyatt, "The BOSS online submission and assessment system", ACM Journal of Educational Resources in Computing, vol. 5, no. 3, 2005, pp. 1-28.

[19] C. A. Higgins, G. Gray, P. Symeonidis, and A. Tsintsifas, "Automated assessment and experiences of teaching programming", ACM Journal on Educational Resources in Computing, vol. 5, no. 3, 2005, pp. 1-21.

[20] The NetBeans IDE BlueJ Plugin, https://edu.netbeans.org/bluej/, accessed: January 15th, 2016.

[21] M.-A. Storey, D. Damian, J. Michaud, D. Myers, M. Mindel, D. German, M. Sanseverino, and E. Hargreaves, "Improving the usability of Eclipse for novice programmers"; In Proceedings of the 2003 OOPSLA Workshop on Eclipse Technology Exchange", ACM Press, pp. 35-39, 2003.

[22] I. Utting, S. Cooper, M. Kölling, J. Maloney, and M. Resnick, "Alice, Greenfoot, and Scratch - A Discussion", ACM Transactions on Computing Education (TOCE), vol. 10, no. 4, 2010, 17.

[23] X. Chen, B. Francia, M. Li, B. Mckinnon, and A. Seker, "Shared information and program plagiarism detection", IEEE Transactions on Information Theory, vol. 50, no. 7, 2004, pp. 1545-1551.

[24] C. McDowell, L. Werner, H. Bullock, and J. Fernald, "The effects of pair-programming on performance in an introductory programming course". ACM SIGCSE Bulletin, vol. 34, no. 1, 2002, pp. 38-42.

[25] R. Lister, and J. Leaney, "Introductory programming, criterion-referencing, and bloom", ACM SIGCSE Bulletin, vol. 35, no. 1, 2003, pp. 143-147.

[26] P. Brusilovsky, E. Calabrese, J. Hvorecky, A. Kouchnirenko, and P. Miller, "Mini-languages: a way to learn programming principles", Education and Information Technologies, vol. 2, no. 1, 1997, pp. 65-83.

[27] N. R. Boyer, S. Langevin, and A. Gaspar, "Self direction & constructivism in programming education", In Proceedings of the 9th ACM SIGITE conference on Information technology education, pp. 89-94, 2008.

[28] F. Motalebi, "Towards Practical C++ Programming Projects for Pre-University Engineering Students", Asian Journal of Computer and Information Systems, vol. 1., no. 4., 2013, pp. 138-149.

[29] J. Huffman Hayes, "Energizing Software Engineering Education through Real-World Projects as Experimental Studies", In Proceedings of the 15th Conference on Software Engineering Education and Training, Covington, KY, USA, IEEE, pp. 192-206, 2002.

[30] R. K. Pucher, "Project Based Learning in Computer Science", In International Conference The Future of Education, 4th Edition, FLorence, Italy, Libreria Universitaria, pp. 245-248, 2014.

[31] A. S. Soares, C. R. Lopes. P. C. Souza, and S. C. Peres, "An Approach for Teaching Logic Programming based on Real-World Applications", In Proceedings of Frontiers in Education Conference, Oklahoma City, OK, USA, IEEE, pp. 861-865, 2013.

[32] K. Fertalj, B. Milasinovic, and I. Nizetic Kosovic, "Problems and Experiences with Student Projects Based on Real-World Problems: a case study", Technics Technologies Education Management, vol. 8, no. 1, 2013, pp. 176-186.

Making Programming Education More Accessible for Visually Impaired

Mario Konecki, Nikola Ivković and Matija Kaniški
Faculty of Organization and Informatics, Varaždin, Croatia
mario.konecki@foi.hr; nikola.ivkovic@foi.hr; matija.kaniski@foi.hr

Abstract - Modern programming paradigms have brought many different challenges for visually impaired programmers. Graphical interfaces have become one of the biggest obstacles for visually impaired programming professionals. The same problems that relate to visually impaired programmers also apply to programming education in which it is very challenging to make educational interfaces accessible for visually impaired and for other students with disabilities. In order to make programming more suitable for students with visual impairment an educational system that enables voice-based control of its elements has been developed for the purpose of their programming education. The results of using this kind of system and its impact on visually impaired students are presented and discussed in this paper.

Keywords **- programming, visually impaired, voice-based control, motivation**

I. INTRODUCTION

From the beginning of computer industry mass popularity, visually impaired have been included into activities of computer usage and into activities of programming. Programming has been an area of computer usage that gained interest among visually impaired computer users and this interest is also present today. Visually impaired have used computers with the help of various text-to-speech converters and the same technology enabled them to be included into programming activities. However, GUI (Graphical User Interface) has created problems for visually impaired programmers and has left them to deal with problems that include interpretation of different development environment parts, creation of graphical interfaces and usage of overall development environments that have become graphical and point-and-click based.

Programming courses have dealt with many problems for years and students have struggled, along with their teachers, to achieve an acceptable level of programming knowledge and skill. Teachers have problems with choosing the right methodology and the right tools for usage in education of programming and students have problems with understanding of abstract programming concepts, which results in fear and low motivation to learn programming.

Attitude of teachers and experts is that programming is hard to learn and teach and this is even more prominent in the case of students with different disabilities. Visually impaired students have problems with remembering the overall context of the program and with using the development environment that in many cases is not compatible with various aiding software that visually impaired use when working with computers. Another problem is that it is hard for visually impaired students to search through a rather large amount of digital online materials that are part of programming courses. In order to provide visually impaired with means that would enable them to go through their programming courses in a more suitable way, a system that features voice control has been proposed and named ABPT (Audio Based Programming Tutor). ABPT and results of its usage are presented in the rest of this paper.

II. DIFFICULTIES OF PROGRAMMING COURSES

Complexity of programming results in problems that affect both students and teachers. Programming requires a special way of thinking and also a special teaching and learning approach in order to achieve desired results. Students have particular problems in their efforts to acquire skills. Since programming is fundamentally a skill, it poses a special problem for them. Programming is also very abstract and not habitual for students. There are tools and attempts to help students in their task to learn programming but problems in programming courses have persisted to this day. It is very hard to pinpoint the reason for this kind of state as there are multiple factors involved. The problem of programming education can be viewed from five different angles [1]:

- The teaching methods
  - Teaching is not personalized
  - Teachers' strategies do not support all students' learning styles
  - The teaching of dynamic concepts through static materials
  - Teachers are more concentrated on teaching a programming language and its syntactic details, instead of promoting problem solving using a programming language
- The study methods
  - Students use incorrect study methodologies

    - Students do not work hard enough to acquire programming competences
- The student's abilities and attitudes
    - Students do not know how to solve problems
    - Many students do not have enough mathematical and logical knowledge
    - Students lack specific programming expertise
- The nature of programming
    - Programming demands a high level of abstraction
    - Programming languages have a very complex syntax
- Psychological effects
    - Students are not motivated
    - Students have to learn programming in a difficult period of their life

Programming courses have opted for object-oriented approach as the very first approach to be used for students, arguing the pedagogical value of this approach [2; 3; 4]. The problems of programming courses associated with procedural programming have however remained more or less the same as always [5; 6]. Students in many cases think that if the syntactic check is done and the code has no errors that the program must work. This points to the fact that students in many cases have not adopted a proper way of thinking required for programming tasks.

The fact is that many programming courses focus their teaching efforts on teaching students the syntax and how to write a program [7; 8], which does not teach them to analyze problems and to comprehend written code, which is necessary to solve programming tasks [9]. In order to promote comprehension of written programs and analytical skills, a solution that is based on test-first approach has been proposed. In this approach, students are required to solve programming tasks and to give a proper data test inputs in order to achieve several goals [9]:

- Assess the validity of the student's tests, giving feedback about which tests are incorrect
- Assess the completeness of the student's tests, giving an indication of how to improve
- Assess the style of the student's code, giving feedback about where improvements can be made
- Assess the quality of the student's code, giving suggestions for improvement or drawing attention to potential problems

In order to help students to learn programming in an easier manner, some authors suggest using game-based approach [10] and others ontology-based approach [11; 12].

Many different tools have been designed with the goal of making programming education easier for students. Some of these tools are [13]:

- Narrative tools which support programming to tell a story (e.g. Alice [14], Jeroo [15])
- Visual programming tools which support the construction of programs through a drag-and-drop interface (e.g. JPie [16], Alice [14], Karel Universe [17])
- Flow-model tools which construct programs through connecting program elements to represent order of computation (e.g. Raptor [18], Iconic Programmer [19])
- Specialized output realizations that provide execution feedback in non-textual ways, like multimedia or kinesthetic robotics (e.g. Lego Mindstorms [20], JES [21])
- Tiered language tools in which novices can use more sophisticated versions of a language as their expertise develops (e.g. ProfessorJ [22], RoboLab [23])

Despite of a large number of these tools, difficulties in programming courses have remained present and to learn how to program is considered as hard [24].

Another aspect that is of significant importance is the fact that students have very different preferable learning styles that should be considered. One of more known learning models that describes these styles is the Felder-Silverman Learning Style Model [25], which includes the following learning styles:

- Active learners (learn by trying things out, working with others) <--> Reflective learners (learn by thinking things through, working on their own)
- Sensing learners (concrete, practical, oriented toward facts and procedures) <--> Intuitive learners (conceptual, innovative, oriented toward theories and meanings)
- Visual learners (prefer pictures, diagrams, flow-charts) <--> Verbal learners (prefer written or spoken explanations)
- Sequential learners (learn in incremental, orderly steps) <--> Global learners (holistic, learn in large leaps)
- Inductive learners (prefer explanations that move from the specific to the general) <--> Deductive learners (prefer explanations that move from the general to the specific)

Students have adopted certain fear from learning programming that passes from one generation of students to another [26]. This makes the task to teach students programming even harder since they become skeptic about their programming courses.

There are many different approaches when talking about solving the problems of programming courses that

suggest different steps that should be incorporated in order to help students understand how to program. One of these approaches proposes the following set of steps [27]:

- Introduce additional programming course prior to introductory programming course that would promote algorithmic way of thinking
- Increase motivation of students for learning programming
- Explain to students that programming is a skill, not merely knowledge
- Introduce elements of constructivism into the teaching process
- Introduce learning by example
- Introduce animation and other visualization techniques combined with interaction
- Introduce interactive visual simulations
- Include support for multiple learning styles

III. AUDIO BASED PROGRAMMING TUTOR

One of the main problems of visually impaired programmers can be related to problems of visually impaired students who try to gain their programming education. Existing text-to-speech synthesizers that were sufficient before the GUI era have encountered a series of problems, such as interpretation of graphical elements and interpretation of overall context of the screen. Visually impaired students encountered several problems:

- Going through the learning materials
- Searching for particular information within the learning materials
- Controlling the development environments

These problems are of current nature because learning materials in most cases are not audio-based nor they support any interaction with text-to-speech synthesizers that are only able to read the text. The same applies for development environments that do not provide interaction with aiding technology and these environments in most cases also lack the support for any kind of voice control.

In order to try giving visually impaired students of programming courses and their teachers an aiding tool that will help them in learning of programming the ABPT (Audio Based Programming Tutor) has been proposed and developed as a potentially effective mean of teaching and learning programming for visually impaired students. ABPT is an audio-based tutoring system that supports voice control input. The system itself consists of the following elements:

- Audio lectures
- Lessons presentation materials
- List of programming tasks for homework and practice
- Programming tasks generator and compiler
- Virtual assistant that enables asking questions
- Message system for communication with the teacher

By using ABPT, visually impaired students are able to listen to audio lectures and go through presentations of different lessons by using voice control. They are also able to call the function that generates the list of programming tasks by using voice control and programming tasks are then read for them by using external text-to-speech synthesizer (such as JAWS or some similar tool). ABPT also enables visually impaired students to use built-in development environment and compiler that supports voice control and generates programming tasks. It supports controlling of the system and asking questions about programming by using natural language and it features a messaging system which supports voice created messages that are send to the teacher. ABPT provides visually impaired students with all necessary elements needed to make learning programming more convenient and easier. All mentioned elements also have their graphical representation, which enables deaf students to also use the system. Voice control is also more convenient for students with other disabilities. Part of ABPT is shown in Fig. 1.



Figure 1. Part of the ABPT

In order to test the ABPT, its prototype was made available to 31 visually impaired novice programmers from several different countries. ABPT prototype consisted of all mentioned functionalities of ABPT. Lectures, materials, and programming tasks that were included have covered all programming concepts that are part of most introductory programming courses. All features of ABPT were developed as built-in functionalities except the compiler and text-to-speech synthesizer for which external tools were used.

All visually impaired novice programmers who participated in the research have been included based on their active participation in several Internet groups of visually impaired programmers and their willingness to participate in the research. There were 22 male and 9 female participants. 24 participants have had less than 1 year of programming experience and 7 participants have had almost no prior programming experience. Average age of all participants was 29.

All participants were asked to go through the ABPT and to report about their experience with the system by answering several questions graded using Likert scale from 1 to 5 (1 - I disagree completely, 5 – I agree

completely). Results of conducted research are given in Table I.

TABLE I. RESULTS OF CONDUCTED RESEARCH

Questionnaire item	Mean	Std. dev.
ABPT is a promissing solution for visually impaired students.	4,15	0,37
I was able to use ABPT without any major problems.	4,03	0,62
Voice controls enabled me to use ABPT in an intuitive and easy manner.	3,98	0,97
ABPT is well-rounded learning tool.	4,18	0,65
I would recommend using ABPT in programming education.	4,01	0,53

According to the research results it can be concluded that ABPT is a promising aiding mean that could be used in programming education and that could help visually impaired and other students with disabilities to learn programming in a more suitable and more convenient way.

Continuous monitoring of students' advancement and possible obstacles during teaching and learning process should enable further improvement of the tool and adjustment of its inclusion in the educational process.

IV. CONCLUSION

Students and teachers of programming courses have been dealing with various problems in achieving certain level of knowledge and skills for years. Problems in programming courses have been even more prominent in the case of visually impaired students and students with other disabilities. In order to provide these students with a potentially usable solution in their programming education, an ABPT (Audio Based Programming Tutor) has been proposed and evaluated. Research results have shown that ABPT is useful and applicable tool and that it can be used in programming education for visually impaired students and students with other disabilities. Further testing and improvement of ABPT will be part of future research.

REFERENCES

[1] A. Gomes, and A. J. Mendes, "Learning to program-difficulties and solutions", In International Conference on Engineering Education – ICEE, 2007.

[2] J. Bergin et al., "Resources for next generation introductory CS courses: report of the ITiCSE'99 working group on resources for the next generation CS 1 course", ACM SIGCSE Bulletin, vol. 31, no. 4, 1999, pp. 101-105.

[3] M. Kölling, and Rosenberg, J., "Guidelines for teaching object orientation with Java", ACM SIGCSE Bulletin, 2001, pp. 33-36.

[4] M. Kölling, and J. Rosenberg, "BlueJ – the hitchhiker's guide to object orientation", The Maersk Mc-Kinney Moller Institue for Production Technology University of Southern Denmark Technical Reports, 2002, 2, 1-8.

[5] R. Decker, and S. Hirshfield, "The top 10 reasons why object-oriented programming can't be taught in CS 1", ACM SIGCSE Bulletin, vol. 26, no. 1, 1994, pp. 51-55.

[6] K. L. Krause, "Computer science in the Air Force Academy core curriculum", ACM SIGCSE Bulletin, vol. 14, no. 1, 1982, pp. 144-146.

[7] D. Buck, and D. J. Stucki, "JKarelRobot: a case study in supporting levels of cognitive development in the computer science curriculum". ACM SIGCSE Bulletin, vol. 33, no. 1, 2001, pp. 16-20.

[8] D. Buck, and D. J. Stucki, "Design early considered harmful: graduated exposure to complexity and structure based on levels of cognitive development", ACM SIGCSE Bulletin, vol. 32, no. 1, 2000, pp. 75-79.

[9] S. H. Edwards, "Rethinking computer science education from a test-first perspective". In Companion of the 18th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications, pp. 148-155, 2003.

[10] S. Leutenegger, and J. Edgington, "A games first approach to teaching introductory programming", ACM SIGCSE Bulletin, vol. 39, no. 1, 2007, pp. 115-118.

[11] C. Pierrakeas, G. Solomou, and A. Kameas, "An Ontology-Based Approach in Learning Programming Languages", In Proceedings of the 16th Panhellenic Conference on Informatics (PCI), pp. 393-398, 2012.

[12] M. Konecki, and S. Lovrenčić, "Ontology-based approach in education of programming", In International Academic Conference on Social Sciences and Humanities in Prague 2015, pp. 347-352, 2015.

[13] K. Powers et al., "Tools for teaching introductory programming: what works?", ACM SIGCSE Bulletin, vol. 38, no. 1, 2006, pp. 560-561.

[14] B. Moskal, D. Lurie, and S. Cooper, "Evaluating the effectiveness of a new instructional approach", ACM SIGCSE Bulletin, vol. 36, no. 1, 2004, pp. 75-79.

[15] D. Sanders, and Dorn, B., "Classroom experience with Jeroo", Journal of Computing Sciences in colleges, vol. 18, no. 4, 2003, pp. 308-316.

[16] K. J. Goldman, "A concepts-first introduction to computer science", ACM SIGCSE Bulletin, vol. 36, no. 1, 2004, pp. 432-436.

[17] J. Bergin, "Karel universe drag & drop editor", ACM SIGCSE Bulletin, vol. 38, no. 3, 2006, pp. 307-307.

[18] M. C. Carlisle, T. A. Wilson, J. W. Humphries, and S. M. Hadfield, "RAPTOR: a visual programming environment for teaching algorithmic problem solving", ACM SIGCSE Bulletin, vol. 37, no. 1, 2005, pp. 176-180.

[19] S. Chen, and S. Morris, "Iconic programming for flowcharts, java, turing, etc.", ACM SIGCSE Bulletin, vol. 37, no. 3, 2005, pp. 104-107.

[20] J. S. Kay, "Teaching robotics from a computer science perspective", Journal of Computing Sciences in Colleges, vol. 19, no. 2, 2003, pp. 329-336.

[21] L. Rich, H. Perry, and M. Guzdial, "A CS1 course designed to address interests of women", ACM SIGCSE Bulletin, vol. 36, no. 1, 2004, pp. 190-194.

[22] K. E. Gray, and M. Flatt, "ProfessorJ: a gradual introduction to Java through language levels". In Companion of the 18th Annual ACM SIGPLAN Conference on ObjectOriented Programming, Systems, Languages, and Applications, 170-177, 2003.

[23] B. Erwin, M. Cyr, and C. Rogers, "Lego engineer and robolab: Teaching engineering with labview from kindergarten to graduate school", International Journal of Engineering Education, vol. 16, no. 3, 2000, pp. 181-192.

[24] A. Robins, J. Rountree, and N. Rountree, "Learning and teaching programming: A review and discussion", Computer science education, vol. 13, no. 2, 2003, pp. 137-172.

[25] L. Thomas, M. Ratcliffe, J. Woodbury, and E. Jarman, "Learning styles and performance in the introductory programming sequence", ACM SIGCSE Bulletin,vol. 34, no. 1, 2002, pp. 33-37.

[26] M. Konecki, "Improving Students' Focus in Introductory Programming Courses", DAAAM International Scientific Book, 2015, pp. 165-172.

[27] M. Konecki, "Problems in programming education and means of their improvement", DAAAM International Scientific Book, 2014, pp. 459-470.

Use of Computer Programs in Teaching Photography Courses at Schools of Applied Arts and Design in Croatia

Zvjezdana Prohaska, BFA*, Zdenko Prohaska, PhD**, Ivan Uroda, PhD***
* Senior Adviser for Visual Arts, Education and Teacher Training Agency, Office Branch Rijeka, Rijeka, Croatia
** Full Professor, University of Rijeka, Faculty of Economics, Rijeka, Croatia
*** Senior Assistant and Research Associate, University of Rijeka, Faculty of Economics, Rijeka, Croatia
zvjezdana.prohaska@azoo.hr, zdenko.prohaska@ri.t-com.hr, ivan.uroda@gmail.com

Abstract - By going digital photography becomes very popular in general and more and more special photography courses are introduced at Croatian schools of applied arts and design.

In this article different computer programs for optimizing photographs, which are actually used in class, will be presented and their advantages and eventual disadvantages will be discussed and statistically evaluated by using average values and standard deviation values.

I. Introduction

This year's paper is based on survey that helped authors to identify and pinpoint both the most significant and the most critical issues regarding computer programs usage in 3 Croatian Schools of Applied Arts and Design that are actually vocational secondary schools. Survey with a total of 33 questions was distributed in 1st, 2nd, 3rd and 4th grades of all 3 aforementioned schools. Statistical analysis of gathered data by average value and standard deviation has been performed on 2 levels in order to properly identify key reasons for and repercussions of pupils' answers.

II. Characteristics of Computer Programs for Photo Editing from the Survey

In the survey that is the integrative part of this paper authors have offered to pupils a total of 4 computer programs for photo editing. Moreover, 2 of them were commercial (Adobe Photoshop and CorelDraw) and 2 of them were free (Gimp and Picasa) [1] [2] [3] [4]. Tabular representations of their most distinguished and comparable characteristics are located below this paragraph.

Name of program		**Adobe Photoshop**	
Developer		**Adobe Systems**	
Operating system(s)		**Windows, OS X**	
Type		**Commercial**	
First version	**1990**	Latest version	**2016**

Name of program		**CorelDraw**	
Developer		**Corel**	
Operating system(s)		**Windows**	
Type		**Commercial**	
First version	**1989**	Latest version	**2014**

Name of program		**Gimp**	
Developer		**GIMP Development Team**	
Operating system(s)		**Linux, OS X, Windows…**	
Type		**Free**	
First version	**1995**	Latest version	**2015**

Name of program		**Picasa**	
Developer		**Google**	
Operating system(s)		**Windows, OS X**	
Type		**Free**	
First version	**2002**	Latest version	**2016**

While every one of those programs has its own advantages and disadvantages, yet they all represent software with more than 10 years of development tradition, while first 3 of them have development tradition that spans over 3 decades. That is exactly what users expect from this type of software i.e. reliability, consistency and continuity.

III. Aggregate Survey Results on Programs for Photo Editing

Specific scientific research conducted in this paper required the entirely new and original survey. This survey was carried out in 3 Schools of Applied Arts and Design in Croatia, Europe that exclusively provide study program for occupation called Photographic Designer.

Significantly, those vocational secondary schools are located in 3 separate Croatian cities and 3 separate Croatian counties. Names of these 3 schools and their cities as well as their counties are:

- School of Applied Arts and Design in Zagreb (Zagreb County)
- School of Visual Arts in Split (Split-Dalmatia County)
- School of Applied Arts and Design in Zadar (Zadar County).

One might notice that school in Split has a slightly different name, but that is because there is a school named School of Design, Graphics and Sustainable Building within the same city. It has certain elements in its name that may mislead uniformed ones, but this school does not provide program for Photographic Designers whatsoever.

In addition, there is also the definition regarding Photographic Designer program that is included in curriculum by School of Applied Arts and Design in Zadar. It states that they "… try to encompass the entire area of photography as art and design discipline, by following technical and technological advancements within photography profession. Pupils work with…digital cameras, video camcorders, … Part of the program is performed on computers. …" what implies that working knowledge of modern technologies and computers is one of the key competences that pupils should acquire and outcomes that pupils should achieve within this program.

Hence, this survey was intentionally directed towards future photographers that also have design competences for whom computer programs, mainly those specialized for photo editing should not be thing of the future, but they should be something that they consider as primary and contemporary tool of their trade. What is more, since those pupils went for vocational secondary school and education per se, those schools should be(come) the place where they will learn both solid foundations and applicable knowledge that will properly prepare them for specialized jobs once they enter the workforce.

All this is exactly why this aforementioned survey was primarily directed towards (computer) programs for photo editing and prevailing pupils' attitudes towards their various features and functionalities. Hence, through 3 general information questions and subsequent series of 30 specific questions pupils were able to anonymously and appropriately voice their opinion on various aspects covered in the survey. Thus, they could identify their gender, grade and city in first 3 non-numerated questions together with 27 Likert scale questions where they could numerically rate each statement from 1 (the lowest grade) to 5 (the highest grade) along with 3 open-type questions in which pupils could freely write their responses.

In the terms of analysis, there were essentially 2 main approaches, types of reasoning and calculations. Firstly, for 3 non-numerated questions and 3 open-type questions, all of the responses have been counted, sorted and used for further analysis. Secondly, for 27 Likert scale questions, statistical measures called average and standard deviation have been calculated in order to display and to elaborate the arithmetic means of pupils' grades in certain questions and the dispersion of values around arithmetic mean value.

Tabular data representations for questions below represent the integrative part of the text above and the text below. In those terms, asterisk was used to denote label with number of survey respondents, capital letters were used to denote labels for non-numerated questions and numbers were used to denote all of the labels for numerated questions. As a final point, labels do not include the whole questions that pupils were asked, however they do include key points that were essentially included in all of the related questions.

*) Number of survey respondents among pupils

RESPONDENTS	82	100,00%

According to the numbers stated in [5], [6], [7] and [8] it can be estimated that the total number of pupils who currently attend specific Photographic Designer program at 3 Schools of Applied Arts and Design is currently somewhere between 96 (4 grades×8 pupils×3 cities) and 144 (4 grades×12 pupils×3 cities). Hence, survey sample that includes somewhere between 56,94% and 85,42% of maximum number of respondents can be considered as more than representative.

A) Pupils' genders

MALE	19	23,17%
FEMALE	60	73,17%
N/A	3	3,66%

Currently there are more than 3 times more female pupils than male pupils among respondents of this survey in Photographic Designer programs. Surprisingly, 3 pupils did not select their gender.

B) Pupils' grades

1ST	1	1,22%
2ND	21	25,61%
3RD	35	42,68%
4TH	25	30,49%

Distribution of those pupils who were survey respondents by grades is not equal, but that was not even authors' goal. This survey was sent to 3 aforementioned schools while distribution to pupils was entrusted to teachers therein. Thus, these numbers were neither planned nor forced.

C) Pupils' cities

ZAGREB	28	34,14%
SPLIT	29	35,37%
ZADAR	25	30,49%

In contrast with the previous questions, distribution of pupils between cities is much more harmonized. That is excellent for this entire scientific research because there is absolutely no bias among aggregate survey results that are in favour of certain city.

1) Pupils' interest in classes in which programs for photo editing are used

AVERAGE	4,06	**ST. DEV.**	0,95

The highest average value and almost the lowest standard deviation value indicate that pupils are unquestionably interested in classes in which programs are used and that their opinions do not vary too much. This is the first indicator that this year's scientific research was spot on since pupils' feedback should be considered as one of the most important in this regard.

2) Pupils' self-assessed general knowledge of computer usage

AVERAGE	3,91	**ST. DEV.**	0,88

High average value and the lowest standard deviation value indicate that these pupils hold their general knowledge of computer usage in a rather high regard. It is important to emphasise that these numbers are pupils' self-assessment that should be tested to get accurate and actual results. Nevertheless, such level of self-confidence should be a good starting point if those pupils will have to learn some new computer-related skills and/or techniques.

3) Pupils' self-assessed knowledge of programs for photo editing usage

AVERAGE	3,57	ST. DEV.	0,97

Once again, combination of rather high average value on top of low standard deviation value more or less implies the same conclusion like in the previous question, but somewhat lower average value clearly indicates pupils' awareness that programs are harder to master in comparison to general computer usage.

4) Pupils' estimate of measure to which programs for photo editing should be introduced in classes

AVERAGE	3,94	ST. DEV.	1,06

Very high average value and standard deviation value that is on the borderline between low and medium indicates that generally pupils are in for the introduction of programs in classes, but their opinions differ more than in the previous questions. This is actually very stimulating dataset that makes all 30 questions in this scientific research justifiable so that the most significant factors that directly influence introduction of programs can be properly discovered and elaborated.

5) Pupils' estimate of whether the introduction of programs for photo editing in classes would make their learning materials easier to master

AVERAGE	3,88	ST. DEV.	1,17

Average value is once again high for this question, but moderate standard deviation value indicates certain amount of differentiation among pupils' estimates on whether proposed innovations would make one of the most demanding pupils' tasks easier. In those terms, value for standard deviation would decrease if pupils were presented with real advantages of modern learning model.

6) Pupils' estimate of whether the introduction of programs for photo editing in classes would make classes more interesting

AVERAGE	3,91	ST. DEV.	1,17

Incredibly similar values like in the previous question, with only 0,03 difference in average value actually imply the same conclusion. Thus, it is completely understandable that pupils may have difficulties with comparing their current learning model with the proposed one that they know little or nothing about.

7) Pupils' estimate on how much does their weekly number of school hours enable usage of programs for photo editing

AVERAGE	3,50	ST. DEV.	1,00

High average value and low standard deviation value indicate that pupils do recognize that there is definitely space for improvements that should be smartly and strategically integrated in Photographic Designer program.

8) Pupils' estimate on how much things that they learn in current grade enable usage of programs for photo editing

AVERAGE	3,66	ST. DEV.	1,11

Rather high average value in indicative conjunction to moderate standard deviation value actually represent excellent platform for the proposed innovation. It is completely understandable that pupils' estimate differ to certain extent because every grade actually learns something different and not all teaching themes support usage of programs.

9) Pupils' assessment of former level of programs for photo editing usage in their classes

AVERAGE	3,61	ST. DEV.	1,00

With somewhat high average value and fairly low standard deviation value, pupils' answers to this question once again support the claim that there is particular room for innovations. In those terms, it is quite encouraging that pupils perceive former level of usage of programs as average to more than average.

10) Pupils' estimate on how often does their teacher realize the whole teaching theme in computer lab

AVERAGE	3,63	ST. DEV.	1,32

Somewhat high average value and rather high standard deviation value indicate that this concept is not something completely new to the pupils, but there is a great divergence in their answers. There are many possible reasons for that, but previous scientific research by one of the authors [9] has shown that it mostly depends on grade, teacher and the availably of computer classroom in a certain school.

11) Pupils' estimate on how often does their teacher realize the whole teaching theme with the help of programs for photo editing (NOT in computer lab)

AVERAGE	3,07	ST. DEV.	1,18

In contrast with the previous question, nothing more than moderate both average value and standard deviation value indicate that most pupils get only average amount of whole teaching themes with the help of programs. However, what is worrying is that certain number of pupils who get almost no chance for this type of classes.

12) Pupils' opinion on to which extent would they want their teacher to use programs for photo editing for teaching new learning materials

AVERAGE	3,83	ST. DEV.	1,02

Quite high average value coupled with reasonably high standard deviation value both indicate that pupils want innovations in their classes, even when they will be learning something new.

13) Pupils' opinion on to which extent would they want their teacher to use programs for photo editing for repeating learning materials

AVERAGE	3,49	ST. DEV.	1,09

High average value in conjunction with just moderate standard deviation value indicate that the extent of innovations that pupils want in their classes decreases when it comes to repetition of something previously learnt.

14) Pupils' opinion on to which extent would they want their teacher to use programs for photo editing for knowledge checks

AVERAGE	3,57	ST. DEV.	1,13

Once again, there is high average value in conjunction with moderate standard deviation value, but in this question they can indicate that pupils' maybe even anticipated a slight discomfort with teachers who could use programs for knowledge checks.

15) Pupils' opinion on to which extent would they want their teacher to use programs for photo editing for competitions (inside classes)

AVERAGE	3,06	ST. DEV.	1,26

Only moderate average value and high standard deviation value indicate that the average pupil does not appreciate competitions inside classes. However there is a great dispersion of pupils' answers what indicates that there were ones who are almost utterly for and against this.

16) Pupils' opinion on to which extent would they want their teacher to use programs for photo editing for competitions (outside classes)

AVERAGE	3,49	ST. DEV.	1,26

Fairly large average value and the same and high standard deviation suggest that pupils generally prefer program supported competitions, but again the repeated dispersion of answers once again indicates the significant differences in pupils' opinions.

17) Pupils' opinion on whether programs for photo editing are the most useful for work at school

AVERAGE	3,60	ST. DEV.	1,06

Rather high average value and standard deviation value on the borderline between high and moderate indicate that pupils would generally like using programs at school.

18) Pupils' opinion on whether programs for photo editing are the most useful for work at home

AVERAGE	3,49	ST. DEV.	1,11

High average value and moderate standard deviation value means that pupils would pretty much like using programs at home, but there is greater dispersion in their answers.

19) Pupils' opinion on whether the exchange of files from programs for photo editing that teachers create should be encouraged

AVERAGE	3,44	ST. DEV.	1,09

High average value and moderate standard deviation value in this question suggest that pupils would generally like their teachers to exchange files that they created.

20) Pupils' opinion on whether the exchange of files from programs for photo editing that pupils create should be encouraged

AVERAGE	3,29	ST. DEV.	1,05

Moderate average value and high standard deviation value in this question suggest that pupils would rather like to exchange files that they created.

21) Pupils' opinion on whether the exchange of files from programs for photo editing between teachers and pupils should be encouraged

AVERAGE	3,80	ST. DEV.	1,10

Surprisingly high average value and moderate average value represent the positive improvement from the previous 2 questions since pupils would like to exchange files with their teachers. In the context of this scientific research this is excellent finding because cooperation between pupils and teachers is key to successfully facilitate usage of programs in classes.

22) Pupils' satisfaction with computer classroom in their school while using programs for photo editing

AVERAGE	2,43	ST. DEV.	1,23

Almost the lowest average value in combination with moderate standard deviation value is rather worrying, especially in this specific question that pinpoints one of the longstanding Achill's heels of Croatian education system. In general, computer classrooms are at least somewhat outdated, but such standard deviation value shows that there is also a great discrepancy in this regard.

23) Pupils' observation on how often do they use computer classroom in their school when they are using programs for photo editing

AVERAGE	3,68	ST. DEV.	1,33

Quite high average value and high standard deviation is both encouraging and worrying because some pupils definitely have better access to computer classroom. Difference could be at both school level and grade level what makes this question essential for more detailed scientific researches on this topic.

24) Pupils' observation of the most frequent problems that have happened to their teachers or themselves while using programs for photo editing in classes

In short, 44 out of 82 pupils (53,66%) have reported various hardware issues and 12 out of 82 pupils (14,63%) have reported various software issues. Even though younger generations have much less tolerance for slow(er) computers, over 50,00% respondents should be taken seriously in the terms of hardware improvements.

25) Pupils' list of programs for photo editing that they have used thus far

Numbers-wise:
78 out of 82 pupils (95,12%) have used Photoshop
19 out of 82 pupils (23,17%) have used Picasa
1 out of 82 pupils (1,22%) have used Corel
1 out of 82 pupils (1,22%) have used Gimp.
Quite understandably, Photoshop by Adobe has taken the first place. Encouraging is the fact that free software Picasa by Google has taken the second place.

26) Pupils' opinion on what is the best program for photo editing that they have used thus far

Once again, numbers-wise:
78 out of 82 pupils (95,12%) have opted for Photoshop
19 out of 82 pupils (23,17%) have opted for Picasa
1 out of 82 pupils (1,22%) have opted for Corel
1 out of 82 pupils (1,22%) have opted for Gimp.
As the matter of fact, comparison has proven that those values are in perfect correlation with the ones from the previous question. Looks like path of least resistance…

27) Pupils' opinion on whether all legally acquired programs for photo editing should be paid for

AVERAGE	2,15	ST. DEV.	1,44

Absolutely lowest average value and absolutely highest standard deviation value is the amazing coincidence that occurred in this question. Pupils could not differ more in opinions than in this question, but it is quite possible that the number of them already know of freeware as well as of open source software and the fact that there is certain commercial software that is available to pupils for free.

28) Pupils' opinion on whether there are quality and free programs for photo editing

AVERAGE	2,51	**ST. DEV.**	1,42

Quite low average value in conjunction to rather high standard deviation value show that in general pupils definitely lack education in this regard, but there are also some exceptions that prove utter ignorance, but also pretty elevated attitude towards quality and free programs.

29) Pupils' interest in education in advantages and features of free programs for photo editing

AVERAGE	3,21	**ST. DEV.**	1,26

Mixture of moderate average value and standard deviation value that is on the borderline between moderate and low indicates that pupils are, in general, just averagely interested in education that is directly related to their program and that can save them large amount of money.

30) Pupils' interest in further researches related to introduction of free programs for photo editing in classes

AVERAGE	3,54	**ST. DEV.**	1,33

In this case average value is somewhat high, while standard deviation value is quite high. While this may seem strange to some, these values reflect what is almost always present in this type of scientific research. While respondents are generally willing to participate in further researches, there are always some who are extremely interested in it and some who do not want know anything else about it.

IV. Analysis of Whether Grade Matters in Aggregate Survey Results

Under this heading authors have made the inquiry into whether grade matters in aggregate survey results. For that purpose both average values and standard deviation values by grades have been subtracted from their aggregate values. That way, a quick numerical reference to whether those values are larger or smaller tin certain grades' was formed. Then, 14 out of 27 questions with numerical answers were filtered out and elaborated because of certain upward trend or downward trend depending on the grade(s). Regarding that matter, values from 2nd, 3rd and 4th grades will be taken into account because only they have representative number of pupils for this analysis.

1) Pupils' interest in classes in which programs for photo editing are used

A (1ST)	*-0,06*	*S (1ST)*	*N/A*
A (2ND)	-0,11	S (2ND)	+0,08
A (3RD)	+0,05	S (3RD)	-0,05
A (4TH)	+0,02	S (4TH)	+0,05

It seems that pupils of 3rd grades and 4th grades are little-bit-more interested in this matter.

4) Pupils' estimate of measure to which programs for photo editing should be introduced in classes

A (1ST)	*-1,94*	*S (1ST)*	*N/A*
A (2ND)	-0,03	S (2ND)	+0,20
A (3RD)	+0,00	S (3RD)	-0,15
A (4TH)	+0,10	S (4TH)	+0,00

Interestingly, average value of 3rd grade pupils' corresponds with the average value of the entire pupil population, whereas average value 4th grade pupils show positive tendency because of increased interest in the topic.

5) Pupils' estimate of whether the introduction of programs for photo editing in classes would make their learning materials easier to master

A (1ST)	*-0,88*	*S (1ST)*	*N/A*
A (2ND)	-0,07	S (2ND)	-0,05
A (3RD)	+0,01	S (3RD)	+0,09
A (4TH)	+0,08	S (4TH)	-0,03

Once again, slight upward trend that is especially visible in 3rd and 4th grades and that is positive tendency.

7) Pupils' estimate on how much does their weekly number of school hours enable usage of programs for photo editing

A (1ST)	*+0,50*	*S (1ST)*	*N/A*
A (2ND)	-0,74	S (2ND)	+0,09
A (3RD)	+0,19	S (3RD)	-0,13
A (4TH)	+0,34	S (4TH)	-0,20

While average value of 2nd grade pupils may indicate that 3rd and 4th grade indeed have more hours at which they can use programs.

8) Pupils' estimate on how much things that they learn in current grade enable usage of programs for photo editing

A (1ST)	*+0,34*	*S (1ST)*	*N/A*
A (2ND)	-0,61	S (2ND)	+0,05
A (3RD)	+0,28	S (3RD)	-0,03
A (4TH)	+0,10	S (4TH)	-0,14

It looks like 3rd grade pupils learn the most things that are related to programs, with 4th grade pupils slightly lagging.

9) Pupils' assessment of former level of programs for photo editing usage in their classes

A (1ST)	*+1,39*	*S (1ST)*	*N/A*
A (2ND)	-0,47	S (2ND)	+0,11
A (3RD)	+0,16	S (3RD)	-0,31
A (4TH)	+0,11	S (4TH)	+0,17

Similar to pervious question, 3rd grade has even better experience than 4th grade, but with smaller difference.

11) Pupils' estimate on how often does their teacher realize the whole teaching theme with the help of programs for photo editing (NOT in computer lab)

A (1ST)	*-2,07*	*S (1ST)*	*N/A*
A (2ND)	-0,79	S (2ND)	-0,34
A (3RD)	+0,38	S (3RD)	-0,01
A (4TH)	+0,21	S (4TH)	-0,08

Very comparable results with the ones from the previous 2 questions, but even higher average values indicate that 3rd grade and 4th grade pupils are in even better position here.

16) Pupils' opinion on to which extent would they want their teacher to use programs for photo editing for competitions (outside classes)

A (1ST)	*-0,49*	*S (1ST)*	*N/A*
A (2ND)	-0,11	S (2ND)	+0,10
A (3RD)	-0,14	S (3RD)	+0,09
A (4TH)	+0,31	S (4TH)	-0,22

Interestingly, this question and the following 4 questions i.e. questions 16), 17), 18), 19) and 20) all follow exactly the same pattern. That pattern shows that 4th grade pupils definitely have more mature attitude towards some of the vital questions within this scientific research.

17) Pupils' opinion on whether programs for photo editing are the most useful for work at school

A (1ST)	*+1,40*	*S (1ST)*	*N/A*
A (2ND)	-0,03	S (2ND)	+0,01
A (3RD)	-0,23	S (3RD)	-0,00
A (4TH)	+0,28	S (4TH)	-0,05

Explanation from question 16) is applied here.

18) Pupils' opinion on whether programs for photo editing are the most useful for work at home

A (1ST)	*-1,49*	*S (1ST)*	*N/A*
A (2ND)	-0,25	S (2ND)	+0,15
A (3RD)	-0,06	S (3RD)	+0,03
A (4TH)	+0,35	S (4TH)	-0,26

Explanation from question 16) is applied here.

19) Pupils' opinion on whether the exchange of files from programs for photo editing that teachers create should be encouraged

A (1ST)	*-1,44*	*S (1ST)*	*N/A*
A (2ND)	-0,11	S (2ND)	+0,02
A (3RD)	-0,15	S (3RD)	-0,04
A (4TH)	+0,36	S (4TH)	-0,01

Explanation from question 16) is applied here.

20) Pupils' opinion on whether the exchange of files from programs for photo editing that pupils create should be encouraged

A (1ST)	*-0,29*	*S (1ST)*	*N/A*
A (2ND)	-0,29	S (2ND)	-0,05
A (3RD)	-0,29	S (3RD)	-0,05
A (4TH)	+0,67	S (4TH)	-0,16

Explanation from question 16) is applied here.

27) Pupils' opinion on whether all legally acquired programs for photo editing should be paid for

A (1ST)	*+0,85*	*S (1ST)*	*N/A*
A (2ND)	+0,28	S (2ND)	+0,06
A (3RD)	-0,12	S (3RD)	-0,10
A (4TH)	-0,11	S (4TH)	+0,13

In this question, negative average values related to pupils from 3rd and 4th grade are more keenly aware of ways how to freely and legally obtain needed software.

28) Pupils' opinion on whether there are quality and free programs for photo editing

A (1ST)	*+1,49*	*S (1ST)*	*N/A*
A (2ND)	+0,34	S (2ND)	+0,04
A (3RD)	-0,23	S (3RD)	-0,10
A (4TH)	-0,03	S (4TH)	+0,09

There is no straightforward explanation of the fact that younger pupils actually hold quality and free programs in higher regard than the older ones. This specific tendency can also be attributed to the fact that pupils develop fascination with trademarks, but somewhat higher average value shows that 4th grade pupils have more sombreness in this regard.

V. CONCLUSION

First part encompassed the quick introduction to software pieces that were used in survey. Intentionally, focus was put on the most visible and the most comparable characteristics. Alas, Picasa has been retired / discontinued by Google's development team [10] as authors were putting the final touches on this paper.

In the second part, the thorough analysis has shown that pupils recognize that there is deficiently some room for innovations. Through a number of various questions pupils have shown good, very good and excellent attitudes towards various innovations that would primarily include programs for photo editing as the main driving force. What is more, pupils also acknowledged that they would use programs for photo editing both at school and at home, but their willingness to share files created by them and by their teacher is somewhat lower than expected. However, they have positive attitude towards file sharing with their teachers what is even more important. Worrying is the fact that majority of pupils have allegedly reported various hardware issues with computers at their schools. While pupils selected commercial software as the best one, they selected free software as the second best and that is great for the innovations that authors propose. As a final point, overall pupils' attitudes towards free and commercial types of software have yielded some encouraging results that will definitely be included in new research by the same authors.

Results of specific analysis from the third part were based on carefully selected questions. It has shown that in some regards lower grades i.e. 2nd grade are ahead of 3rd and 4th grade, but in some key regards analysis has shown that older pupils are positively more mature and sensible.

To sum up, this paper has proven that pupils want innovations in their classes, namely they are certainly interested in education that incorporates the benefits of computer programs for photo editing. Hence, regardless of pupils' diverse levels of interest in various innovations that were proposed in this survey, overall and grade-specific differences in their opinions and issues that they listed, it can be concluded that pupils wish for and would welcome all the positive innovations that would be properly, mindfully, gradually and attractively introduced in their classes.

REFERENCES

[1] http://www.adobe.com/aboutadobe/pressroom/pdfs/timeline.pdf, visited on 17.02.2016

[2] https://www.corel.com/content/pdf/corporate/History_of_Corel.pdf, visited on 17.02.2016

[3] https://www.gimp.org/about/history.html, visited on 17.02.2016

[4] http://picasa.en.lo4d.com, visited on 17.02.2016

[5] Škola primijenjene umjetnosti i dizajna, Popis razreda i broj učenika po zanimanjima (2010/11), Zagreb, 2011

[6] Škola primijenjene umjetnosti i dizajna, Popis razreda i broj učenika po zanimanjima (2011/12), Zagreb, 2012

[7] Škola primijenjene umjetnosti i dizajna, Popis razreda i broj učenika po zanimanjima (2012/13), Zagreb, 2013

[8] Škola primijenjene umjetnosti i dizajna, Popis razreda i broj učenika po zanimanjima (2013/14), Zagreb, 2014

[9] Prohaska, Z., Main Prerequisites for Introduction of ICT in Visual Arts Education in Croatian Primary Schools, Opatija, 2011

[10] Sabharwal, A., Moving on from Picasa, online, 2016

Universtiy Search Engine

Željko Knok* and Mark Marčec **
* Polytechnic of Međimurje in Čakovec/Computer engineering, Čakovec, Croatia
mmarcec1@student.mev.hr

Abstract – On university's web portals are available data about the study program, but it is really hard and time wasting for a high school student to find the data that is need. For easier choosing of a study program, there is a need for a dynamic web application. This application is implemented in open source technologies such as (HTML5, JavaScript, PHP and MySQL). They support installation on Windows and Linux environments. XAMPP is required for the development. This application gives information to a future collegian about the study programs and university's in one place. Information like requirements for admission, terms for admission, location of the university shown on map, and workflow for admission which was barely available. The application could be used on all devices that have access to the internet, no matter which operating system or screen size they have.

I. INTRODUCTION

After many students finish they high school, than they have to choose a study program to finish their education and eventually to get a great job. And many students often don't know what to choose and they are worried about something they don't know. Students also often don't know about requirements for each university and study program.

Presentations are organized by universities for interested students, but someone has to come to the presentation, either the university is sending somebody to a high school to present the university or some students are allowed to travel to the university to see the presentation and to speak to other students. This is nice but isn't effective. Because not everyone has great presentation, not everyone can travel to university to see the presentation and not everyone is interested is interested in the presentation.

There is also a bunch of portals on the web of universities that give an overview of their study programs and requirements but it isn't so easy to find all the important data and is a time waster.

Postani Student [6] is a search engine for universities and study programs, but their search options often confuse students and at the end it shows only the basic data about the study program. And this data are, where the study program is performing, its name and basic requirements.

In order to help students and anyone who is interested to study at a university, to find the right study program quick and easy, we developed USE (University Search Engine). It is a dynamic web application that offers all the important data about study programs in one place. Useful search filters depending by some criteria like study type, location, fields and types of universities. The filtered results will contain a course overview, all the important dates, map locations and also the universities.

II. APPLICATION OVERVIEW

USE or University Search Engine is a dynamic web application developed in PHP with an MVC framework called CodeIgniter 3[2]. PHP is a popular general-purpose scripting language that is especially suited to web development. Fast, flexible and pragmatic, PHP powers everything from your blog to the most popular websites in the world. [4]. A server is required for running PHP scripts. First, the application is tested on a local server that can be installed with XAMPP a PHP development environment which includes Apache server, MySQL, Perl and other useful applications and features.

The key features of the University Search Engine are searching all study programs through the internet based by most common search terms and conditions and carries over to USE (University Search Engine). A nice and attractive view of available study programs with important data, well designed study program detailed overview, also an overview of the universities and its important data. Therefore, the PHP libraries where added to simplify programming and debugging of the web application.

A. Set up the styling and responsive web site template

Responsive templates are key for today's web applications. W3 layouts are Free Responsive Mobile Website Templates that can be downloaded for free from [3]. Importing a template is done by following steps. In the application folder create an assets folder. In this assets folder is all the CSS, JS, jQuery, Bootstrap and the images which is necessary for the web page to be responsive and attractive. After that load the url helper in controllers file like this:

```
$this->load->helper("url");
```

Or in application/config/autoload.php find this array and add "url" to the array $autoload['helper'] = array('url');

In the view page link the files like this:

```
<link href="<?php echo base_url(); ?>assets/css/font-awesome.css" rel="stylesheet">
```

After that, the web pages are responsive and attractive in the PHP Framework.

B. Controller

Controllers are the heart of any web application, as they determine how HTTP requests should be handled. A Controller is simply a class file that is named in a way that can be associated with a URI [1].

Example of a controller:

```
<?php
class Blog extends CI_Controller {
    public function index()
    {
        echo 'Hello World!';
    }
}
?>
```

This file has to be saved in the application/controllers/ folder. To setup a default controller in Codeigniter open your application/config/routes.php file and set this variable: $route['default_controller'] = 'Blog';

C. Model

Models are PHP classes that are designed to work with information in your database. For example, let's say you use CodeIgniter to manage a blog. You might have a model class that contains functions to insert, update, and retrieve your blog data [1].

```
class Blogmodel extends CI_Model {
    var $title   = '';
    var $content = '';

    function __construct()
    {
        // Call the Model constructor
        parent::__construct();
    }
    function get_last_ten_entries()
    {
        $query = $this->db->get('entries', 10);
        return $query->result();
    }
}
```

D. Loading a Model

Here is an example how to load a model in the controller class:

```
class Blog_controller extends CI_Controller {
    function blog()
    {
        $this->load->model('Blog');
        $data['query']=$this->Blog->get_last_ten_entries();
        $this->load->view('blog', $data);
    }
}
```

E. Loading a View with data from model

Here is an example how to load a model in the controller class:

```
class Courses extends CI_Controller {
    public function __construct() {
        parent::__construct();
        $this->load->model('course_model');
    }
    public function index() {
        $this->load->database();
        $this->load->model('course_model');

        $data['courses']=$this->course_model->get_courses();
        $this->load->view("courses/courses", $data);
    }
}
```

Now load the data onto the view like this:

```
<html>
    <head>
        <title><?php echo $title;?></title>
    </head>
    <body>
    <div>
        <?php foreach ($'courses' as $course):?>
        <li><?php echo $course->name;?></li>
        //And so on…
        <?php endforeach;?>
    </div>
    </body>
</html>
```

F. User interface as Views

Typical web applications have pages or views in MVC. Here is an example of a view that displays details from a study program.



Figure 1 Selected view displays details about a study program

Displaying a view in MVC framework is achieved with an url helper from the url helper library like this:

```
<a href="<?php echo site_url("courses/course_details") ?>">
```

First parameter in this function is the controller and the second one is the function of the controller which sends the data to the view.

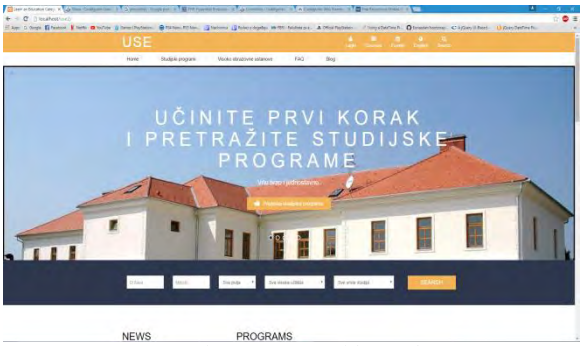


Figure 2 Home with search

For loading a default view we'll need to specify a default controller which is the Home.php controller file in controller folder. To set up a default controller go to config folder then routes.php file and change $route['default_controller'] = 'home';

The search abilities also are working with AJAX and jQuery which provides the right search results.

All of the views are responsive. What means that it can be displayed on a large variety of devices including smartphones, tablets, PC's and etc. Bootstrap allows that ability.

G. Bootstrap for responsive web sites

Bootstrap is the most popular HTML, CSS, and JS framework for developing responsive, mobile first projects on the web. Also it makes front-end web development faster and easier. It's made for folks of all skill levels, devices of all shapes, and projects of all sizes. Bootstrap ships with vanilla CSS, but its source code utilizes the two most popular CSS preprocessors, Less and Sass. Quickly get started with precompiled CSS or build on the source. With Bootstrap, you get extensive and beautiful documentation for common HTML elements, dozens of custom HTML and CSS components, and awesome jQuery plugins. [8]

H. Get content form othe pages

To get content form another page to your page, you need a url from which you want to get the content and a HTML tag where the content is stored. There is and PHP library called simple_html_dom [5] which simplifies getting the required content from other pages. This example shows how to get content from other pages:

```
get_university_course($courseLink, $contentTag){
        $raw = file_get_html($courseLink);
        return $raw->find($contentTag);
}
```

Parameter courseLink is a link taken from the browsers url field, and the content tag is basically a HTML tag like p, div etc. The second parameter is something like this **'div[id=content]'.** This is the expression for the library to find the specific element in the rendered HTML document.

III. CONLUSION

We have developed the web application USE (University Search Engine) that searches the internet for universities and study programs. The web application also has the ability to filter study programs by certain criteria which are country, city, type of study and etc. The application will be available online. Currently Croatian universities will be shown on this web application but, we look forward to expand this experience to Europa.

Students find this web application very useful, because it is efficient and doesn't waste time and it is really easy to use and find the right study program.

Currently USE is available only in Croatian language and it holds data about the Croatian universities. However, in the later stages it will feature more universities from other Countries and will be available in English also.

REFERENCES

[1] Controller. [Online]. Available: https://ellislab.com/codeigniter/user-guide/general/controllers.html. Accessed Jan. 12, 2016

[2] Codeigniter. [Online]. Available: https://www.codeigniter.com/. Accessed Jan. 12, 2016

[3] W3C Layouts. [Online]. Availeble: https://w3layouts.com/. Accessed Jan. 12, 2016

[4] PHP. [Online]. Availeble: https://secure.php.net/. Accessed Jan. 13, 2016

[5] Simple DOM class for PHP [Online]. Availeble: http://simplehtmldom.sourceforge.net/. Accessed Jan. 13, 2016

[6] Simple DOM manual. [Online]. Available: http://simplehtmldom.sourceforge.net/manual.htm. Accessed Jan. 15, 2016

[7] Postani student. [Online]. Available: https://www.postani-student.hr/Ucilista/Default.aspx. Accessed Jan. 12, 2016

[8] Bootstrap. [Online]. Available: http://getbootstrap.com/. Accessed Jan. 12, 2016

Experience with Usage of LMS Moodle not Only for the Educational Purposes at the Educational Institution

D. Paľová
Department of Applied Mathematics and Business Informatics
Faculty of Economics, Technical University of Košice, Slovakia
dana.palova@tuke.sk

Abstract – In general, definition describes the learning management system (LMS) as a software application for the administration, documentation, tracking, reporting and delivery of e-learning courses or programs. At the market with LMS can be found a wide range of LMSs from systems for managing educational records to software for distributing online or blended/hybrid courses over the Internet with features for online collaboration. The use of LMS at the educational institution can help teacher to concentrate on students' needs. The on-line access to all data needed for learning is immense benefit not only for students but for teachers too. Faculty of Economics, Technical University of Košice (FoE) use LMS Moodle with success by full-time students as supplementary educational tool and by part-time students as main system for distance learning for about 10 years. FoE staff actively participate on different kind of projects and need to share data, cooperate on design of new products via on-line tools, mostly because of its short reaction time. Familiarity with the LMS Moodle acquired through education lead to decision use LMS Moodle also for not only above mentioned educational purposes. Within the paper will be shortly presented our experience with using LMS Moodle for education and also for different non-educational purposes.

I. INTRODUCTION

The 21th century is characteristic by a growing trend towards more decentralized organizational structures, social networking. Widespread digital technologies are solid and invaluable part of the way how people living in the present can interact, work and trade. In a changing and globally competitive world the requirement for education and ongoing training has arisen [4]. At the same time, cost reduction pressure, boosting return on investment to training activities and education, as well as expanded participation on higher education, is present.

We can find multiple benefits of the digital revolution in education: easy search and acquirement of knowledge from different sources, often free for individuals; learning is no longer confined to specific classroom timetables or methods and can be personalized; establishment of new education providers; easy sharing and content development in cooperation with colleagues and learners from different countries. [5]

Currently available, and what is important to underline, open technologies allow everybody who wants to learn, to learn anywhere, anytime, via different devices, with the support of anybody. It is possible to develop new solutions for more personalised learning by up-to-date technologies. The efficiency of learning, students' involvement in education process and their motivation can be increased by stronger use of blended learning where technologies can be used for simulation of real-life situations. Through these learning experiences learners improve their technical skills and their problem-solving abilities. [9]

Within [4] authors state that educational institutions are adopting online learning and information systems at a rapid pace - 65% of higher education institutions are identifying online learning as a critical part of their long term strategy.

Today, e-Learning represent a $56.2 billion valued industry, and this number is going to be doubled in by next few years. Now 41.7% percent of global companies use some form of educational technology to instruct employees during formal learning hours. [3] As a result, there is a great interest for Learning Management Systems (LMSs) to systematically implement, improve and manage eLearning, regardless of whether it is institution providing primary and secondary education, higher education, continuing education or professional industry training programs. The continuing development of LMSs keep improving the access and tracking of learning activities as well as support organizational growth and development. Connected to present social and cloud networking trends the next generation of LMSs will be open, mobile, personal, social, flexible, and will have more tools for learning analytics. This new generation of LMSs as new education approach must be able to meet the need of the changing environment of business and education. [4]

The changes in education approach are not being fully exploited in education and training systems across Europe. About 70% of teachers in EU use Information and Communication Technologies (ICT) just for preparation of their teaching materials but not for their face-to-face work with students. The ICT is used by digitally confident and supportive teachers in education process only in case of 20 – 25% of students. [5]

From the Bruges Communiqué on enhanced European Cooperation in Vocational Education and Training for the period 2011-2020 [6] follows that European Union needs to attract new talent, equip citizens with relevant skills,

promote science and research and fuel innovation, creativity, entrepreneurship, productivity, employment and growth, in the environment of global competitive marketplace. Europe needs to provide the right policy framework and introduce innovative learning and teaching practices in schools, universities, vocational education and training (VET) and adult learning institutions. The Open Method of Coordination in Education and Training 2020 and programmes like Erasmus+, Horizon 2020 help to create the framework conditions for this to happen. [6]

II. TYPES OF MANAGEMENT SYSTEMS

A. Learning Management Systems

The Learning Management System (LMS) is a software application for the administration, documentation, tracking, reporting and delivery of electronic educational technology (also called e-learning) courses or training programs. Traditional LMS are teacher or institution centric, i.e. the educational process, course content is developed by teacher and students activities connected to course content design are limited. Current educational reform as a whole, done across EU, moves from traditional teacher centred education to learner centred and informal learning. This movement does not present the complete replacement. In practical education we can find something what is rather complementary. The LMSs can support both of mentioned education approaches, because they provide not only authoring tools, but features support the learner needs (social activities, resource and learning management, activity controlling, and personal publishing). [4]

From [18] follows that: 73% of users declared that mostly used LMS software feature is trainee testing, followed by training administration (68%); 99% of surveyed users agreed that LMS software positively impacts the organization of content and training efficiency; the top challenge for LMS users is the integration of LMS software with other systems, cited by 32% of survey respondents; considering the kind of technology used for study, 89% of respondents uses the LMS on their desktops, 76% on laptops and just 19% via smartphone; 90% of survey respondents plan to maintain or increase their investment in LMS software over the next year.

Generally said, the LMS represents mostly web-based framework handling all aspects of the learning process – delivery and management of content within the courses, identification of education process objectives and goals, tracking the progress, reporting of data from education process. [8], [13]

The common features of Learning Management Systems can be summarized as [4], [17]:

- *Content authoring/resource management* gives the ability to users to design their own content and deliver courses within a LMS. Authoring and sharing tools are often part of an LMS.
- Ability of *control the user's activities.* The LMS provides set of tools and rules for setting permissions of access, setting up activities, determining sharing, initiation of study group etc.
- *Proficiency testing and reporting* helps administer tests to measure employee/student knowledge or skill. Analytics and reporting functionality can identify learning gaps.
- *Learning management* tools helps to organize and simplify training or learning administration, including distribution of content, management of user information, schedule and oversee course enrolment. Learners can set learning objectives, monitor and evaluate their learning process and achieved knowledge.
- *Course and personal library* provides a pre-made library of training courses for generic training needs, such as those on sexual harassment policies or management techniques. At the same time, LMS provides place for user's creating and publishing various kind of content via blog, wiki, forums.
- *Certification and compliance management* – including setting up, tracking and managing certification programs for industries that require employee certification.
- *Virtual classroom* via functionalities like video conferencing, live course/lesson leading, remote classes through the platform.
- *Extended enterprise* means that LMS allows organizations to train or teach external users. For that purposes the E-commerce functionality may be included.

B. Content Management Systems

The Content Management System (CMS) or software is used to organize content comprising a website. [11] It represents an application providing method for managing content from a central location. Based on user's permissions is it possible to edit, add, and view content, or just only have preview of the content. These systems are used by companies because of very easy managing versions of a document and centralizing all aspects of something in one place. These systems are managing manage various content forms, including files, images, electronic documents, audio files, and many more. Content can be easily tagged using metadata, which is the best for searching and using content quickly and efficiently. [12]

C. LMS vs.CMS

As was described in previous subchapters, both of mentioned systems is used for content management. And both have their special place within organisation. From the point of view of possibility to provide eLearning, both can be used, in principle. But the difference is significant and results from their core functionality. CMS is more passive and students are allowed just to view the shared documents. From previous facts follows, that for learners it is important to have available functionality for keep all types of eLearning content and courses, as well as tracking who has completed them, when, and what kind of

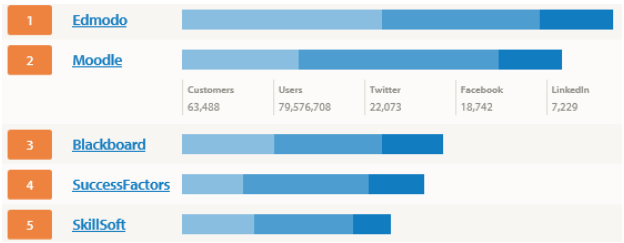
score received on assessments. These functions are provided by every existing LMS, so students have the ability to view, listen, and interact with the data. They have the options to try examples, complete assessments, and give evaluation or feedback on courses. [12]

III. USE OF LMS MOODLE AT THE FACULTY OF ECONOMICS

The Faculty of Economics (FoE) was established on July 8th 1992 within the framework of transformation of the Technical University to a "full-fledged" university covering also non-technical disciplines. Mission of the Faculty of Economics is to develop knowledge and provide high-quality, professional education in the field of economic sciences and to contribute to the development of society and shaping of the regional economy through the provision of research and consultancy activity, continuing education and services for the local and regional communities [9].

The changes in the world of education (described for example in [4], [6]) and widening the implementation of ICT and eLearning methods into education process [10], [14] brought faculty top management to decision to introduce ICT and useful methodology to all types of the education provided by the faculty (bachelor and master study programs and continuing education courses). This decision was supported by achieved staff skills and experiences by participating at many international projects (for example SIGOLD, SAGE+, FASTER, OpenInn) and realisation of internal project "*Preparation and realisation of the distance learning bachelor study in the field of Finance, Banking and Investment*". In 2003, we started with eLearning courses development using LMS uLern, which we found as not really user friendly (especially for those who have no previous experience with ICT). In 2004 the uLern support and development was suspended and we need to change the LMS. Based on our previous negative experience, we did serious selection of the best LMS fitting out needs [21]. Finally we chose LMS Moodle (www.moodle.org) due to requirements of today's web development (portal structured webs) and easy navigation.

The well-known LMS Moodle was founded at 2002 in Australia and for a long time belongs to the top 5 LMS (Figure 1.) wide world [22]. At the moment it has 81.4 million active users, while at Dec.2015 it was about 79.9 million [15]. Based on research done by G2 Crowd [2], [7] the LMS Moodle is member of "high performers" group at the market of LMSs. It provides products that are highly rated by their users.



Ranking and bar segment lengths are determined using a weighted scoring system that awards points for total metrics and rank within a data set.

Figure 1. Top LMS Softwares (Dec. 2015), source: [2]

A. *LMS for educational purposes*

As was mentioned in previous chapters, the LMS is great option for teachers, entrepreneurs or subject matter experts who have no knowledge of coding and design. [1] The LMS Moodle is one of the open source options, which don't necessarily work like "out of the box" products. It offers great LMS skeleton, but administrator needs to set it up first. It needs to be configured to meet user's design and technological requirements. But then it is an excellent tool for eLearning providing, especially for those, who are not very well skilled in ICT use – doesn't matter whether it is teacher or student.

At 2004 FoE started to use LMS Moodle as the supporting environment for teaching subjects Informatics and Spreadsheets, first. From 2005, after achieving the positive feedback from students and also from teachers, the management of the FoE decided to implement LMS Moodle to support educational process of the other subjects at the faculty (e.g. Microeconomics, Macroeconomics, Economic Analysis, Risk and Uncertainty, Counting, Economic Information System…) and from 2005 we already use the system also for providing whole distance learning education. It means that we have two education installation of LMS Moodle at the faculty, at the moment. Based on internal statistics done in January 2016, there are 325 courses created (195 courses for full-time students, 130 courses for part-time students) and our LMSs Moodle are used by 1271 active users (and by 3667 users in total). The number in TABLE I. is presenting a disproportion between registered users and active users. It is caused by education system principle: students usually stop use the system after their graduation. That is why there is included also parameter "active users".

TABLE I. FOE MOODLE STATISTICS

	Moodle installation for			
	full-time students		**part-time students**	
	Total	**Active users (last access latest Sept.2015)**	**Total**	**Active users (last access latest Sept.2015)**
Users	2376	998	1291	273
Students	2155	906	1235	227
Teachers/ faculty staff	221	92	56	46

In Figure 2. I would like to point out the interesting comparison of numbers of newly created courses during the years of existence of the LMS. It is copying all the changes of the educational system at the FoE. The LMS is used by every kind of provided study form: full-time and part-time (currently being realized by blended learning methods). As was mentioned above, the LMS Moodle was introduced to the faculty's educational environment in 2004. Under both installations there was created one course: Course Informatics I for full-time student's installation and Introductory course for teachers for part-time student's installation.

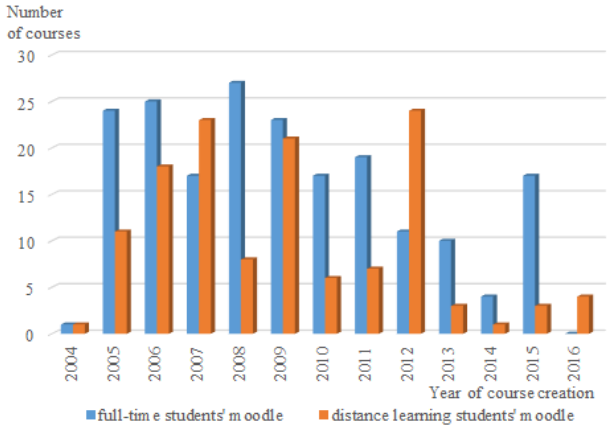


Figure 2. Comparison of the number newly created Moodle courses within timeline

Next academic year (2005) we started by full support of part-time students education in their first year of their study via LMS Moodle. As can be seen from the chart above, next five years represent the growth of new courses, what was caused by continuing development of new distance learning courses for bachelor and master study programmes. During the years 2010 – 2012 we again needed to create new courses for newly developed study program (Economy and Management of Public Administration). In following years, as can be seen, number of new courses increased by creation of the courses for optional subjects. By this activity we tried to make the part-time study more attractive. The provision of part-time study via LMS Moodle is defined as internal faculty project and this process is continuously evaluated every year by faculty management. Because of restriction set for length of contribution cannot be this process and its management described in more details.

The situation in installation of full-time students' LMS Moodle is little bit different – teachers are not obliged to use for their teaching LMS Moodle. But as is visible on the graph, every year we created approx. 20 - 25 new courses based on requirements from the teachers (most of them were inspired by students' demand). The year 2015 was special, because of implementation of new accreditation document and its implementation to education process at the faculty and there was a need to create new courses for new subjects provided by our faculty. That was the reason of design of 17 new courses.

B. LMS for non educational purposes

Question is, whether LMS could be used for CMS purposes. Next chapter in short shows our decision to use LMS Moodle, which is default set of tools used for eLearning purposes, for creating and handling different kind of documents and files.

For the decision making purposes, faculty management and IT department, did serious analysis and comparison of standard functionalities of CMS and LMS first. The results of the comparison, based on own research, information published in [4], [16], [20] and our needs analysis, are shown in TABLE II. For the final decision there was prepared also SWOT analysis taking into account certain factors such as skills and abilities of the faculty staff, investments needed, features availability of the systems, possibilities of integration with other systems.

TABLE II. COMPARISON OF CMS AND LMS FUNCIONALITY

Criteria/Fact/Functionality	**LMS**	**CMS**
Controled target	training control	version control
Manages	content, learners and their activity	content
Avoid need for hand coding	+	+
Easy editing via WYSIWYG editor	+	+
Scalable feature set	+	+
Possibility to send e-mail/ message via environment	+	+
User management		
User registration	+	+
Store users profile	+	-
Authentication via external database	+	+
Users group management	+	+
Access control	+	+
Content management		
Store users data	+	+
Share users data	+	-
Content creation	+	+
Templates	-	+
Organize reusable content	-	+
Workflow tools	+	+
Processes and procedures	+	+
Editing, format management	+	+
Modifying	+	+
Organizing	+	+
Search	+	+
Version controlling	-	+
Deleting	+	+
Maintainig	+	+
Available activities		
Users collaboration	+/- (via chats, forums, wikis, …)	+
On-line uploading data	+	+
Event scheduling	+	-
Competency mapping	+	+
Content publishing	+	+
Type of supported content		
Text	+	+
Graphics	+	+
Photos	+	+
Video	+	+
Audio	+	+
Code	-	+
News	+	+
Forums	+	-

NOTE: In case of LMS the user is teacher or learner, but it not can be used for description of CMS, that is why we used just term user;
+/- means that functionality is present but with some exceptions and restrictions

One of most important positive factors was that using LMS Moodle for teaching purposes made teachers more ICT skilled and opened them new horizons in possibilities of cooperation and data sharing using various ICT products. The staff of the faculty regularly participates on different kind of projects and struggles with the problems of design, sharing and publishing the non-public outcomes. Most of projects have partnership made up by not ICT skilled, but highly specialised people (sociologists, psychologists, foreign language teachers, etc.). For them seems classical CMS or cloud solution

very complicated, but they are already familiar with environment of LMS Moodle via their education (as was mentioned above). And the situation is same with our partners from other education institutions, where LMS Moodle is wide-spread. That was the reason why IT department decided to propose to project participants to test LMS Moodle instead of other ICT products for project's documentation sharing and collaboration among the partnership. It should help to save time and finance needed for education faculty staff to use any other system.

Over the years we have created and used 38 courses for the 186 people originating from 18 countries (for example: AT, BE, BG, CY, CZ, DE, DK, ES, SE, GR, PT, etc.) and participating on national, international projects (SIGOLD, FASTER, CRENAR, SAGE+, OPENINN, CLUSTERING, etc.).

Over the time, we found another advantage: many of projects with existing courses at this installation of LMS Moodle needed to provide also some pilot eLearning course. So it was very easy to create new course for this purpose and do pilot testing without installation new instance of the system and new registration to another one system.

The positive feedback for the application LMS for project data sharing and collaboration instead of CMS inspired top management of the faculty also as intranet solution for FoE. At present is ongoing the pilot testing, where the structure of courses in LMS Moodle was improved to needs of the faculty documentation and structure. After one year of pilot testing the results will be summarised and evaluated.

C. LMS as tool for cooperation between educational institution and companies – project SP4CE

The project SP4CE addresses the aims and needs identified in Bruges Communiqué on enhanced European cooperation in vocational education and training [6], especially that one "*improving the quality and efficiency of vet and enhancing its attractiveness and relevance*" and "*enhancing creativity, innovation and entrepreneurship*". Project activities are planned to connect with career-oriented continuing voluntary education and training (C-VET) principles. All project's results and project's actions will be concerned to promote take-up of innovative practices in education, training by supporting personalised learning approaches, collaborative learning and critical thinking, strategic use of ICT, Open Educational Resources (OER), open and flexible learning, virtual mobility and other innovative learning methods. [19] The expected project outcomes are *SP4CE learning portal*, *SP4CE pedagogical concept, Guideline for consultants* (enterprises), *Guideline for mentors* (schools teachers) and *Guideline for general public* „How to use SP4CE portal". The project has three target groups [19]:

- students from vocational schools and universities, entering in the near future the labour market,
- HR staff of the enterprises from companies looking for new workers especially those just entering the labour market,
- teachers (mentors) from different education institutions.

All three target groups will have access to multi-user platform SP4CE portal (Figure 3.) containing training materials about coaching students and tools for establishing direct contact with students interested in cooperation with company. Using this system, mentors will help students to establish successful cooperation with coaches from companies.



Figure 3. Illustration of SP4CE portal and SP4CE LMS Moodle installation

The SP4CE Portal will provide the space for:

- problems to be solved, questions to be answered,
- development of the teams which want to work towards the problem solution,
- work for teams or individuals to develop the solution,
- mentoring & coaching,
- presentation of developed solutions,
- and publishing of the chosen solution. [19]

Besides the expected positive impact on employment of young people in industrial companies, designed portal proposes other benefits, for example: repository of the accepted and implemented solutions to the various problems accessible to general public; support of international cooperation towards the solutions to various problems; encouraging young people to work as designers, inventors, and creators; use of new learning and cooperation methods implemented into daily activities of enterprises; promoting take-up of innovative practices in education by supporting personalised learning approaches, collaborative learning and critical thinking; enhancing the professional development of those working or active in the education by innovating and increasing the quality and range of initial and continuing training, including new pedagogies, ICT-based methodologies.

The project SP4CE started on Sept.2014 and will end on August 2017. There were already done following activities: design of SP4CE pedagogical concept and pilot testing of innovation eLearning rooms by each partner. Pilot testing was done in LMS Moodle environment with the aim to check the relevance of particular sources and activities for the SP4CE portal objectives. The pilot tests evaluation is carried out at present time.

IV. CONCLUSION

ICT presents a great help not only in education institutions, but impacts all parts of our lives. Different kinds of IT solutions make our lives easier manageable and available from different places at different time.

A CMS is basically designed to manage content, for example documentation of an enterprise. In the past it was also used to support educative or academic courses. It allows the instructor to create a course website, where documents can be uploaded in popular formats such as MS Word, MS PowerPoint, pictures, etc. without converting them to a HTML format. But from the point of view of tracking and reporting education process it has more weaknesses than benefit. This is the reason the LMS were designed. Modern LMS is a set of functionality supported by traditional content management and learning management, but also includes more learner focused personal, social, and collaboration features.

The information published in the paper has no ambition to act in favour of the use of LMS for whatever, but I'd like to present our reasons and experiences with using LMS Moodle for the learning purposes, but from the 2005 also in different way – just as traditional CMS. More traditional process is integration CMS into institution and afterwards transition to LMS (in case of education process), but as is described on our example it is possible also vice versa. In addition the different ways of usability of LMS is widened by research within project SP4CE, which uses LMS as tool supporting the collaboration between educational institutions and enterprises. It is obvious, that the most ideal case is to use specialised information systems for the purpose they are designed, but in our case we tried to find win-win solution, where the highest request was to not stress faculty staff by using another new system. The LMS Moodle as CMS is currently under pilot testing and we collect users' feedback for evaluation this modernisation step at the faculty.

ACKNOWLEDGMENT

This work has been kindly funded by the European Commission under the ERASMUS+ Programme Project SP4CE: Strategic Partnership for Creativity and Entrepreneurship (No. 2014-1-PL01-KA200-003341). This publication reflects the views only of the authors, and the National Agency and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

REFERENCES

[1] Academy of mine, "Learning Management System Comparison: LMS Features Chart", [online], Available at <https://www.academyofmine.com/learning-management-system-comparison/>, [Accessed 30. January 2016].

[2] Capterra, "Top LMS Software", [online], Available at <http://www.capterra.com/learning-management-system-software/#infographic>, [Accessed 25. January 2016].

[3] Ch. Pappas, "The Best Learning Management Systems", [online], Available at <http://elearningindustry.com/the-20-best-learning-management-systems>, [Accessed 5. February 2016].

[4] D. E. Stone, Z. Guangzhi, "Learning Management Systems in a Changing Environment", [online], Available at <https://www.researchgate.net/publication/267625754_Learning_Management_Systems_in_a_Changing_Environment>, [Accessed 10. January 2016].

[5] European Commission DG Communications Networks, Content & Technology, "Survey of Schools: ICT in Education", [online], Available at <http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=1800>, [Accessed 5. February 2016].

[6] European Commission, "Supporting vocational education and training in Europe: the Bruges Communiqué", [online], Available at <http://ec.europa.eu/education/library/publications/2011/bruges_en.pdf>, [Accessed 15. January 2016].

[7] G2 Crowd, "Best Software review", [online], Available at <https://www.g2crowd.com/>, [Accessed 3. February 2016]

[8] K. Gilhooly, "Making e-learning effective". Computerworld 35 ed., vol.29, 2001, pp. 52–53.

[9] L. Révészová, "Informatics education according to the latest trends and ICT development". In: eXclusive e-Journal, vol. 1 , 2015, pp. 1-8, ISSN 1339-4509.

[10] L. Révészová, J. Bucko, P. Mihók, "Modelling of e-services". In: Ambient Intelligence Perspectives, Vol. 1, 2009, pp. 203-210, ISSN 1875-4163.

[11] Litmos at CallidusCloud, "Learn more about content management systems (CMS) vs learning management systems (LMS)", [online], Available at <http://www.litmos.com/training-platform/content-management-system-vs-learning-management-system>, [Accessed 25. January 2016].

[12] M. Dubowy, "LMS vs LCMS vs CMS...changing one letter makes a big difference", [online], Available at <https://www.opensesame.com/blog/lms-vs-lcms-vs-cmschanging-one-letter-makes-big-difference>, [Accessed 3. February 2016].

[13] M. Szabo, K. Flesher, "CMI Theory and Practice: Historical Roots of Learning Management Systems". In Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2002 (White Paper), pp. 929–936. ISBN 1-880094-46-0.

[14] M. Vejačka, "Introduction to electronic banking by elearning methods", In jubilee 35th International convention MIPRO 2012, 21 – 25. May 2012, Opatija, Croatia, pp. 1406-1411, ISBN 978-953-233-069-4.

[15] Moodle.org, "Statistics data", [online], Available at <www.moodle.net/stats>, [Accessed 3. February 2016].

[16] R. Ellis, "Field Guide to Learning Management Systems. Learning Circuits - American Society for Training & Development (ASTD), [online], Available at <http://www.astd.org/NR/rdonlyres/12ECDB993B91-403E-9B15-7E597444645D/23395/LMS_fieldguide_20091.pdf>, [Accessed 10. December 2015].

[17] Software advice Inc., "Compare LMS Software", [online], Available at <http://www.softwareadvice.com/lms/?deployment_id=&market_products_sort_order=&market_products_sortby=nb_reviews&more=true&price_ranges=1&stars=3&segment_id=1246&platforms=&int_site_code=&layout=var_c0&size_id>, [Accessed 3. February 2016].

[18] Software advice Inc., Learning Management Systems UserView/2015", [online], Available at <http://www.software-advice.com/hr/userview/lms-report-2015/>, [Accessed 3. February 2016].

[19] SP4CE consortium, official web page of the project, [online], Available at <http://www.sp4ce.piap.pl/index.php>, [Accessed 27. January 2016].

[20] T. A. Dousay, "LMS vs CMS", [online], Available at <http://www.slideshare.net/ebloomsie/lms-vs-cms-a-summary>, [Accessed 3. February 2016].

[21] T. Urbanová, D. Šiškovičová, "Commercial LMS versus open sources LMS", In Generic issues of knowledge technologies, HUBUSKA proceedings, Sept. 14th 2005, Budapest, Hungary, ISBN 954-91700-2-0, pp.: 109-119.

[22] W. Fenton, "The Best Learning Management Systems (LMS) for 2015", [online], Available at <http://www.pcmag.com/article2/0,2817,2488347,00.asp>, [Accessed 25. January 2016].

Using Robot Simulation Applications at the University – Experiences with the KUKA Sim

D.Lukac*
* Rheinische Fachhochschule Köln gGmbH,
University of Applied Sciences, Köln, Germany
dusko.lukac@rfh-koeln.de

Abstract - **In order to work with robots and in order to test differed application conditions, are in the meanwhile, real robot systems not necessary any more. At the current market, all name-like robot manufacturers offer suitable simulation software. This concerns simulations referred to the industrial robot systems as well as to the mobile robot systems. In this work important robot-simulation applications are introduced. A focus is directed upon the simulation software KUKA Sim and its suitability for the education practice at Rheinische Fachhochschule Köln gGmbh, University of Applied Sciences in Germany. Based on experiences with the students it is discussed, how such an application can be used in the daily educational tasks.**

Keywords — KUKA Sim Pro, Industrial Robot Systems, Simulation, Robotics

I. INTRODUCTION

Production and process innovations based on the robot systems, require in view of the competition of the companies, an exact knowledge of the program, communication and movement possibilities and abilities of the robot system in adaptable workrooms. The essential demand for secure industrial robot planning, can be still well realized by prior three-dimensional graphics simulations and feasibility study of the planned robot application. With the simulation of motion sequences and peripheral objects the possibility exists, without cost intensive hardware-robot solution, to plan the assembly and manufacturing process, but also to predict the needed cycle time, modifying and optimize it during visualization process. Today is, beside a wider applicability of an open simulation program, in particular the high user comfort with easy learnability one of the main judgement criteria about the quality of the software. It is beside other criteria, a relevant and decisive factor for evaluation of the goodness of the used robot simulation program. It means in the practice, that not only special user groups like engineers should be able to use such simulation programs, but also all user with far less technical knowledge, meaning also those with only basic technical knowledge. On the subject of the robot simulations, it was already written since middle of the 80-s [1], [2]. Accordingly various research teams have been assigned with the task, to develop suitable user-friendly robot simulation frameworks [3]. The first useful simulations as e.g. so called “Robocad” application, have been actively used in the technical departments since the beginning of the 90-s [4] [5]. In the middle and at the end of the 90-s, object-oriented robot simulation applications based on Java were developed and used [6]. In addition and consequently, suitable robot libraries were developed for simulation, also of older robot models served with it [7]. Besides, applications were developed "to place the emphasis on using the knowledge of the geometry of the parts and machines in a production cell to automate those aspects of the traditional design method that can be performed by the computer" [8, p.21]. Following section gives a short view at the industrial robotics market and the criteria for choosing appropriate robot simulation application.

II. APROPRIATE ROBOT SIMULATION APPLICATIONS FOR EDUCATION PURPOSES

At the market one can find different robot sim simulation systems. To find right one for the use at university, requires the knowledge of the technical systems and also the knowledge about the industrial robot market. It is an interest of the university to educate students in that way that they become prepared for requirements of the industrial market. Therefore it is understandable that such robot systems and simulation applications should be used at the university, which are mostly implemented in the industrial companies. Due to those criteria and the vast amount of today's robot simulation applications today, in the following are briefly introduced four simulation applications, which are shortly mutually evaluated. They are presented in the following table:

TABLE I. INDUSTRIAL SIMULATION SYSTEMS FOR INDUSTRIAL ROBOTS

Name of the application	Producer
RoboGuide	FANUC K.K.
RobotStudio	ABB AG
KUKA Sim	KUKA AG
CODESYS	3S-Smart Software Solutions GmbH

These above named simulation applications of the companies FANUC K.K., ABB AG, KUKA AG and 3S-Smart-Smart Software Solutions GmbH are often used in

the industrial applications. Furthermore it is to be pointed out that intermutually translation applications exist, as e.g. "SubitoRun", which enable translation of the application of the one company into other, here of FANUC GUI into KUKA GUI. Such translation modules exist also in the case of KUKA C4 interfaces which are translated into Siemens S7 GUI. Even translation of Siemens TIA portal exist, which is offered by "FTP Robotik" [9]. It indicates the necessity of the development of such specified translation-software market. The Codesys (CODESYS, CoDeSys) is an application development system for programmable logic controller (SPS), developed according to IEC 61131-3 standard for the application development in the industrial automation. All above mentioned applications allow simulation of robot systems, with the exception of the CODESYS software application, which offers basically a simple simulation module connected with the internal PLC kernel and it is not specified application on robot systems. For simulations of other kinds of robot systems as e.g. humanoid running robots, mobile robots or other movable systems, it is pointed out to applications specialized for such applications like Microsoft Robotics Developer Studio, Anykode, SimplyCube, V-REP or Eureka, which are mostly offered as a freeware. Four stated robot simulation systems which illustrate the market portion of the industrial robot manufacturers at the European market were selected from the fullness of the industrial robot simulation applications. According to topical studies [10] the following robot manufacturers are market leaders at the European market in the segment automotive industry: Fanuc, Yaskawa, ABB, Nachi, Comau, Reis. The company 3S-Smart-Smart Software Solutions GmbH is not a robot manufacturer, but for it, this company is permanently stronger representing supplier of universal PLC and simulation application solutions and becomes a counterpart at the German market to the big companies like Siemens. Alike it looks at the Chinese market. According to statement of International Federation of Robotics (IFR) Statistical Department „The robot market in China is topically still by foreign robots of 4 most important manufacturers ABB, Fanuc, KUKA and Yaskawa dominated"[11]. Hence, it makes sense, on account of the industrial requirement and inquiry, to look at robot simulations of these stated manufacturers for the use in the educational purposes. Yaskawa application was took out of consideration with the process of choosing at Rheinische Fachhochschule Köln gGmbh (RFH), because of the less spread at the European and especially German market. With the choice of software - as an essential subrange of the software procurement - for an economic sphere, the following demands / criteria can be laid in support of the study [12]:

(1) Functional demands, as e.g. functionality

(2) IT-technical demands, as e.g. performance or IT-liability

(3) Demands to the architecture of the software, as e.g. Software Multitenancy

(4) Other criteria, as e.g. price, reputation, support, technical quality of simulation result

Before of any purchase of the software, comparative study according to defined criteria has to be made. The basic criteria stated, can been enhanced with the specific demands for educational purposes, especially by students' and lecturers' view an on-site circumstances. Furthermore the choice of the robot simulation software can strongly be impacted by the already used hardware, in this case in the laboratories of RFH. It means that robot simulation software has to be compatible with already installed and used robot systems. Therefore following two additional criteria can been taken into consideration by choosing the appropriate robot simulation software system.

5. Trial version of the software for min. use of 30 days is available
6. Hardware demands has been taking into consideration

III. BENEFITS AND WEAKNESSES OF ROBOT SIMULATION APPLICATIONS

Compared with other technologies, also simulation technology has its benefits and drawbacks. Using modelling tools to design robot systems is an easy and low-priced technique to shape complex robot applications, at least virtual robot systems of different kind, as humanoid, mobile or industrial robot systems. The main pros and cons by using of the robot simulations software are listed and shortly explained below:

A. Benefits for user, by using of robot simulation software

(1) Lower costs in the case to produce a robot system from scratch

(2) In many cases, sometimes indirectly, it is possible to test programming code according to specifications

(3) Possibility to modify design of robot cell without additional costs

(4) In some cases, some robot parts and peripheral components can be tested and simulated

(5) In a multifaceted project the robot systems can be replicated in stages

(6) A whole simulation can define if the robot solution meet the wanted conditions

(7) Nearly all simulations software are well-matched with a extensive variety of programming languages and interfaces ,

(8) Elapsed period stuck between the start of the project and its finishing point can be reduced

B. Weaknesses for user, by using of robot simulation software

(1) The real-world robot systems may focus the robot at too various more situations than a virtual-world robot systems

(2) All simulation, also the robot simulation programs, simulate what are automated to simulate with a specific level of accuracy

Summarizing the most important benefits for the practical decision to buy simulations software, the most included argument for acquisition of simulation software is the lowering of the general costs of the robot system, if preplanning phase of robot supported production plant development is properly done. While all of the simulation utensils provide the option to act out the robot system in diverse scenarios, the programming code can also be tested to conclude the compatibility with the mandatory conditions, and many more structures. Surrounded by other disadvantages, as a main weakness of robot simulations software are regarded the programming problems that can't cover all the scenarios that may occur in the real world. In the most cases 1:1 transmission of the simulation into the physical robot system is not possible. Also, tests of the simulations have to be done in any case.

IV. UNIVERSITY EDUCATION THROUGH USING ROBOT VISUALISATION SOFTWARE WITH HARDWARE

Analysing the appropriate literature, it is evidential that combining the simulations software application with practical exercises with hardware brings the best educational results [13]. In the case of robot systems is in the normal way both, hardware and software very expensive and not in any case applicable at university. Also, with hardware and software only, the appropriate education is not directly possibly, because the instructor or lecturer have to be qualified and educated to use the simulation software in combination with the hardware (robot system). By merely reading the instructing guides is, in normal case not possible, appropriately to use such complex systems in the educational institutions.



Figure 1. Components of KUKA Education Bundle according to [11, p.2]

It means that, behind the acquisition and asset costs also training costs have to be provided by the institution, in order to prepare qualified stuff to proper transfer the necessary knowledge to the students. Summarizing it, it can be stated, that in the case of appropriate education in industrial robotics many obstacles exists. At the RFH such financial means have been provided, so that up-to-date robot system KUKA KRC 4 with fitting simulation software, including seminars for programming of the robots as well as for using simulations software has been purchased (cf. Figure 1.). The system purchased is provided by KUKA Roboter AG Company, the worldwide market leader in automotive industry. It is so called "KUKA Education Bundle" which according to [14, p.3] "…you enable students to acquire highly valuable skills at an early stage. It is the ideal addition to theoretical tuition and allows practical robotics training, identical to real applications, in accordance with globally recognized industry standards." In order to provide students-close effective education, at RFH has been developed mobile robot system presented in the following figure (cf. Figure 2.). It is very good and very explaining way to explain complex function of robot system, including programming and visualisation functions.



Figure 2. Industrial robot mobile system at RFH

First experiences with this system show very good acceptance by the students including very good learning results. Not only practical work with the robot hardware, mainlining programming of the robot system is possible. Also all boundary aspects in industrial robotics education can be covered with this solution. For example presented robot system works with modern, well-tried bus technology as e.g. EtherCAT [15] which can be additionally, as e.g. student project work, including into

the robotic subject module at university. Possible solution of system connection of the EtherCAT, made with the additional free-ware offered by KUKA company, called WorkVisual, which also offers offline programming possibilities, is presented in following figure (cf. Figure 3.).

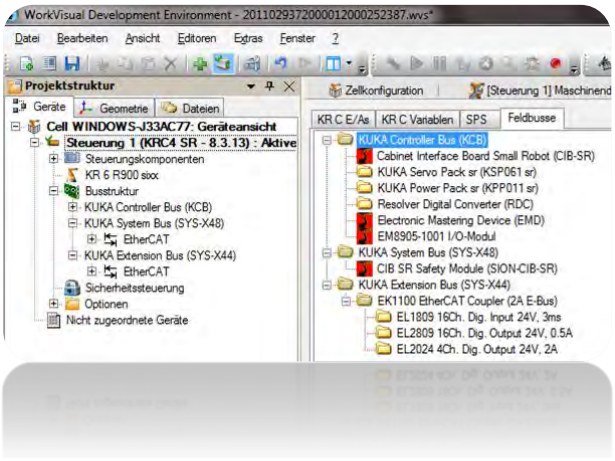


Figure 3. Configuring the KUKA KRC 4 EtherCAT interface

In order to exercise programming and simulation skills, students work with the simulation software KUKA Sim Pro combining with so called KUKA.OfficeLite application. An example of such simulation made in the laboratory of RFH is presented in the following figure (cf. Figure 4.).

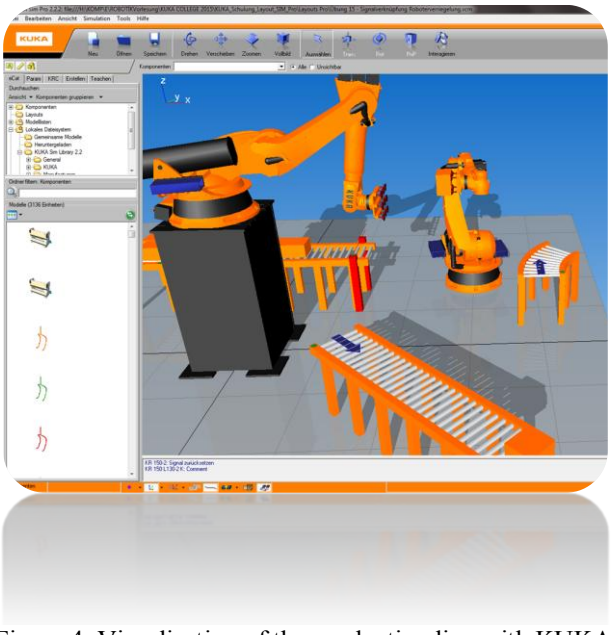


Figure 4. Visualisation of the production line with KUKA SIM Pro

With KUKA.OfficeLite for KUKA robots programmes can be provided and be optimised on every PC off-line. KUKA.OfficeLite is nearly identical with the system KUKA software at the KR C4 control unit. By the use of the original KUKA SmartHMI and the KRL linguistic syntax, the off-line service and programming corresponds precisely to that of the robot. KUKA.OfficeLite is delivered as a preloaded virtual image and can be used with a Vmware software. Not only visualization of the production line and merely programming can be presented, but also the kinematics of the robot, as a typical part of the robotic movement theory. So, it is possible to build own components in the visualization system or to analyze e.g. composition and structure of the industrial robot. Following figure shows one industrial robot as a so called kinematic chain (cf. Figure 5.).

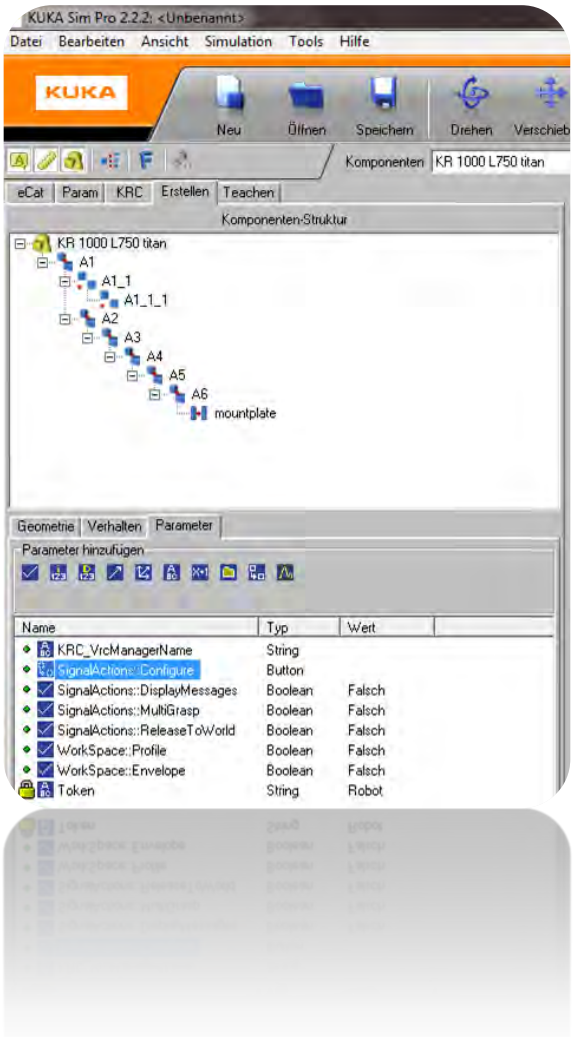


Figure 5. Robot component as kinematic chain

Quite often it is necessary to have corporate identity related components in the simulation program or components with special properties which are company-specific. As a part of the educational concept and as one possible student project simulations of so called KUKA SIM pro components can be made (cf. Figure 6.). One such exercise is integrated into the simulation exercises offered to students at RFH. Students may, on that way, learn to adapt software application to real circumstances and with it, to work on projects which serve to solve a real problem. Furthermore, the whole robotics laboratory can be simulated as a real case and even handling process in the laboratory itself. The virtual robot controller can be

connected to the simulation program KUKA.Sim Pro, and on this way, it can allow the KUKA robot to be programmed directly in KRL (KUKA Robot Language), without having to intervene directly in the process sequence. Following figure shows simple "Rheinische Fachhochschule Köln"- visualization component made in KUKA SIM Pro in the laboratory of RFH (cf. Figure 6.).

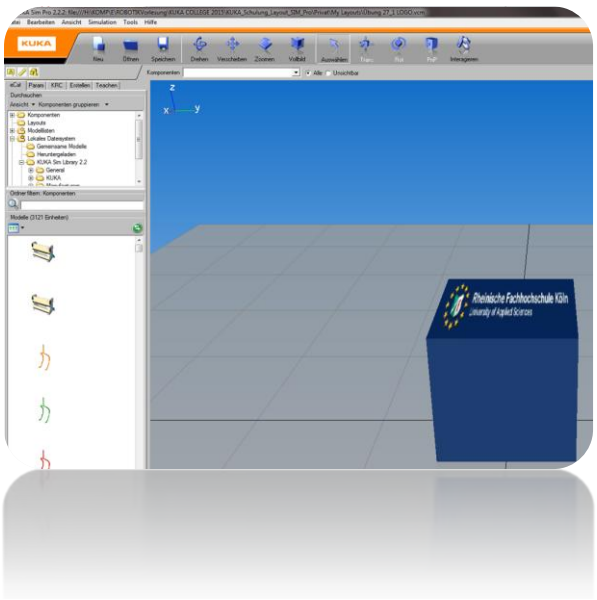


Figure 6. Component design with KUKA Sim Pro

V. CONCLUSION

Planning the proper use of practical and theoretical exercise with robot simulation software and robot system hardware enables effective and efficient education in the field of industrial robotics. It is necessary that in the case of tertiary education, lectures becomes not only practical, but with lot of practical explanations which highlights the theoretical backgrounds. Very important in the organizational educational sense is forming of the smaller student project groups, in order to shape frame for suitable transfer of the knowledge by using software simulations and real hardware systems. This paper indicates that at many universities such an efficient work is not possible, because due to lacking of financial, organizational and human assets. Collaboration between universities at European level, may be a way to define models of sharing the laboratories as a part of European educational project.

REFERENCES

[1] Metternich, J. (1987), Mathematische Algorithmen und dreidimensionale Robotersimulation mit HIKAWA-Robotern (Mathematical algorithms and three-dimensional robot simulation with HIKAWA robots), Diplomarbeit, FH Koblenz

[2] Degen, H.-J., Fraunholz, L., Klein, J (1992), CBT-gestützte Demo-Version von RobotoCAM (CBT-supported demo version of RobotoCAM.), BIS Allgemeine Elektrotechnik 1992

[3] Wloka, DW (1991), Robotersimulation, (Robot simulation), Springer-Verlag Berlin Heidelberg, ISBN: 978-3-540-53828-8

[4] Breitenbach, K. (1991) "Simulation und Offline-Programmierung von Industrierobotern mit ROBOCAD" (Simulation and Offline-Programing of industrial robots with ROBOCAD) in Wloka,DW (1991) „Robotersimulation", pp.65-88

[5] Laloni, C. , Plank, G., Rieseler, H. and Wahl F. (1991), PC-basierte Modellierung und Simulation zur Off-line-Programmierung von Industrierobotern, (Modelling based on PC and simulation to the out of vision line programming of robots), Affiliated with Institut für Robotik und Prozeßinformatik, Technische Universität Braunschweig,Springer-Verlag Berlin Heidelberg,pp 121-152

[6] Rauh,O. (2000) "Objektorientierte Programmierung in JAVA", Fallstudie: Robotersimulation, (Programming oriented to object in JAVA", case study: Robot simulation), Vieweg+Teubner Verlag|Springer, pp. 107-120

[7] Tewkesbury, G. and Sanders,D. (1999) "A new robot command library which includes simulation", Industrial Robot: An International Journal, Vol. 26 Iss: 1, pp.39 – 48

[8] Owens, J. (1996) "Task planning in robot simulation", Industrial Robot: An International Journal, Vol. 23 Iss: 5, pp.21 – 24

[9] SPS Magazin (2015), "Wenn der Roboter mit der Werkzeugmaschine...", (If the robot with the tool machine..) available at <http://www.sps-magazin.de/?inc=artikel/article_show&nr=101125>, 12.01.2016, 12:15h, SPS-MAGAZIN 7 2015 date 09.07.2015

[10] KUKA AG - Company Presentation (2013) "Leading positions in Robotics and Systems", Robotics - Automotive, March/ April 2013, http://www.kuka-ag.de/res/ag/company_presentation/2013/company_presentation_en_20130326.pdf, 31.01.2016, 21:33h

[11] International Federation of Robotics (IFR) Statistical Department (2015) „CEO Kommentare zur Veröffentlichung von World Robotics 2015 30. September 2015" (Comments of CEOs to publication of World Robotics 2015 from 30. September 2015) , available at <http://www.worldrobotics.org/uploads/tx_zeifr/CEO_Kommentare_World_Robotics_2015_deutsch.pdf>, 31.01.2016, 21:42h

[12] Universität Münster (2012), Kriterien der Softwarebeschaffung – Arbeitsbericht (Criteria for purchasing of software, Working paper), available at <http://www.wi.uni-muenster.de/institut/arbeitsberichte/ab62.pdf>31.01.2016, 22:23h

[13] Chen,Li-Chiou and Lin,Chienting (2007) "Combining Theory with Practice in Information Security Education", Proceedings of the 11th Colloquium for Information Systems Security Education Boston University, Boston, MA June 4-7, 2007, pp.1-8

[14] KUKA AG (2016), KUKA Education Bundle: Preparing students for the markets of the future, available at < http://www.kuka-robotics.com/res/sps/a737ee03-5832-4c95-9d91-84e0de80c664_KUKA_Educational_Package_en.pdf>, 04.02.2016, 13:46h

[15] Fiebiger,Bernd and Munz,Heinrich (2014), KUKA-Robotersteuerung KR C4 setzt auf EtherCAT (KUKA robot control unit uses EtherCAT), PC-Control 01 2014, https://www.pc-control.net/pdf/012014/solutions/pcc_0114_kuka_d.pdf, 04.02.2016,14:03h

Implementation and analysis of open source information systems in Electronic business course for economy students

*H. Jerković, *P. Vranešić, **G. Slamić
Zagreb School of Economics and Management,
*IT department, Zagreb, Croatia,
**MBA graduate student,
{hrvoje, pvranesi, goran.slamic} @zsem.hr

Abstract - Electronic business course at Zagreb School of Economics and Managements (higher education economy faculty) is highly practical course with aim to educate students how to use variety of open source information systems for online business purposes. As final project students have to use open source content management system and implement their own business idea using principles taught in theoretical and practical part of the class. Limited versions of various content management systems are offered online but they are not sufficient to fully feature various important business activities. This paper presents testing, analysis and implementation of hosting environment, applications and class activities for course purposes in own educational private cloud environment.

I. INTRODUCTION

A content management system (CMS) is a computer application that allows publishing, editing, modifying, and organizing, deleting, and maintaining content from a central interface. Today CMS are used in almost every industry by millions of end-user organizations [1]. Adapting business process-oriented approach to enterprise content management and finding framework that will allow organizations successful management of content is main starting reasons for using CMS. Immense overload of mainly unstructured data, frequently very long search time, problems fulfilling compliance requirements, maintaining information quality and data representation are some of the challenges that organizations are facing.

This paper analyzes experiences of teaching Electronic Business course at ZSEM. Course goal is to introduce modern open source CMS to economy students and teach them how to use them for implementation of business ideas on the web. Basic theoretical lectures and similar learning activities are not enabling students to get proper understanding of purpose and possibility that various CMS are offering. Course is mainly focused on online business activities related to small and medium enterprises with focus on open source CMS since activities related to large enterprises are mainly covered through other courses. This paper analysis related research in field of education students in working with CMS. Initial sections defines course objectives and then proceeds to compare virtual machine management software which is needed for installation of students' virtual machines. After that main open source CMS solutions and their installation procedures are compared. Lesson workflow for most important lessons is explained in following chapter. Final sections brings analysis of students' experiences while working with CSM and ends with qualitative and quantitative research analysis.

II. CURRENT RESEARCH

Majority of body of research in area of CMS is focused on ECM (Enterprise Content Systems) and their comparison and benchmarking. It's almost impossible to find study on how to empower students to use CMS adequately in always changing online business environment. Research in [2] demonstrates importance of creativity in managing enterprise content and propose it as core value. Researchers in [3] shows how CMS plays very important role enabling users to became creators and editors of the content thus bridging gap between creation and creativity. Other scientific literature and research is scarce in this area and it's mainly focused on CMS statistical and performance comparison or around building various web applications (learning management systems, library information systems, simple web applications) using CMS [12-17]. Researchers in fields of exploring other learning possibilities of CMS for students had various results with their work. Authors in [7] explored usage of CMS in order to analyzed enhance students self-learning experience. Other authors admit that foreknowledge of students plays [8, 14] important role.

Although current body of research offered some useful insights we didn't discovery research that is sole focused on analyzing teaching experience in teaching students CMS.

III. DEFINING COURSE OBJECTIVES

Lead by problematics described in introductions, throughout years of lecturing this course we have redefined main course goals as follows:

- Help students understand and comprehend how modern Internet networks and cloud computing services function,

- Present to students various internet business services and help them to make proper assessment and analysis of those services against needs of their business processes,
- Teach students procedures necessary for starting business online: understanding process of registering their own business internet domain and web hosting site,
- Teach students how to analyze company activities and define business processes,
- Teach students to analyze and test features of available open source applications and information systems,
- Teach students how to create and deploy their business idea in cloud environment.

Aim of Electronic business course is to actively engage students in building business oriented web portal or site. Students can get firsthand experience with creation and deployment of various business processes like creating web based business content, web and video galleries, multilingual site, web shops, understanding and deploying payment gateways, securing web content, search engine optimization best practices etc. Since students are economy students we cannot expect them to build features like that by coding. So we decided to focus on content management systems which became powerful enough to implement all main online business activities without exhausting knowledge of web programming languages. In last class lessons we focused on customization problems and understanding how modest knowledge of web programming languages can help us properly customize CMS in order to satisfy needs of specific online business process.

TABLE 1. COMPARISON OF MAIN CONTROL PANEL SOFTWARE

Control Panel	DNS	Email	FTP	Databases	Multi-Server
ZPanel	x	x	x	x	
ISPConfig	x	x	x	x	x
Virtualmin	x	x	x	x	x

IV. REQUIRED SYSTEMS AND APPLICATIONS FOR IMPLEMENTATION

Next step in process is to define list of requirements regarding implementation of activities described in previous chapter. While analyzing available open source content management systems we decided to compare three main possible solutions from standpoint of satisfying course goals but also satisfying other relevant technical aspects of study: Joomla, Drupal and WordPress.

Regular students can follow course on their own instance of content management system of choice which supposed to support all necessary online business goals described previously. In order to support that we had open subdomain for course purposes and set up virtual machine with Linux operating system, web server, database and DNS server.

A. Virtual server requirements and VM management software

Linux Debian virtual server was set up on faculty cloud infrastructure powered by VMware. Since this is optimal configuration without need for any significant changes regarding choice of Linux operating systems because all Linux distributions support what is needed for installation of required systems. CMS are supporting PHP applications environment, MySQL databases and Apache and Nginx web servers. Second important issue is choice of virtual machines management software which can allow simple management of virtual machines. Beside need for CMS application, students can learn how to open email accounts, create database, understand domain basics and learn how to use SFTP client software (Secure File Transferee Protocol) to transfer files to and from server. TABLE 1. shows comparison for main Control Panel software for management of virtual servers. In first generation we have used ZPanel software which showed lack of important options like ability to quickly create new virtual server and also not so good support at the time of usage. ISPConfig were software of a choice in next generation but in comparison with Virtualmin it proved to be more complicated and with less automated options for end users, also ability to back up and move servers is more easier to use and transparent in Virtualmin as well as many other options. Customization of control panel with themes was important option since it allows better structuring of all possible options. This is important for users with poor foreknowledge since it gives them simplified and visual more appalling approach to common system administration tasks.

B. Deployment of open source applications

Ease of deployment of web applications to virtual server environment was another challenge. How to deploy easily open source application without going through process of installation? Professional solutions like cPanel have integration with automated installers of open source application like Installatron, Softaculous and Fantastico [4, 5] but they lack support for open source solutions like Virtualmin. Virtualmin on other hand has options for manual creation of scripts for automated installation of various popular open source software like PhpMyAdmin and options for adding other scripts for installing software manually. Since we encountered problems in creating scripts for automated installation of common CMS we've took another approach to matter. Problem was resolved by updating course syllabus with additional, and on the end proven to be very useful lesson "Installation of open source web applications". After opening students account we guided them step by step through whole process showing how it's done. Steps included:

- Downloading web application,
- Uploading web application in own web folder using FTP client program,
- Unzipping application in web folder,

- Creating database for application,
- Creating username and password for database,
- Assigning user to database and defining permissions,
- Start of application installation,
- Configuring application during installation,
- Application launch and logging to administration area of application.

Although this procedure is somewhat complicated for economy students it still allows them to understand system processes behind it much better. Students daily access various web applications without basic knowledge of how information is being retrieved.

V. ANALYSIS OF CURRENT CMS PLATFORMS

Main question is which CMS can be used so 25 students can at same time run their CMS administration? After review of existing research in the field [6, 7] we had to start reviewing professional resources since scientific research and resources were poor and in many cases not aligned with our topic of the research. Expert resources and publications of professional hosting companies are largely summed up in and defined in TABLE 2. which shows basic comparison of most used open source software. Current share of the market is following: WordPress 60%, Joomla 6% and Drupal 4%.

TABLE 2. COMPARISON OF MOST USED OPEN SOURCE CMS

	Drupal	Joomla	WordPress
General	Drupal is a powerful, developer-friendly tool for building complex sites. Like most powerful tools, it requires some expertise and experience to operate.	Joomla offers middle ground between the developer-oriented, extensive capabilities of Drupal and user-friendly but more complex site development options than WordPress offers.	WordPress began as an innovative, easy-to-use blogging platform. With an ever-increasing repertoire of themes, plugins and widgets, this CMS is widely used for other website formats also.
Ease of Use	Drupal requires the most technical expertise of the three CMSs. However, it also is capable of producing the most advanced sites. With each release, it is becoming easier to use. If you're unable to commit to learning the software or can't hire someone who knows it, it may not be the best choice.	Less complex than Drupal, more complex than WordPress. Relatively uncomplicated installation and setup. With a relatively small investment of effort into understanding Joomla's structure and terminology, you have the ability to create fairly complex sites.	Technical experience is not necessary; it's intuitive and easy to get a simple site set up quickly. It's easy to paste text from a Microsoft Word document into a WordPress site, but not into Joomla and Drupal sites.
Features	Known for its powerful taxonomy and ability to tag, categorize and organize complex content.	Designed to perform as a community platform, with strong social networking features.	Ease of use is a key benefit for experts and novices alike. It's powerful enough for web developers or designers to efficiently build sites for clients; then, with minimal instruction, clients can take over the site.
Best use cases	For complex, advanced and versatile sites; for sites that require complex data organization; for community platform sites with multiple users; for online stores	Joomla allows you to build a site with more content and structure flexibility than WordPress offers, but still with fairly easy, intuitive usage. Supports E-commerce, social networking and more.	Ideal for fairly simple web sites, such as everyday blogging and news sites; and anyone looking for an easy-to-manage site. Add-ons make it easy to expand the functionality of the site.

B. Comparing CMS performances and user friendliness for class purposes

At this point we were left with choice of Drupal or Joomla. Drupal was more flexible and offered powerful features for implementation of all needed business processes but it also required knowledge of web technologies. We can assume it could be good start for teaching students web programming [8] or even as basis for creating lightweight web applications [9, 10] as existing research already successfully confirmed. Statistical research on CMS performance also bears

Further testing of CMS is split in two main categories:

- Testing three main CMS platforms against defined business process requirements,
- Comparing CMS performances and user friendliness for class purposes.

A. CMS testing against defined business process requirements

We've tested main business process defined by individual objectives of the course. Advancements in CMS and templating frameworks enabled surprisingly simple procedures for implementation of core business processes. But in early years (2012, 2013) when Electronic Business course was started our analysis shown that Joomla or Drupal could be CMS of the choice. Although they are more complicated for using than WordPress they offered more options for structuring content. WordPress at the time seemed unable to handle more complex content and lacked proper plugin support in various important features like multilanguage site and advanced front end content structuring. Therefore, when analyzing from perspective of support for required business processes (course objectives) we had to conclude that WordPress, at that time, was not suitable to be CMS of the choice.

weight because we need 25 operational hosts on single virtual machine with limited capabilities (4G RAM). Research, testing practice and experience with usage showed that for informative and intranet site Drupal is better, for intranet with multiple functionality site Joomla is better and for live site WordPress offers best results [11]. But main criteria was definitely user friendliness and ease of use. If comparing Drupal and Joomla based on that criteria Joomla was certainly more used friendly and it became only choice for teaching economy student with limited foreknowledge in year 2012 and 2013.

Although WordPress could be winner in this category for previously explained reasons we could not use him in early years of the course but it was definitely CMS of the choice in recent years of the course (2014, 2015) because newer versions supported implementation of all necessary teaching goals.

VI. CORE OBJECTIVES CLASSROOM WORKFLOW

Most challenging part while educating students was figuring out proper workflow during man lessons. First lesson explains and demonstrates virtual server environment and its relevance in today's businesses. Also students are presented with examples of virtual servers in every day usage. So in that lesson student are thought meaning and purpose of virtual servers in cloud infrastructure and on the end of the lesson, using Virtualmin, student create their own virtual server one by one using professors administration credentials.

Figure 1. shows simplified workflow of "Web application databases" lesson which comes as second lesson in "Installing web applications" module. Students are first introduced to concept of databases and their current understanding is questioned. After that in class discussion we are testing students' comprehension of databases before we continue with next step. Similar is applied in step "Examples of real life usage of databases". Primary reason for such thorough approach lies in fact that first time we run this course we had just theoretical explanation of databases and purposes of the databases and then we can go straight to procedure of installing web application. Although students managed to install applications successfully they still lacked understanding of connections and meaning of database and web application that is using it.

After creating database, user and user credentials student assigned user to database. After installation of CMS we returned on this lesson and finished it by showing students how plain blog post or page looks inside database after they create it in CMS administration. This procedure has high relevance because successfully links missing understanding and knowledge for non-technical students and gives them pretty good comprehension of matter at hand.

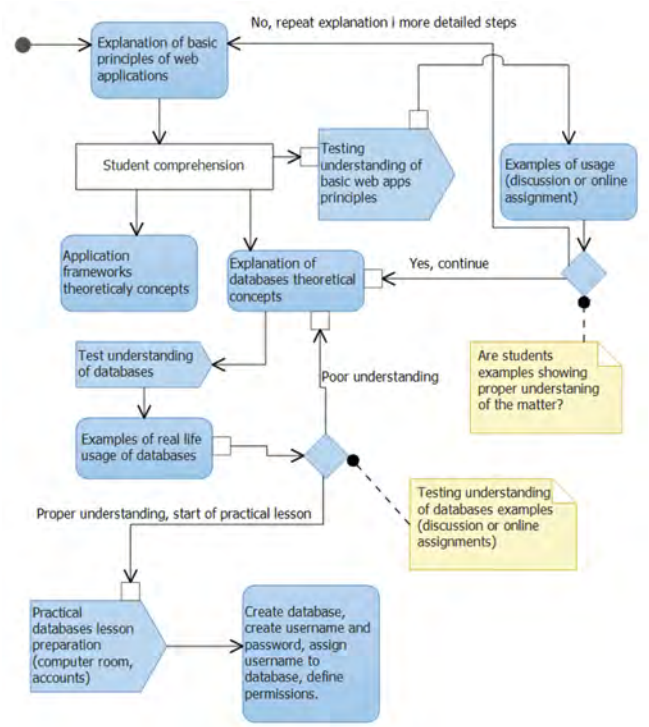


Figure 1. Workflow of "Web application databases" lesson

In similar way other objectives are implemented in form of practical lesson after few cycles of trial and error. Hugh advantage of such approach is that students get to do process themselves and at the same time understand the meaning of individual steps.

VII. STUDENT SATISFACTION ANALYSIS

We conducted research with students that were actively engaged in course. Around 20-25 students each year was actively participating in the class and class activities. Surveys allowed better understand of students' experiences during class in regard with setup, implementation and deployment of various business processes while using CMS applications. While reviewing these statistics it's important to note that in years 2012 and 2013 we have used Joomla and in 2014 and 2015 WordPress is used.

Figure 2. shows student survey on complexity of CMS installation procedure. It's clear that perception of complexity changed in last year showing increase of around 10%-20% of students shifting from "very hard" to "hard". It's important to note that majority of students were exchange students from all parts of the words with very different skills and their foreknowledge is non-technical for large majority.

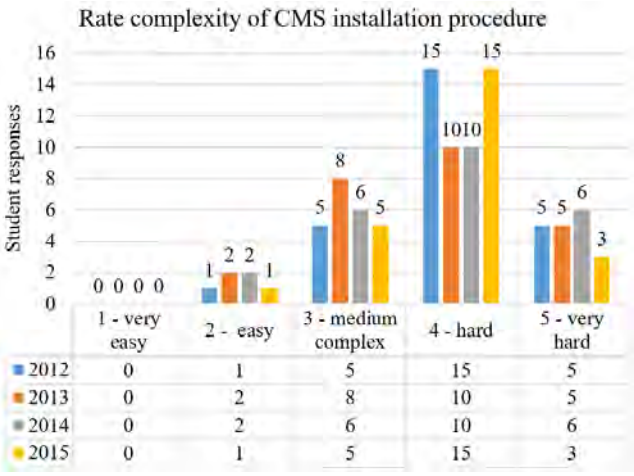


Figure 2. Student survey on complexity of CMS installation procedure

Figure 3. shows that students experience while using CMS shifted from dominantly "medium complex" to "easy".

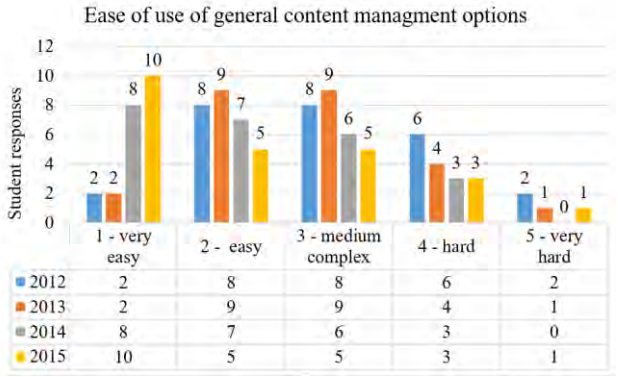


Figure 3. Student survey on ease of use of general content management options

Figure 4. shows students satisfaction with their final project and completeness of their business idea. Figure shows that students were less satisfied with their final achievement.

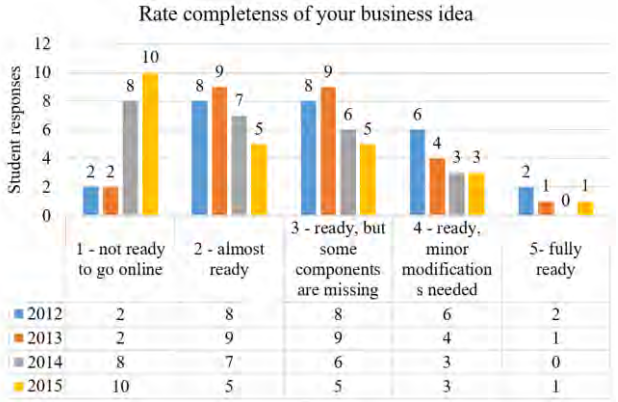


Figure 4. Student survey results on rating completeness of students' business project idea

One of the reasons for that lies in the fact that project itself become far more demanding since we forbid project ideas that has no business value and allowed submission of partially finished projects as long as students implemented everything that was taught in the class. Although is seems like this is not a good way to teach course if final objective is only partially met but fact is that final teaching goals are fully met and students after finish of the class have understanding of what is missing and what exact knowledge they need in order to make their business idea fully realized.

A. Course and student projects qualitative analysis

Individual student presentations were good source of qualitative information retrieved directly from individual students. First part of student project presentations gave us rounded insight in students' complete and comprehensive understanding of the matter and proper way of implementation of that understanding in finalization of their projects. Assessment in middle of semester only gave us idea on parts of course we need to explain better so students could implement that particular feature properly. But it was not possible to get final picture of their project and therefore insight in their overall understanding.

We are listing five main students' responses on our questions that gave us good idea on how to improve course. Statements are listed by perception as most important by student and most frequently mentioned during final presentations:

- CMS is easy to use and I had just few problems which I managed to resolve,
- Some procedures regarding usage of plugins are easy to follow but hard to repeat on my own, or end up being too complicated but I managed to understand them by exploring additional resources by myself,
- Limitations posed by themes are sometimes frustrating. It should be simple to change simple things like layout, font, sizing and spacing,
- I've faced problems I could not resolve while finalizing my business idea because I don't have enough of HTML, CSS or programming knowledge from PHP,
- I found course to be challenging but extremely useful. I've learned and understood so many useful and practical lessons which will help me in my business.

VIII. EVALUATION OF THE RESULTS AND COMPARISON TO CURRENT RESEARCH

Compared to current research this paper brings novelty approach in teaching economy students management of open source applications from installation to practical implementations of business processes. There are no similar relevant attempts in current body of research in regard with practically teaching economy students how to install and use CMS applications in business purposes. From given results we can see that majority of students are still perceiving installation process hard, medium complex or very hard. This didn't change significantly with shift on WordPress simply because WordPress,

although simpler to use, doesn't have simpler installation procedure than Joomla. Unfortunately installation cannot be simpler that it already is. Our conclusion of this part is that there is simply too many gaps in students' foreknowledge so even when they are introduced with each step properly, still much more time is needed to elaborate on whole procedure so students could have proper understand of matter. Automated installation procedures are needed to simplify this part of the course. Student survey on ease of use of general content management options shows significant shift towards "very easy" for last two years when WordPress is used, from "medium complex" and "hard" for previous years when Joomla is used. Student survey results on rating completeness of students' business project idea shows that students in previous two years were less satisfied with completeness of their own ideas. In our opinion overall implementations of business ideas are more complete than when build with Joomla. Students understand WordPress quickly so they gain better understanding faster when compared to case where Joomla is used. Problems beyond their understanding and beyond scope of course – advanced CSS, HTML or PHP editing, are main cause of issues. Nevertheless students have desire to finalize projects and try to understand all that's needed for project finalization. In either cases intrinsic motivation is present with positive desire to explore technology deeper and finalize project on their own.

CONSLUSION

Students' understanding of core features of CMS is what got positive results in our experience. CMS offers a lot of functionalities and many educators don't have full understanding of all underlying procedures, possibilities and processes. Majority of mainstream CMS are modern and continually expanding frameworks with extensive plugin support. As such they offer countless possibilities for education of students on every level of foreknowledge.

REFERENCES

[1] W. Maass and T. Kowatsch, *Semantic Technologies in Content Management Systems: Trends, Applications and Evaluations*: Springer Science & Business Media, 2012.

[2] J. Vom Brocke, A. Simons, and A. Cleven, "Towards a business process-oriented approach to enterprise content management: the ECM-blueprinting framework," *Information Systems and e-Business Management,* vol. 9, pp. 475-496, 2011.

[3] J. V. Brocke, S. Seidel, and A. Simons, "Bridging the gap between enterprise content management and creativity: A research framework," in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, 2010, pp. 1-10.

[4] S. Raharjo, "Building Web Journal Directory And Its Articles With Drupal" *DASI,* vol. 14, p. 64, 2013.

[5] A. Moore, *The Creative Person's Website Builder: How to Make a Pro Website Yourself Using WordPress and Other Easy Tools*: HOW Books, 2013.

[6] A. Iglesias, L. Moreno, P. Martínez, and R. Calvo, "Evaluating the accessibility of three open‐source learning content management systems: A comparative study," *Computer Applications in Engineering Education,* vol. 22, pp. 320-328, 2014.

[7] G. Lust, N. A. J. Collazo, J. Elen, and G. Clarebout, "Content Management Systems: Enriched learning opportunities for all?," *Computers in Human Behavior,* vol. 28, pp. 795-808, 2012.

[8] X. Wang, "Design, develop and teach the second web programming course in computer science curriculum," *Journal of Computing Sciences in Colleges,* vol. 29, pp. 52-59, 2014.

[9] B. M. Pandiyan and M. Singhal, "Development of a Lightweight Learning Management System using Open Source Software Drupal," *Journal of Current Trends in Library and Information Science,* vol. 1, pp. 1-5, 2015.

[10] S. J. Ling and B. R. Lan, "A Research on Development of Drupal-Based Topic Service Website," in *Advanced Materials Research*, 2014, pp. 1840-1843.

[11] S. K. Patel, V. R. Rathod, and S. Parikh, "Joomla, Drupal and WordPress-a statistical comparison of open source CMS," in *Trendz in Information Sciences and Computing (TISC), 2011 3rd International Conference on*, 2011, pp. 182-187.

[12] Z. A. Adhoni, H. Al Hamad, and A. A. Siddiqi, "CBQ-API: A Cloud-Based Programming Interface for Qur'anic Applications," in *IT Convergence and Security (ICITCS), 2013 International Conference on*, 2013, pp. 1-5.

[13] F. McKelvey, "FCJ-128 A Programmable Platform? Drupal, Modularity, and the Future of the Web," *The Fibreculture Journal,* 2011.

[14] M. B. Rosson, E. Thiry, D. Zhao, and J. M. Carroll, "Teachers as designers: enabling teachers to specify dynamic web programming projects for students," in *End-User Development*, ed: Springer, 2011, pp. 233-238.

[15] L.-s. Qi and A.-k. Tian, "Design and Application of Hybrid Learning Platform Based on Joomla," in *Advances in Computer Science and Education Applications*, ed: Springer, 2011, pp. 549-556.

[16] M. Abdulwahed and Z. K. Nagy, "Developing the TriLab, a triple access mode (hands‐on, virtual, remote) laboratory, of a process control rig using LabVIEW and Joomla," *Computer Applications in Engineering Education,* vol. 21, pp. 614-626, 2013.

[17] R. Chowdhary, "Programming Technology and Web," 2012.

Virtual Firms as Education Tool in the Field of eCommerce

M. Vejačka*
* Department of Applied Mathematics and Business Informatics
Faculty of Economics, Technical University of Košice, Slovakia
martin.vejacka@tuke.sk

Abstract – The use of information and communication technologies brings many opportunities to innovate educational process. In the case of education of potential future entrepreneurs, it might be suitable for them to try out firm establishment and operation in virtual environment. Usage of virtual companies is supported by ICT means. The aim of this paper is to present how virtual companies serve for educational purposes at our faculty. At our faculty, we have started several years ago with such a subject of study that allows students to experience setting up and running their own small virtual business in the first period of its existence. Students set up small trade firms in form of limited liability companies, which operate e-shops. Virtual bank of our faculty provides services of electronic payments in e-shops. Also electronic communication with different virtual institutions supported by use of digital signature is used. Our results show virtual firms as very good mean to allow students to gather experience with the establishing and operating firms and develop practical entrepreneurial skills.

I. INTRODUCTION

At present, the use of information and communication technologies (ICT) influences multiple aspects of economic life. Entrepreneurs might use it to communicate with administration institutions, for distribution of their products to customers, during usage of banking and payment services or even for establishing their firm. Development of ICTs invoked formation of electronic businesses and electronic commerce.

Electronic commerce (e-commerce) refers to the facilitation of trading of products or services using information and communication technologies. E-commerce allows distribution of products or services for companies that use it to wider range of potential customers. E-commerce often uses other information and communication technologies such as electronic funds transfer, online marketing, electronic data interchange, enterprise resource planning, data collection systems etc. Electronic commerce uses the web for at least one part of the transaction and it may also use other information technologies. E-commerce activities of businesses may include use of online shopping web sites (e-shops) for retail customers, online marketplaces for customers or other businesses, online marketing and sales promotion and also business-to-business electronic data interchange [1].

Electronic commerce by its nature can be connected with electronic payments and electronic banking. The characteristic feature of electronic banking is continuous all-time access of payment services. The access that is independent of business hours with convenient and rapid exchange of information between the retailers, banks, card companies and clients. Electronic payments also allow to conduct transactions virtually from anywhere in the world at any time. This is very beneficial for retailers bringing cash flow without need of brick-and-mortar shop branches, what can save time and costs.

Electronic businesses routinely have to communicate with public administration institutions to fulfill their legal obligations, such as providing tax declarations, applying for permits and licenses etc. If such communication and interaction with public administration institutions is allowed by electronic means, we speak about electronic government (e-government). The firms interact with the government-to-businesses (or commerce) level of e-government (G2B level of e-government) that is specifically adjusted for the needs of businesses [2].

It is appropriate to include all these aspects so that students can acquire experience the operation of such businesses as closely as possible, when training future entrepreneurs in the field of e-commerce. Information and communication technologies besides of above mentioned effects on businesses also routinely support educational process and help to develop knowledge and skills of students.

At our faculty we try to use ICT intensively in educational process. They support education process not only in computer sciences and informatics, but also in many other subjects of study [3]. Information technologies allowed emergence of e-learning that is very frequent in distant learning form of education in our conditions [4]. E-learning might help to improve ICT skills and knowledge of university teachers or support other innovative forms of education [5]. ICT might help future entrepreneurs to analyze the processes in their businesses, reprocess and improve them with aim to streamline the operation of their firms.

II. VIRTUAL FIRMS

All above mentioned aspects of electronic business are reflected in our education of future entrepreneurs. At Faculty of Economics, Technical university of Košice, we provide education in facultative subject of study to the students in third year of study in fields of Finance, Banking and Investments, and Business Informatics. The

course is open for all students in both distant and face-to-face study forms. The course is named Economic Information Systems (EIS) and it allows simulation of doing electronic business in conditions of Slovak economy with use of virtual firms. The course of EIS is based on e-learning methods to allow course completion without necessity to visit lectures personally. Face-to face parts of the course have only consulting character and they are not obligatory for students what is similar to some other courses provided at our faculty. The course is support by learning management system (LMS) Moodle.

The course of Economics Information Systems (EIS) allows students in the form of establishing and operation of virtual firms to try and experience the basics of entrepreneurship. The use of virtual firms or businesses for training and education was investigated before and its use has been shown beneficial [6].

These students' firms have form of one-person limited liability company as very common form of business in local conditions. Within their business activities students are restricted to operate e-shop with one of three groups of goods (computer equipment, office equipment or office furniture). Lecturers operate central faculty firm, which serves as wholesale, where students' firms might buy their virtual goods for further selling.

Students establish their firms according to legislative conditions of Slovak republic. Virtual electronic registry operated by lecturers is used to emulate public institutions (business register, tax office, social insurance office etc.) and electronic communication with them. During this communication digital signature is intensively used. Digital signature certificates for students are issued by faculty's certification authority. Within the process of establishing the company students have to develop and prepare the business plan for the first year of their firm's existence. Another important step is to create company's web site and upload it to web place provided by faculty's web server.

The important part of company's web site is also e-shop, where all firm's products are sold. Basic solution of e-shop for students is provided by faculty, however they have to adjust it to their needs. E-shops have to be filled by their goods and maintained by students until the end of the course. The source of goods is faculty's main e-shop serving as wholesale. Clients are emulated by other students, which are buying goods not sold in their e-shops. In this case they act as private persons to generate some demand for goods. Demand is also supported by lecturers buying in students' e-shops. Payments for sales in e-shops are conducted via faculty's virtual bank, which also provides all firms with starting loans.

Before the end of the EIS course students are required to book all their operations into accounting. During last few years various accounting softwares in trial versions were used in this part of our course. The last task in course is to develop the strategic plan into the future of their company. It contains also proposals of changes in processes of their firm and prospects into the future of their virtual firms. At the end students provide feedback on course to lecturers.

Following electronic systems are intensively used during the course duration to support operation of virtual firms.

A. Learning Management System

During the course learning management system Moodle as main medium for information provision to students is used [7]. LMS Moodle is used for e-learning purposes at our faculty as standard. It serves also for basic communication with students. Schedule of the EIS course and all tasks during course are published here. Likewise all study materials are provided by lecturers in Moodle. Students may check if they are meeting course's deadlines and their tasks are evaluated via Moodle.



Figure 1. Learning Management System Moodle and the course of Economic Information Systems

Some of the tasks (e.g. business plan of company, outputs from accounting, strategic plan) are uploaded directly into the LMS by students and then evaluated by teachers.

B. Electronic registry and digital signature

While establishing their virtual firms students have to communicate electronically with virtual public administration authorities. These authorities (such as business registry) are represented by electronic registry and student have to communicate with them using electronic communication. All processes and legal forms are adjusted according to valid Slovak legislation. Students have to register their future business activities with Trade Licensing Register and their firms in Business Register. Later they have to register their firms with Tax Office, Social Security Office and Health Insurance Company. All necessary forms are electronic and after filling they can be sent to virtual office via faculty's electronic registry.

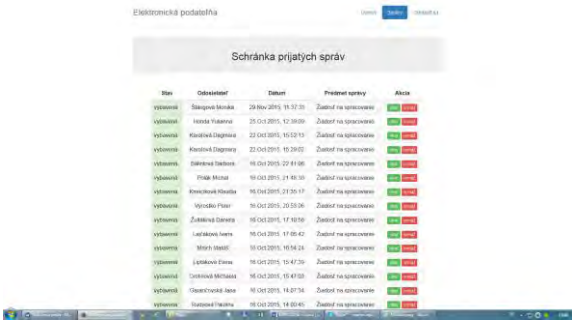


Figure 2. Electronic reqistry system

These forms have to be authorized by digital signature. Important part of course is dedicated to the area of

electronic communication with use of electronic or digital signature. In this part of our course following basic principles are discussed:

- electronic data interchange,
- symmetric and asymmetric encryption,
- digital signature and electronic signature,
- electronic signature legislation,
- Public Key Infrastructure (PKI),
- certification and registration authority.

Information and instructions about the use of electronic signature for communication are provided via web page which also incorporates the certification authority of our faculty (available at ca.ekf.tuke.sk). Our certification authority provides private certificates of digital signatures for students for the purposes of our course. It is based on EJBCA open source PKI certificate authority. The singing of applications forms and documents which are in pdf form is finally conducted in Adobe Acrobat Reader DC.

The possibilities of electronic signature use in communication and authentication in general is discussed. Possible roles of electronic signature in other areas e-government is briefly described. The difference between the terms of electronic and digital signature according to our legislation is explained. While electronic signature is broader legislative term, digital signature is one of technological solutions corresponding with theoretical definition of electronic signature [2].



Figure 3. Certification authority of Faculty of Economics

Public Key Infrastructure (PKI) is explained theoretically. Students get also practical experience with PKI by using faculty's certification authority, when generating their own digital signature certificates. Web site of certification authority also contains manuals for use of certificates.

Practical use of digital signature in the EIS course consists of singing all necessary electronic documents and sending them to corresponding virtual authorities. Such a communication reflects the real possibilities of communication with given public administration authorities in Slovakia basically.

C. *E-shop*

The main substance of e-commerce part of the EIS course rests in setting up and managing electronic shop where all goods of virtual firms are merchandised. All the goods are have their source in central wholesale e-shop of faculty's firm managed by lecturers. The goods have only virtual character and they are only included in the inventory of central firm with no production costs. Students' firms buy any goods which they are selling in their e-shops according to the group of goods (computer equipment, office equipment or office furniture) of their interest. To simulate demand after their goods students should buy goods from the e-shops of other students from two other groups of goods sold with aim to furnish their company's virtual premises with equipment or furniture. Lecturers also support demand with random purchases in all students' e-shops.



Figure 4. E-shop of virtual firm

Students have to represent their firm also by its website. Website should be design to attract potential customers to buy goods in company's e-shop and it is main tool of company's marketing. If customer is interested in buying some product from given firm, he should go to e-shop part of website, where registration is needed. After registration and logging in classical e-purchase process is used. Customer picks up goods from various categories into virtual shopping cart and when finished he proceeds to online payment.

Students in this part of the course have the opportunity to experience both sides of selling and buying process in e-shop using virtual firms.

D. *Electronic payments*

Electronic payments supporting the purchase and sale process in e-shops are conducted via internet banking system of faculty's virtual bank. Electronic Online Bank (EOB – ebanka.tuke.sk) is virtual bank provided to faculty for free in outdated version by its developer company UNICOM. Although it is old, it serves well for educational purposes of various courses at our faculty. Students get loans of virtual euros in this bank to allow buying and selling process.



Figure 5. Virtual bank – Internet banking interface

Realization of online payments in e-shop will lead buyer to virtual bank's internet banking prefilled with parameters of payment. Virtual bank's online payment

system supports also digital signature as authorization element.

Also identification of bank user is supported by digital signature. The security of the whole payment in e-purchase process is the key issue. Students by practical use of online payments familiarize themselves of this crucial step of e-commerce. Technology is developing to provide online payments in secure way. Safety and security of their use is also a question of users' safe behavior. At least, clients should observe safety rules while using online payments.

Technological approach to security of online payments brings variety of methods for authentication and authorization of electronic payments. Main forms of authorization possibilities are GRID card, One-Time Password generator (token) or SMS code or digital signature. Highest level of security is represented by using digital signature for authorization purposes [8].

The main issue in the EIS course regarding electronic payments in e-shop was to interconnect basically independent systems of virtual bank with e-shop to allow automatic transfer of payment data from e-shop to banking system. Then students only have to set up an account number in e-shop administration menu.

III. SURVEY

In year 2015, survey about students' experience with electronic communication, e-commerce and electronic payments was conducted as component part of feedback gathering within the Electronic Information Systems course. Survey was performed before and after the course completion. Students' knowledge and experience with establishing and managing small trade firm, electronic commerce, electronic business, electronic communication with public administration and electronic payments was investigated. The survey results are source of relevant suggestions to improve the education process within the EIS course.

The survey was executed electronically via the web questionnaire and respondents were addressed by link in the course in LMS Moodle. We gathered 129 completed questionnaires from students of the EIS course what represents 87 percent of recovered questionnaires. It is very high percentage of successful recovery of responds what was probably caused by filling out the questionnaires during lectures.

Age composition of respondents was caused by fact that both internal and distant students were responding to the survey. While external students are often older persons already having job in opposite with internal students, the variety of ages between respondents was bigger. However the majority of respondents were students in age between 20 and 25 years, where the most students of internal form of study is.

Gender composition of group of respondents shows that over 60 percent of our survey respondents were women. This reflects also the composition of students at our faculty. The area of residence often affects availability of broadband internet services in Slovakia, which is better in bigger cities. Above one third of respondents is from smaller towns under 40 000 citizens, where availability of broadband internet lower at their residence.

All relevant demographic data gathered is available shown in following table (TAB. 1):

TAB. 1: DEMOGRAPHIC DATA OF STUDENTS IN EIS COURSE

Age	**Count**	**Percent**
20 - 25	95	73.6
26 - 30	23	17.8
31 and over	11	8.6
Gender		
Men	51	39.5
Women	78	60.5
Residence		
Large-sized city (over 100 000 citizens)	83	64.3
Medium-sized city (40 000 to 100 000 citizens)	8	6.2
Small town (5 000 to 40 000 citizens)	12	9.3
Village (under 5 000 citizens)	26	20.2

Source: author

Following results refer to survey before the start of EIS course.

Only 8 percent of respondents indicated any experience with the establishing or managing firm. All other respondents did not have any experience with entrepreneurship. All respondents indicated to have theoretical experience with business legislation including the process of establishing the firm, while they already have taken part in course regarding Slovak business legislation in previous years. Only 3 percent of respondents indicated practical experience with establishing businesses.

Almost 18 percent of respondents had some practical experience electronic communication with public administration authorities, but mostly not in position of firm's executive, but only as private persons.

Approximately 84 percent of respondents adduced some experience with preparing the business plan, what again is related to the content of study subjects in first two years of study at our faculty. No one has indicated using electronic or digital signature before course, while using any e-government applications in Slovak conditions. Although electronic communication with use of digital signature is favored by price directly in legislation conditions. This reflects low level of usage of digital signature in e-government in Slovakia at the time [9].

The most of students (87 percent) had experience with developing the simple website, while it was mandatory tasks in subject of Informatics in their first year of study. However connecting an e-shop to their website and management of this e-shop was indicated only in one case. Given respondent had small real life firm operating e-shop. Around 71 percent of students indicated some experience with buying in real e-shops. Approximately 91 percent of respondents noted that they have already used electronic payments on the web, but only one respondent reported to be recipient of electronic payments in his own electronic business.

All of the students had theoretical experience with accounting, but practical experience with accounting in real firm had only 6 percent of them. Over 41 percent indicated experience with preparing strategic plan of firm for its future development.

These results of survey before the start of EIS course show that whole course uses the theoretical experience of students acquired in other subjects and lower years of study and allows them to get some practical experience although only in virtual environment.

After EIS course completion another survey was conducted and its following results are presented further.

Over 98 percent of respondents adduced gaining valuable experience with the establishing or managing firm during EIS course. Remaining 2 percent indicated that the course did not bring them any new experience with entrepreneurship. It represents part of the students that already did business in real environment before the course.

Almost 87 percent of students reported that they gained new information on the process of establishing the firm according currently valid Slovak legislation through the EIS course and 92 percent indicated new experience with establishing firms.

All respondents welcomed possibility to acquire practical experience with electronic communication with public administration authorities via faculty's electronic registry, while using digital signature. We consider this area as one of the most beneficial for course attendants in future praxis.

Over 39 percent of students noted that preparing the business plan for their virtual firm was beneficial even when they already had experience with business plan preparation. Similarly 22 percent adduced that creating the simple company website was beneficial and brought them some new deeper experience in this area.

Setting up and managing e-shop of their own virtual firm was new and valuable experience for around 61 percent of students. They appreciated in particular the possibility to try out operation of e-shop from the side of e-shop's operator. As the most important issues of this part of operating their virtual firms they considered implementation of electronic payments system into e-shop and inventory monitoring and management.

Respondents got also practical background in accounting of first financial year of their newly established virtual firm. Almost 59 percent of respondents found it potentially beneficial for their future. Even more (72 percent) welcomed creation of strategic plan for their virtual firms into following years.

Almost all respondents (92 percent) welcomed the opportunity to try out entrepreneurship in form of setting up and operating a virtual firm in form course. This result indicates that virtual firm is a very good form of interposing the experience with operating the firm from the students' point of view.

IV. CONCLUSION

Participants of the Economic Information Systems course have the opportunity to gain practical experience with entrepreneurship in the field of e-commerce using virtual firm. Students try out establishing a company within the conditions of Slovak legislation using electronic registry and digital signature during communication. Later they prepare business plan for their newly established firms. Developing company's website with e-shop, its management and operation is further step in the course. After the notional end of financial year, participants also book their operations into accounting and prepare strategic plan of their firms for future economic periods.

According to the results of survey conducted in recent year, the most participants highly appreciate the opportunity to practically experience the operation of company and to develop practical entrepreneurial skills in the form of virtual firm.

REFERENCES

[1] Y. Bakos, "The Emerging Landscape for Retail E-Commerce," in Journal of Economic Perspectives, vol. 15, no. 1, 2001, pp.69-80.

[2] M. Yildiz, "E-government research: Reviewing the literature, limitations, and ways forward", in Government Information Quarterly 24, 2007, pp. 646–665.

[3] L. Révészová, "Informatics education and requirements of current practice" MIPRO 2015, Rijeka : MIPRO, 2015, 38th International convention proceedings: Opatija, Croatia, pp. 868-873. May 2015

[4] D. Paľová and L. Révészová, "Importance of ICT and education for knowledge society development," ICT as a Determinant Factor in the Development of the Public Services in the Knowledge Society, Targu Mures: Petru Maior University, pp. 38-74, 2013.

[5] I. Korobaničová, D. Paľová and N. Urbančíková, "Knowledge Generating House – a Tool for Education Based on Open Innovation Principle," MIPRO 2015. 38th International convention proceedings: Opatija, Croatia, pp. 832-837. May 2015.

[6] M. Bitter-Rijpkema, P. B. Sloep and D. Jansen, "Learning to change: The Virtual Business Learning approach to professional workplace learning," in Journal of Educational Technology & Society, vol. 6, no. 1, Training Issues for Successful ICT Innovation in Companies, 2003, pp. 18-25.

[7] Moodle.org (available at: http://www.moodle.org), last visit: February 5th, 2016.

[8] T. Bálint and J. Bucko, "Comparative analysis of handwritten, biometric and digital signature," in International Review of Social Sciences and Humanities, vol. 4, no. 2, 2013, pp. 43 - 53.

[9] United Nations Department of Economic and Social Affairs, "United Nations E-Government Survey 2014," UN, 2014 (available at: https://publicadministration.un.org/egovkb/Portals/egovkb/Documents/un/2014-Survey/E-Gov_Complete_Survey-2014.pdf), last visit: February 3rd, 2016.

Systems and Software Assurance – A Model Cyber Security Course

V. Jovanovic, J.K.Harris
Georgia Southern University/Computer Sciences Department, Statesboro, USA
vladan@georgiasouthern.edu
jkharris@georgiasouthern.edu

Abstract – In this paper we present a model Cyber Security course primarily designed to address the educational and training needs for developing secure software intensive systems. The course is designed, within a Computer Science curriculum, to satisfy the national requirements of the Information Assurance Workforce Improvement Program (per DoD8570.01-Mfrom 11.10.2015), for Tiers II and III jobs, primarily in the Systems Architecture and Engineering (IASAE) specialty areas. The course maps to the knowledge bases of two approved Certifications for IASAE, namely the Certified Security Software Lifecycle Professional (CSSLP) and the Certified Information Security Engineering Professional (ISSEP). The course content is focused on four key elements: Security Architecture Analysis with Treat Modeling, Risk Assessment based on the Cyber Security Framework (and the NIST SP 800 series of guidelines), Secure Development Life Cycle under a process improvement framework CMMI, and a standardized set of Secure Coding best practices. At the graduate level, this course also addresses the making of formal Assurance Cases for Security and Safety of Software Intensive Systems, as well as Software Failure Mode Effect Analysis (FMEA).

I. INTRODUCTION

This paper elaborates on the intent and the content of a Systems and Software Assurance (SSA) course, serving as a cornerstone course for software developers interested in supporting cyber security and receiving related certifications.

The System and Software Assurance course is completely redesigned from its first incarnation [41] in an ongoing effort to improve a curriculum [44] specifically within an undergraduate computer (cyber) security certificate in an an area recognized as strategically important due to an increasing demand from both government and industry. Furthermore, our SSA course is designed to serve a critical need in a proposed graduate concentration in Cyber Security.

The SSA course content is organized around the use of relevant standards/guidelines and organizational/system level processes and practices. It is instructive to compare references from the [41] version (designed to cover training requirements such as [6] and [7]), and the current NIST guidelines, see references [20] to [35], as almost 70% of them are either new or have been substantially updated in the last eight years. The SSA course, in the context of our Computer Sciences (CS) curriculum [44] has direct prerequisites (Ethics, Computer Security, Data Communication), as well as supporting coursework (Computer Architecture, System Software, Distributed Web Systems Design, Database Design, Object-Oriented Design, and Software Engineering capstone as required coursework). The Cyber Security Certificate requires three courses out of following electives: SSA, Network Management, Software Testing and Quality Assurance, and Selected Topics relevant to security such as Storage/Cloud Technology courses derived from [42] and [43]. A similarly proposed graduate concentration in cyber security requires the following courses: SSA, Applied Cryptography, Network and Computer Security, Data and Database Security, Ethics and Research in CS, as well as one additional elective from Computer Science, Electrical Engineering or Information Technology.

A number of certificates are required, or at least expected, to reach performance levels required by many jobs (see Table I). The mapping for Table I was derived from [8] by selecting only certificates relevant to CS students. Approved certificates mapped to jobs are shown for Information Assurance (IA), both Technical and Management (IAT and IAM), Computer Networks Defense Service Providers (CNDSP) specializations, and of primary interest, the IA System Architecture and Engineering (IASAE) job specializations.

Our Cyber Security Certificate prepares students for the Certified Information System Security Professional (CISSP) certification (see Table II). Of primary interest to future software developers (i.e. the majority of our students) is the Certified Secure Software Lifecycle Professional (CSSLP), a certification independent of the CISSP. Table III lists mapping of CSSLP to relevant portions of our CS curriculum and our SSA course in particular. In fact the SSA course alone maps very well to the desirable Information Systems Security Engineering professional (ISSEP) certificate (see Table IV).

TABLE I. JOB CATEGORIES VS. SELECTED CERTIFICATION

Job/Level	Certificates
IAT Level I, IAT Level II	A subset of CISSP
IAM Level II, IAM Level III, IAT Level III, IASAE Level I	CISSP
IASAE Level II	CSSLP or CISSP
IASAE Level III	CISSP plus ISSEP/ISSAP
CNDSP: Manager; Analyst, Auditor, Infrastructure Support, Incident Responder	CISSP plus ISSPM; Certified Ethical Hacker (CEH)

TABLE II. **CISSP** DOMAINS VS. RELEVANT CS COURSEWORK

CISSP 8- Domains	Relevant CS Course
Security and Risk Management	System & Sw. Assurance
Asset Security	System & Sw. Assurance
Security Engineering	Computer Security, Ethics-Computers and Society,
Communication and Network Security	Computer Security, data Communication, Network Management, Storage Technology-Cloud Computing
Identity and Access Management	Computer Security
Security Assessment and Testing	Systems and Software Assurance
Security Operations	Computer Security
Software Development Security	Systems and Software Assurance Software Testing and QA, Software Engineering Capstone Project

TABLE III. MAPING **ISSEP** TO ASSURANCE COURSE

ISSEP 4- Domains	Systems and Software Assurance modules/assignments
Systems Security Engineering	M2 HW1- Treat Modeling, M2 HW2- Software FMEA, M4 Research Paper, M3 Secure Coding Exercises, M3 Final Project, Plus - Computer Security course, and elements from -Software Testing and Quality Assurance course
Certification & Accreditation (Risk Management Framework)	M2 Midterm Project - Assessment Phase
US Government Information Assurance Related Policies and Issuances	M2 Midterm Project - Security System Gaps Analysis - Capability Improvement
Technical Management	M2 Midterm Project -Organization and Management Software Engineering capstone project course

TABLE IV. MAPING **CSSLP** TO ASSURANCE COURSE

CSSLP 8- Domains	Systems and Software Assurance modules and/or supporting courses
Security Software Concept	M1 lectures/discussions
Secure Software Requirements	M2 HW1- evaluation only, Software Engineering course
Secure Software Design	M2 HW1- evaluation only, Software Engineering course, Object-Oriented Design course
Secure Software Implementation/Coding	M3 Secure Coding Exercises M3 Final Project
Secure Software Testing	Software Testing & QA course – part of the Certificate
Software Acceptance	Software Testing & QA
Software Deployment, Operations, Maintenance and Disposal	Software Engineering capstone project course
Supply Chain and Software Acquisition	Software Engineering capstone project course

Detail course design is presented in Table V, and includes schedule of topics, key assignments and related recommended readings (shown in italic). The SSA course content and student outcomes are verified against relevant work specializations from the latest edition of the US NICE Workforce Framework [45] that is replacing requirements from [6] and [7]. Out of 31 work specializations of the Cyber Security Workforce [45] our Certificate covers over 13. A detailed mapping is under development and is expected to be presented at the conference and used in the process of applying for our university's recognition as a center of excellence in cyber security education.

TABLE V. SCHEDULE OF TOPICS AND ASSIGNMENTS

Module 1 Intro	Orientation
Week 1	Scope, jobs, terms of reference, standards Cyber Security Risk Management Framework *[17], [23], [1],[36],[39]*
Module 2 System assurance	System Assurance areas for software developers (including at organizational level), regarding Software/Data Intensive Systems
Weeks 2, 3, and 4: Security architecture - best practices in Treat Modeling	- System Architecture, Data Flow Models - Treat Modeling (TM): STRIDE, Treat Trees, CAPAC - Microsoft TM Tool 2016 laboratory practices - Elevation of Privilege Game - Analysis of architecture: six case studies - Linking TM with Risk, DRED. **HW1 Assignment: Treat Modeling-** 3 cases each team using MS TM tool. *[3], [4], [5], [18], [19]*
Weeks 5, 6, and 7: Assessment methodology	- Regulation/Guidelines and Assessment Cases - Methods: IAM, RMF, ISO 27K,Octave, Fair - Assessing and improving a Security System - Defense in depth plan after gap analysis - Evaluation based on Common Vulnerabilities **Midterm Assignment- Assessment Project** *[14], [10], [11], [17], [36], [38], [20] to [35]*
Week 8 Process Improvement	SDL Standardization, Process Capability Maturity Model CMMI, Assurance Maturity Model**,** Security Metrics: *[9], [40], [27], [33]*
Week 9 System assurance review	**- Safety Assurance,** graduate students only - Safety/Hazard Assurance **HW2 Assignment- FMEA for Software** - Preparing and reviewing Assurance Cases: *[37], [2]*
Module 3 Secure coding	Secure application coding in a networked environment, coding standards and common vulnerabilities:
Week 10 Designing security	Principles: Compartmentalization, Reducing complexity, Redundancy, Auditing, and Coding standardization: Java, C/C++ *[12], [13] and numerous industry standards and guidelines (supported by MITRE, CERT, OSWAP, etc.)*
Week 11 Protecting access	- OS access control and least privilege - Validating input
Weeks 12, 13, and 14 Preventing common attacks	- Buffer, integer and float overflows -- Error trapping (C, C++, C#, Java) - Cross-site scripting - SQL Injection - Other problems and remedies such as: race conditions, serialization, software integrity, sandbox, etc. plus using tools to detect flaws;
Week 15	Security Code Review and Security Testing: *[12], [13], [16], [18]*
Module 4 In parallel with Weeks 7-13	Security related research in Software and/or Data Intensive Systems culminating a paper of publishable quality (following IEEE/ACM conference format) with a formal peer review. *Additional references per topic.*
Week 16	**Final Evaluation** Includes: term project presentation, secure coding project review, research paper peer review, and optional mock-up certification exam for selected relevant domains from the CISSP/CSSLP/ISSEP.

II. CONLUSION

This paper presents an outline of a course in Systems and Software Assurance (SSA) together with extensive set of references that may be of use to others interested in similar security course for software system developers. It is our hope that educators may be interested in exchanging ideas, experiences, and relevant course materials. New detailed course outcomes, based on [41] are in process of consolidation, and mapping to [45] as well as expectations per [44], and will be available in May 2016. Most of the course content and material for the SSA course was redesigned, key of which was adding security architecture analysis with treat modeling and secure coding practices, and eliminating network evaluation as it is covered in the required Computer Security course. The modernized SSA course is being offered, as a pilot, in the spring 2016. For illustration, the Appendix shows models from the HW1. The initial reaction by a group of 20 undergraduate CS students is positive. Students appreciate the hands on model/code focus, both breadth and depth of the course, and especially the fact that course is being delivered by two experienced professors covering respectively systems and code level issues. The architecture analysis was considered the most elusive (in both variety of system types and multitude of security patterns [3], [4], [19], [5]) as the volume of material, especially [24], [25], [26], [29], [33], was at times overwhelming. The recommended textbooks [17], and [12] were helpful and additional examples and guidance regarding in depth architectural analysis and treat modeling in [3] and [4] are also quite effective. A related graduate cyber security concentration, within the Master of Science in CS, is proposed that will share, as a core, our SSA course, include additional required coursework from CS (Applied Cryptography, Network and Computer Security, Data Systems Security, and the Ethics and Research in CS), and shared elective coursework with the IT and the EE departments (Cloud Security, Digital Forensic, and Physical Layer Security).

REFERENCES

[1] A. Conclin, D. Shoemaker, CSSLP Certification All-in-One Exam Guide, McGraw Hill 2014.

[2] A. Neufelder, Effective Application of Software Failure Modes Effects Analysis, SoftRel 2014.

[3] A. Shostak, Treat Modeling: Design for Security, J. Wiley 2014.

[4] B. Schoenfield, Securing Systems- Applied Security Architecture and Treat Models, CRC Press 2015.

[5] C. Berg, High-Assurance Design- Architecting Secure and Reliable Enterpsise Applications, Addison Wesley 2006.

[6] CNSS 4014, National Information Assurnace (IA) Training Standard for Information Systems Security Officers, 2004.

[7] CNSS 4016, National Information Assurance (IA) Training Standard for Risk Analysts, 2005.

[8] DoD 8570.01-M, Information Assurance Workforce Improvement Program, updated 11.10.2015

[9] D. Shoemaker, V. Jovanovic, Engineering a better software organization, Quest 2001.

[10] FIPS 199, Standards for Security Categorization of Federal Information and Information Systems,

[11] FIPS 200, Minimum Security Requirements for Federal Information and Information Systems, 2008.

[12] F. Long, D. Mohindra, R. Seacord, D. Sutherland, D. Svoboda, Java Coding Guidelines, Addison Wesley 2014.

[13] G. McGraw, Software Security – Building Security In, Addison Wesley 2006,

[14] G. Mills, Security Assessment Using NSA IAM, Syngress 2004.

[15] I. Williams, X. Yuan, Evaluating Efectiveness of Microsoft Treat Modeling Tool, ISCD Conference 2015, Kennesaw GA, October 2015.

[16] J. Ransome, A. Misra, Core Software Security, CRC Press 2014.

[17] L. Johnson, Security Controls, Evaluation, Testing, and Assessment Handbook, Elsevier 2016.

[18] M. Merkow, L. Raghavan, Secure and Resilient Software, CRC 2012.

[19] M. Schumacher, E. Buglioni, D. Henderson, F. Buchman, P. Somerland, Security Patterns- Integrating Security and System Engineering, J. Wiley 2006.

[20] NIST 800-18, Guide for Developing Security Plans for Federal Informatin Systems, 2006.

[21] NIST 800-27rA, Enginering Principles for Information Security (A Baseline for Achieving Security), Revision A, 2004.

[22] NIST 800-30r1, Guide for Conducting Risk Assessments, 2012.

[23] NIST 800-37r1, Guide for Applying the Risk management Framework to Federal Information Systems, 2010.

[24] NIST 800-39, Managing Information Security Risk, 2011.

[25] NIST 800-53r4, Security and Privacy Contriols for Federal Informatin Systems and Organizations 2013.

[26] NIST 800-53Ar4, Assessing Security and Privacy Controls in federal Informatin Systems and Organizations, 2014.

[27] NIST 800-55, Perfomance Measurement Guide for Informatin Securitty, 2008.

[28] NIST 800-60 V1, r1, Guide for Mapping Types of Information and Information Systems to Security Categories, 2008.

[29] NIST 800-64r2, Security Considerations in the System Development Life Cycle, 2008.

[30] NIST 800-115, Technical Guide to Information Security Testing and Assessment,

[31] NIST 800-137, Information Security Continuous Montiroring (ISCM) for Federal Information Systems and Organizations, 2011.

[32] NIST 800-144, Guidelins on Securty and Privacy in Public Cloud Computing, 2011.

[33] NISTIR 8011 Volume 1, Automation Support for Security Control Assessment, Draft 2016.

[34] NIST 800-160, Systems Security Engineering, 2014.

[35] NIST 800-171, Protecting Controlled Unclassified Information in Nonfederal Information Systems and Organizations, 2015.

[36] M. Talabis, J. Martin, Information Security Risk Assessment Toolkit, Elsevier 2013.

[37] N. Mansurov, D. Campara, System Assurnace- Beyond Vulnerabilities, MK 2010.

[38] R. Anderson, Security Engineering, 2ed, John Wiley 2008.

[39] R. Krutz, R. Vince, The CISSP and CAP Preparation Guide, John Wiley 2007.

[40] V. Jovanovic, D. Shoemaker, ISO 9001 standard and software quality improvement, Benchmarking for Quality Management & Technology, Vol. 4, No 2 1999, MCB University Press, pp 148-160

[41] V. Jovanovic, Systems Assurance Standards and Processes, InfoSec Curriculum Development Conference 2008, September 2008, Kennesaw, GA, pp 148-159.

[42] V. Jovanovic, T. Mirzoev, Teachiong network storage technology-assessment outcomes and directions, ACM SIGITE'08, October 2008, Cincinnati, OH, pp 147-151.

[43] V. Jovanovic, T.Mirzoev, Teachnig storage infrastucture management and security, InfoSec Curriculum Development Conference, October 2012, Kennesaw, GA, pp 41-43.

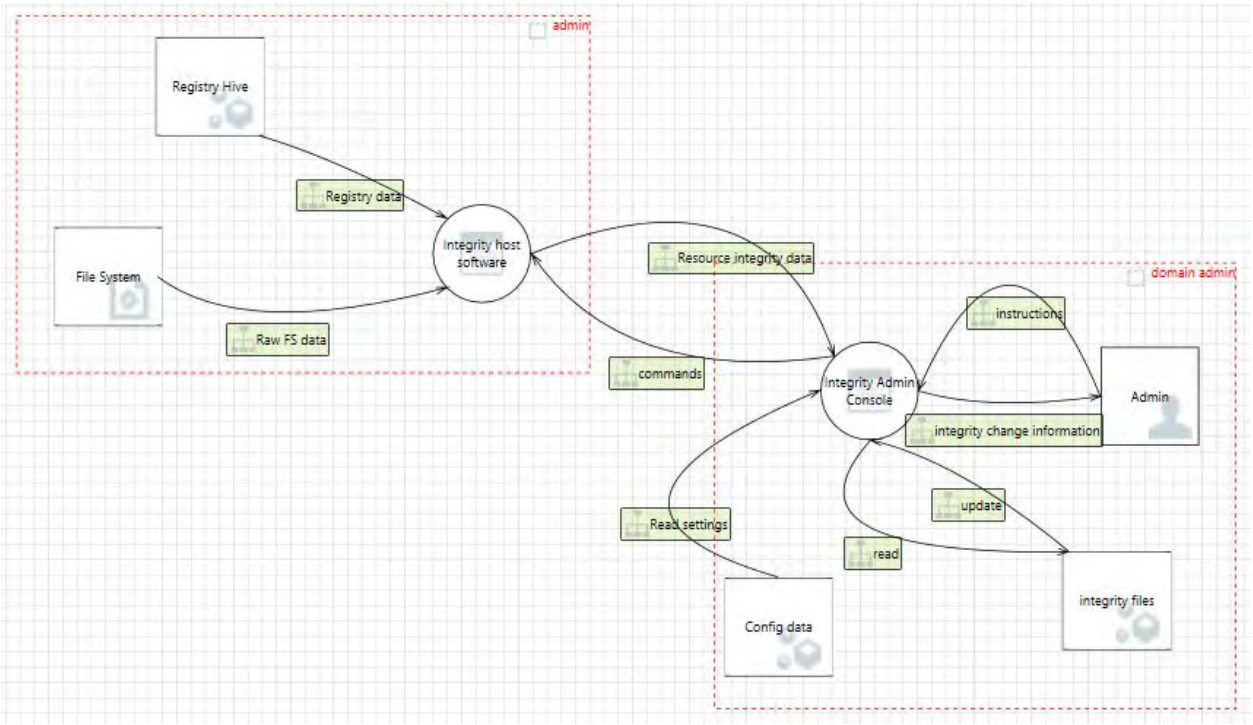
[44] V. Jovanovic, Computer Science Curriculum 2013- An Example, MIPRO 2013, Opatia, Croatia, pp 759- 764.

[45] WorkForce V2, 2015, retrived Febraury 7. 2016 from the: https://niccs.us-cert.gov/training/tc/framework/specialty-ares.

APPENDIX

For illustrative purposes the following details, from a first homework regarding treat modeling are included.

Model A:



STRIDE Threat List and Controls:

Spoofing:

- Incorrect file data could be fed into the software, Software spoofed to feed data into admin console
- Admin console spoofed to access the software
- Incorrect config data could be fed into the admin console, incorrect files could be fed into the admin console
- Admin console spoofed to update the files to be incorrect
- Fake admin could spoof the admin console, Incorrect registry could be fed into the software

Control: More secure authentication

Tampering:

- Connection between software and admin console could be intercepted and data could be tampered with

Control: Verify data on route

Repudiation:

- Outside connections to and from software and admin console could result in a loss

Control: Monitor incoming and outgoing connections

Information Disclosure:

- Unauthorized data flow could tamper with the software and admin console
- Connection between software and admin console could be sniffed by attacker

Control: More secure authorization and better encryption

Denial of Service (DoS):

- Connection between software and admin console could be fed unwanted data leading to DoS

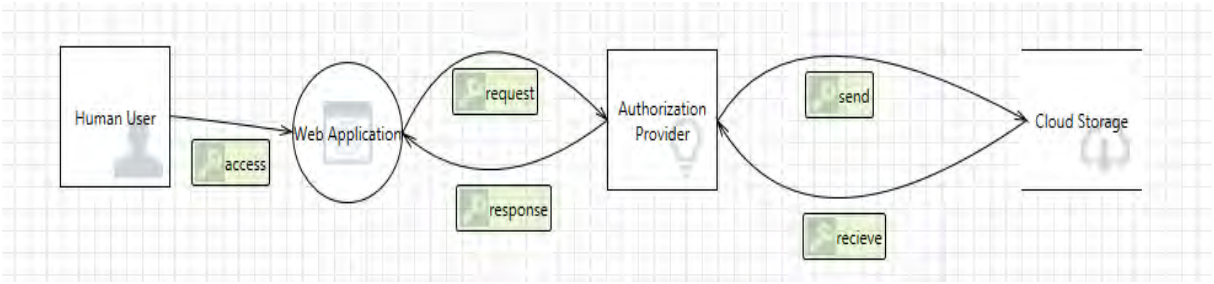
Control: Higher protections from DoS

Elevation of Privilege:

- User can gain access to admin console by posing as an admin
- Admin console and software user could feed data into it from an outside location

Control: More secure authorization for users and administrators

Model B:



Threats:	Controls:
attacker takes advantage of custom encryption	make sure encryption is secure
attacker alters data in storage	have storage monitors and backups
attacker injects code into process	security measures to prevent injecting
attacker writes to storage	security measures on storage units and backups
attacker changes trust parameters	secure authorization on parameters
attacker has access to write to a resource	secure authorization on resources
attacker gains access to account information	secure server and encryption to hold user data
attacker connects to server	secure authorized access to servers
attacker gets access through a port used by the system	have port monitoring and alert when unknown address gets through
attacker uses stolen credentials to access system	secure authorization on system
attacker tries multiple credentials over and over again	lock-out account upon multiple unsuccessful logins
attacker confuses client because of many ways to access server	find more secure way to access server
attacker spoofs server	have authorization codes attached to server that clients connect to
system has default password with no change	have mandatory password change
attacker reads unencrypted files	secure encryption on files
attacker changes data en route to location	secure connection between locations and track the data as it travels
attacker causes dos	identify and cut off the source of the attack
attacker makes server unavailable	have backup servers running
attacker makes client unavailable	
attacker forces data through different paths which gives different results	implement all path security testing before releasing software
attacker provides pointer across trust boundary	implement bounds checking in software development

Analysis of Learning Management Systems features and future development challenges in modern cloud environment

*H. Jerković, *P. Vranešić, **A. Radan
Zagreb School of Economics and Management,
*IT department, Zagreb, Croatia,
**MBA graduate student,
hrvoje, philip, anja.radan @zsem.hr

Abstract - Learning Management Systems (LMS) are commonly used for management of online learning activities inside single or many courses. Usually they are rich with different tools and options in order to enable modern interactive learning environment. But from their early beginnings, before migration on web, LMS were designed as closed systems in order to serve single institution learning content. Such development hindered LMS for proper integrations with other systems and brought many challenges which are present up to this date in many LMS solutions. This research explains why some of the most used LMS in the world today are not able to properly integrate in modern cloud environment. Analysis of core functionalities of most used LMS are explained with propositions for improvement and future development.

I. INTRODUCTION

Sharing and collaboration on creating knowledge should be main feature in any e-learning system. Unfortunately those activities are frequently being hindered by systems which supposed to support them. LMS and similar e-learning systems are designed to structure learning procedures, tools and information but they are often not enabling effective sharing and creation of content in easy to use collaborative environment. Educators are reaching for other services and solutions in order to create, share and collaborate on creation of educational content which, on the end, is being offered to students inside LMS.

This paper explains why LMS are inherently closed systems and proposes open architecture for sharing and reusing learning objects (LO) with explanation of main benefits of such approach. Integration capabilities and challenges for successful integration of current LMS solutions are analyzed from perspective of two most commonly used LMS solutions - Blackboard and Moodle. General proposition for LMS open architecture is given with explanations of communication inside proposed learning network. Special attention is given to analysis and proposition of LO since they are in the core of learning network. In the end, this research brings possible future development based on proposed architecture.

II. CURRENT RESEARCH

Researchers exploring possible integrations of various learning systems and learning repositories have successfully created different solutions which are implemented in government and e-learning networks. Those repositories and services allows creation of easily accessible resources [1]. Extensively mentioned OER (Open Educational Resources) are in focus of various other researchers with aim to facilitate integration of learning resources [2]. Therefore similar solutions for integrated LO search tools or networks already exist as demonstrated in [3], [4], [5], [6] and [7]. However, existing federated LO search engines produced through such research are rarely used. Educators are mainly using dedicated individual sources like professional literature resources with readymade materials or LO which are available through various OER portals. Poor usage comes from fact that majority of high quality LOs are not present in OERs or any other learning repositories but inside LMS. Other researchers explored options for personalization of learning objects in virtual learning environments [7] or gave practice-oriented review of LO which emphasizes need for better management of LO in any learning information system [8]. But majority of research is focused on integration problems in some limited environment [9] or on need for upgrading LMS as individual system [7] without taking into consideration global scope of the problem.

III. ASSESMENT OF CURRENT LMS INTEGRATION CAPABILITIES

LMS integration services should allow educators and students to collaborate not only on institutional level but also with other institutions. LMS integration with other systems like SIS (Student Information System) has been standard way of integration. It is usually fairly limited type of integration for accessing basic information in SIS (about students, professors, departments, courses) in order to replicate same state inside LMS. Blackboard and Moodle, as most used LMS, are following IMS Global standard for LTI (Learning Tools Interoperability [10]) in order to integrate content from providers like McGraw and Pearson [11] or to allow access to full courses and activities from

remote systems (other LMS installations or any other LMS LTI consumer compliant software). LTI is standard way of integrating rich learning applications (often remotely hosted and provided through third-party services) with platforms like LMS, web portals, or other educational information systems [12]. This is also powerful way to integrate LMS with cloud services in order to achieve collaborative and interactive features that often lack inside LMS environment. Integration with Google Docs is good example of that [13]. API (Application Programming Interface) on LMS is offering access to content and tools inside LMS environment and API is usually used to create plugins (in Moodle [14]) or building blocks (in Blackboard [15]) environment which are extending standard functionalities. These are usually ways LMS can be extended but this does not mean they are offering ideal modern e-learning environment. We must keep on mind that e-learning is continuously evolving and learning systems must be ready to integrate new paradigms that consider student as the center of the learning process [16, 17]. This shift will mean changing the tools currently used and giving way to new tools for easy integration of new services and resources which are frequently emerging [18]. Web carries large variety of new kinds of tools and services and many educators are adopting new approach from the closed LMS course space towards Personal Learning Environments (PLE) [19]. PLE's are characterized by its absence of structure, just what is provided by open standards and mashup techniques which are becoming more and more important because they allow effective integration of content or service in modern web environment [20]. Unfortunately integration options like that are either not supported or they are very limited in current mainstream LMS. LMS APIs that should enable described integration are falling short in enabling such dynamic environment inside LMS courses.

A. Metadata problems

Valuable resources like LO have to be discoverable in networked educational environment and with proper metadata automatically defined. Unfortunately LMS is not storing LO with metadata options and usually they have quite week capabilities regarding metadata management in general. Content features do allow easy input of data on course, for example with Moodle you can easily just drag and drop new content in certain content holder areas inside courses. Question is – how is document metadata stored? In usual scenario users will not have option for filling metadata fields and there is no automated mechanisms that are doing that instead of user. Convenient circumstance is that LMS already holds majority of information which can be assigned as metadata or it can be easily fetched from SIS. Problem is that there is no usual workflow that implements this important aspect [21]. Metadata already present in LMS is:

- User information about teachers and students – fetch from authorization system or SIS of the institution,
- Course and department data – is already stored in structure of the LMS while courses are opened, structured and classified.

With such a rich sources of metadata any LO generated in LMS environment could be described with all information that can be fetched from LMS or any other institutional system. It's worth noting that SaaS (Software as a service) LMS solutions are becoming more and more popular [22, 23]. They are usually modern cloud based software created on SOA (Service Oriented Architecture) principals and therefore very flexible, easily scalable and customizable with better pricing systems which is appealing to management of institutions [24]. SaaS LMS features SIS integration but custom integrations are not fully supported [25].

IV. CHALLENGES FOR SUCCESSFUL LMS-CLOUD INTEGRATION

Many proposals tried to demonstrate integration of learning activities in cloud environment [26]. Main challenges in the way are:

- Successful integration with existing learning and other supporting information systems.

We have seen proposed architectures that are giving complete road map on how to migrate existing services into cloud environment but transition in reality cannot easily follow that pattern. There are many legacy systems (not only LMS) that needs to be taken into consideration. SIS is starting point of integration - allowing authorization and data acquisition from cloud applications as well as better integration with LMS.

- Supportive cloud environment with adept tools for educational purposes of different types.

Different cloud platforms are recently launched with "education" emphasis [27]. They usually provide good foundation for creation of new generation of educational cloud content services because they provide stable storage with powerful collaboration features, event management features, email integration on institution level and similar services. What is missing are instructions, case scenarios, tools, additional APIs and examples on how to build and integrate cloud applications with different types of existing standardized infrastructures; LMS and SIS in first place.

- Advanced adaptive schemas for semantics and ontology.

Educational services on the web use different ways of naming various types of data for internal handling and also for communications with other systems. Some are standardized according to SCORM [28] but in many cases situation is still quite challenging since there is no standardized procedure for all activities [29] in formal or informal learning on the web. Therefore problems are present in service communication with any standardized LMS. Procedures, data types and APIs should be defined with agreements between key organizations on how to implement them.

- Tracking, assessment and collaboration.

Implementing ways to assess and track learners' progress during formal and informal learning in PLE is imperative. There has been good attempts recently to assess those activities [30] but we are far away from standardized procedures. We need to be able to preserve learners' privacy and still track activities across web with customizable tools for measuring quantity and quality of learners' contributions across wide variety of tools and services on the web and not only inside LSM. Learning is

happening more frequently on the web than in LMS. Measurement are not enough in this situation since we also need ways to link them with rubrics and assessment procedures in LMS that will allow us to give students feedback. Gamification tools that recently emerged in LMS tools are promising but much needs to be done so we could effectively integrate them with third party tools [31].

V. PROPOSITION FOR LMS OPEN ARCHITECTURE

There is obvious need for better collaboration among institutions for enabling support for teachers' and students' needs in learning environment. Open architecture for connected learning has to support unprecedented agility, flexibility, and personalization in order to fulfil those requirements. Agility, in the IT context, means a focus on applications so that instructors can mix and match tools from a wide array of sources (i.e., traditional vendors, instructors, institutions, open sources) to best support their pedagogical goals [32]. Figure 1. shows schema for institutional collaboration integrated in existing LMS environment. Idea is not to abandon currently used LMS but enrich them with collaborative options which educators from various institutions can use in order to create, share and reuse educational material. LMS directory service could enable discovery of learning materials and activities across many institutions.

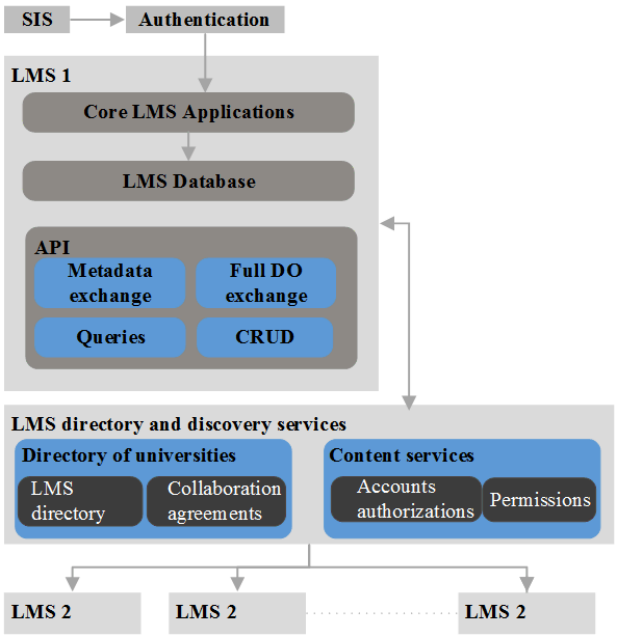


Figure 1. Simplified schema of Open Architecture for LMS network

Based on their learning affiliation, content and activities, institutions could automatically initiate collaboration intention and simplify various agreements based on them. After agreement has been officially signed collaboration can start. So this will enable customizable network of resources built with affiliated partners. This proposed solution brings down many barriers that educators are facing today and could enable richer learning experience for students and educators as well [33].

VI. LMS COMMUNICATION INSIDE PROPOSED LEARNING NETWORK

In order to facilitate communication with proposed LMS learning network, institutions has to be able to discover, describe and open their own resources to learning network in controlled way. We have to keep on mind that any scientific publication system or library information system could as well be part of this network. Figure 2. shows proposed Learning Directory Service for harvesting and discovery of learning resources on own private network with primary focus on LMS and other internal information systems with educational content and services. Such service also has to be able to integrate 3rd party content and service from other LMS or other external repositories.

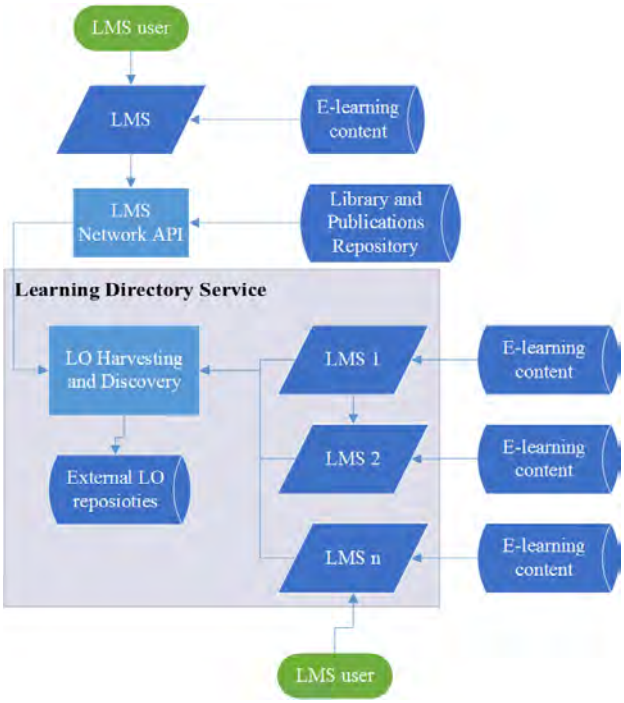


Figure 2. Learning Directory Service

In extended form, this service should be able to expand own API in order to successful integrate various types of services and content providers. But in that case data model should be very flexible and in same time no information should be omitted while communicating with rest of the learning network. Idea is to integrate heterogeneous sources and properly enable technology enhanced learning ecosystem [34]. In order to achieve that we need far better architecture of LO.

VII. LEARNING OBJECT AS CORE OF SERVICE ORIENTED E-ARCHITECTURE

Proposed data model of LO is in the heart of architecture shown on Figure 3. Figure shows proposed Extended Learning Object Architecture. LO core components define its ID, properties, description, relations to other object (in case if it's part of some complex object), his internal structure (in case it is itself complex object), API implementations automatically defined by type of resources that object is consisted of, core datastreams (local or remote) and places where object is integrated

(information systems, local or remote which are using object). Such LO structure enables sustainability, reusability, tracking and versioning supported by corresponding APIs. Also any missing service could be easily integrated since custom API for that specific LO could be easily build. Any object editing could create new versions and any citation or reusing of object would create new record in objects Resource Map [35].

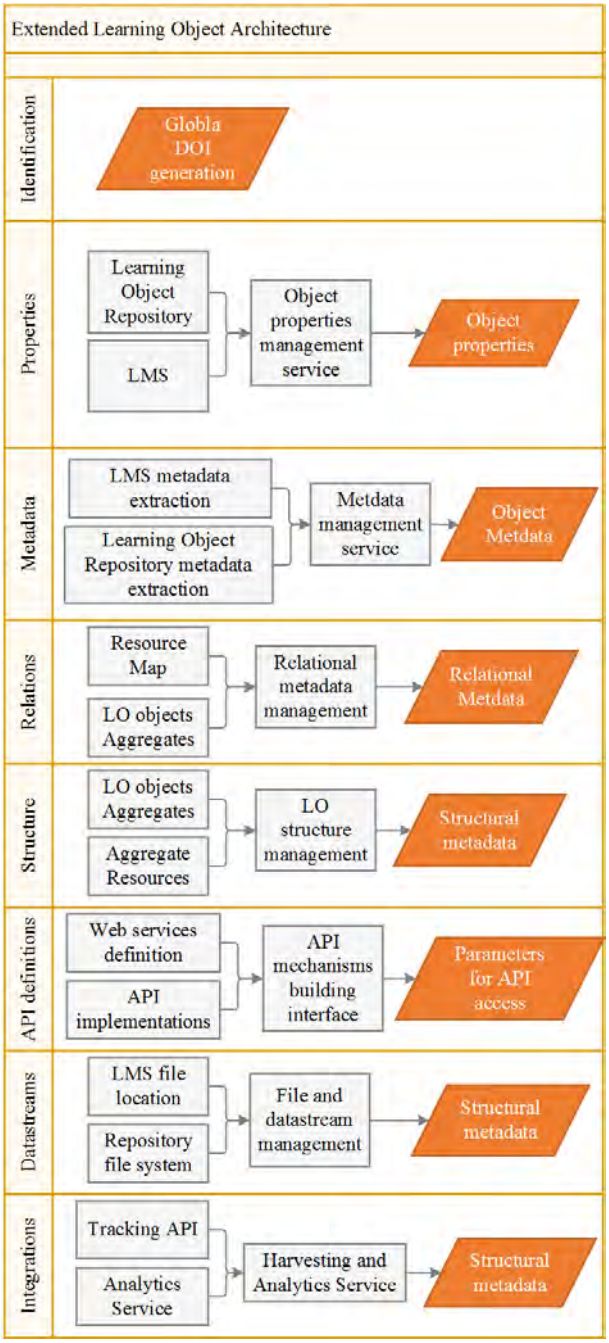


Figure 3. Schematics of proposed Extended Learning Object Architecture

The Service Oriented Architecture approach is applied in form of application of protocols like OAI-ORE (Open Archives Initiative Object Reuse and Exchange) which gives thorough guidelines on how to enable such proposition in own environment [36, 37]. Just like every SCORM compatible LO could be reused in LMS SCORM supporting system in same way proposed extended LO could be reused everywhere through its extensive APIs (including SCORM supported LMSs).

VIII. Discussion on Future Development Possibilities of Proposed Architecture

Purpose of proposed framework would be enabling of advanced features and integrating them in existing learning environment. Examples of new features are:

- enhanced application for collaboration

This would enable creation of customized web environments. Specialized collaboration apps could allow researchers and students collaboration while writing: essays, project, and scientific articles from online apps which would be integrated with enhanced learning content tools and references. Users would be able to collaborate on joint projects inside its own cloud environment or in collaboration with other educational institutions. Integration for other recognized third party service (e.g. Google Docs through Google Docs API) would be enabled with various options like tracking of user contributions, versioning and content integration on courses.

- enhanced existing learning content tools and references

Bookmarked sites, custom libraries, annotations, commenting and highlighting tools, custom web scrapping and other web objects, personal learning objects library etc. , are all cloud tools and services that could be integrated with cloud core storage engine together with custom search. This would allow learners to search their storage, their college resource, college partners' resources and also school library and scientific databases on which school bought access - all from single location.

- educational internal social network

Service integration would allow creation of social-like news feeds where students could see contributions of their project groups or class participants in internal or external learning apps from their own institution or with users from other institutions. This would be enabled since this approach would literally wipe LMS borders allowing direct communication among groups. Instead of just programmatically support various platforms, web services mechanism builder (under API definitions) from Figure 3. could provide interface for less knowledgeable users for visual creation of applications. Apps could harvest types of data records from knows services and then allow users to build custom applications with available building blocks and integrate in them various data sources (users and objects from other LMS or Repository services).

A. Example of learning activity in proposed system

In this chapter we'll give example how proposed architecture could be used to asses learners' informal activities on the web and integrated it in LMS.

Computer science students are learning about web server administration. Group have to set up their own web server with custom features and create comparison between few open source applications using open source

performance benchmark testing tools. While they run into several problem professor is giving them just few guidelines and directing them to post question on Stack Exchange portal [38]. They start using schools LMS account and log with it on Stack Exchange portal. Professor initiates cloud application for Stack Exchange inside LMS which allows professor tracking of students' contributions on Stack Exchange from LMS. Users' contributions and results of tested solutions are recorded as well as user communication.

In this case no special application permissions are required since application only needs to harvest data which is already publicly available on the web site of the service. Additional service is needed to support representation of students' activities in overall grading schema in LMS. This is just matter integrating cloud services with its corresponding building block. That part could be easily developed since interface for such applications is supported in standard way in all major LMS since same interface is used for building plugins in LMS.

This is just one simple lesson in which third party professional questions and answers portal is used for learning activity and in same time students are using professional services for building their reputation in professional environment.

CONCLUSION

Modern cloud environment in combination with SOA principles and proper cross repository data model for e-learning offer new and exciting options for educators. Many project in area offered similar solutions but key for future success is agreement of various parties. E-learning standard makers should be able to understand problematics of closed LMS and needs of users which are frequently frustrated with LMSs' closed environment. Future development should in first place work on bringing together e-learning standard makers, LMS producers and content and repository providers. Goal is to bridge gaps between institutional collaboration and define standardized ways for sharing, creating and reusing e-learning content.

Besides, as we mentioned before, learning is frequently happening outside LMS environment and teaching stuff has no control over it. Far better mentoring procedures could be created if teaching staff had better control over informal learning of their own students.

Migration from traditional platforms is happening slowly but biggest problems are not migration in cloud itself, which is already under way, but effectiveness of integrations of learning activities in cloud environment. Cloud environment offers many advantages but if used with inflexible traditional mind-set it will only be recreation of linear activities taken from LMS to cloud. This will not significantly enrich current learning activities. Advancements in that part greatly depend on teaching staff and institutional stand towards such approach.

We have to truly grasp idea that student's learning already did become ubiquitous and then we'll be able to support it effectively and build form there onwards.

REFERENCES

[1] L. Carvalho and P. Goodyear, *The architecture of productive learning networks*: Routledge, 2014.

[2] N. Butcher, *A basic guide to open educational resources (OER)*: Commonwealth of Learning, Vancouver and UNESCO, 2015.

[3] S. Pallottelli, S. Tasso, M. Rui, A. Laganà, and I. Kozaris, "Exchange of Learning Objects Between a Learning Management System and a Federation of Science Distributed Repositories," in *Computational Science and Its Applications--ICCSA 2015*, ed: Springer, 2015, pp. 371-383.

[4] F. De la Prieta, A. B. Gil, A. J. S. Martín, and C. Zato, "Learning Object Repositories with Federated Searcher over the Cloud," in *Methodologies and Intelligent Systems for Technology Enhanced Learning*, ed: Springer, 2014, pp. 93-100.

[5] H. Spoelstra, P. Van Rosmalen, and P. B. Sloep, "Toward Project-based Learning and Team Formation in Open Learning Environments," *J. UCS,* vol. 20, pp. 57-76, 2014.

[6] M. J. Hannafin, J. R. Hill, S. M. Land, and E. Lee, "Student-centered, open learning environments: research, theory, and practice," in *Handbook of research on educational communications and technology*, ed: Springer, 2014, pp. 641-651.

[7] M. Á. Conde, F. J. García-Peñalvo, M. J. Rodríguez-Conde, M. Alier, M. J. Casany, and J. Piguillem, "An evolving Learning Management System for new educational environments using 2.0 tools," *Interactive Learning Environments,* vol. 22, pp. 188-204, 2014.

[8] J. Sinclair, M. Joy, J. Y.-K. Yau, and S. Hagan, "A practice-oriented review of learning objects," *Learning Technologies, IEEE Transactions on,* vol. 6, pp. 177-192, 2013.

[9] N. Tserendorj, U. Tudevdagva, and A. Heller, *Integration of Learning Management System into University-level Teaching and Learning*: Techn. Univ., Fak. für Informatik, 2013.

[10] I. G. L. Consortium, "Inc.,(2013), IMS Learning Tools Interoperability," ed, 2013.

[11] I. Waßmann, C. Schönfeldt, and D. Tavangarian, "WIKI-LEARNIA: SOCIAL E-LEARNING IN A WEB 3.0 ENVIRONMENT," *Engineering Sciences & Technologies/Nauki Inzynierskie i Technologie,* vol. 4, 2014.

[12] C. Severance, T. Hanss, and J. Hardin, "Ims learning tools interoperability: Enabling a mash-up approach to teaching and learning tools," *Technology, Instruction, Cognition and Learning,* vol. 7, pp. 245-262, 2010.

[13] J. Tsay, L. Dabbish, and J. Herbsleb, "Influence of social and technical factors for evaluating contribution in github," in *Proceedings of the 36th international conference on Software engineering*, 2014, pp. 356-366.

[14] I. A. Kautsar, S.-I. Kubota, Y. Musashi, and K. Sugitani, "Developing Moodle plugin for creating learning content with another REST function call," in *Global Engineering Education Conference (EDUCON), 2014 IEEE*, 2014, pp. 784-787.

[15] F. Huertas and A. Navarro, "Integration mechanisms in e-learning platforms," *International Journal of Computer Information Systems and Industrial Management Applications,* vol. 5, pp. 714-721, 2013.

[16] A. Butt, "Student views on the use of a flipped classroom approach: Evidence from Australia," *Business Education & Accreditation,* vol. 6, p. 33, 2014.

[17] M. Baeten, K. Struyven, and F. Dochy, "Student-centred teaching methods: Can they optimise students' approaches to learning in professional higher education?," *Studies in Educational Evaluation,* vol. 39, pp. 14-22, 2013.

[18] R. Jorge, "The Combined Effect of Problem-Based Learning and the Flipped Classroom," 2014.

[19] N. Dabbagh and A. Kitsantas, "Personal Learning Environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning," *The Internet and higher education,* vol. 15, pp. 3-8, 2012.

[20] S. Roy, A. De Sarkar, and N. Mukherjee, "An Agent Based E-Learning Framework for Grid Environment," in *E-Learning Paradigms and Applications*, ed: Springer, 2014, pp. 121-144.

[21] C. Batanero, S. Otón, and J. Holvikivi, "Considering Student Personal Needs and Preferences and Accessible Learning Objects to Adapt Moodle Learning Platform," 2014.

[22] R. Gurunath and R. A. Kumar, "SAAS EXPLOSION LEADING TO A NEW PHASE OF A LEARNING MANAGEMENT SYSTEM," *International Journal of Current Research and Review,* vol. 7, p. 62, 2015.

[23] M. Despotović-Zrakić, K. Simić, A. Labus, A. Milić, and B. Jovanić, "Scaffolding Environment for e-Learning through Cloud Computing," *Educ. Technol. Soc,* vol. 16, pp. 301-314, 2013.

[24] S. Ros, R. Hernandez, A. Robles-Gomez, A. C. Caminero, L. Tobarra, and E. S. Ruiz, "Open service-oriented platforms for personal learning environments," *Internet Computing, IEEE,* vol. 17, pp. 26-31, 2013.

[25] C. Milligan, A. Littlejohn, and A. Margaryan, "Patterns of engagement in connectivist MOOCs," *MERLOT Journal of Online Learning and Teaching,* vol. 9, 2013.

[26] Y.-T. Lin, M.-L. Wen, M. Jou, and D.-W. Wu, "A cloud-based learning environment for developing student reflection abilities," *Computers in Human Behavior,* vol. 32, pp. 244-252, 2014.

[27] V. Stantchev, R. Colomo-Palacios, P. Soto-Acosta, and S. Misra, "Learning management systems and cloud file hosting services: A study on students' acceptance," *Computers in Human Behavior,* vol. 31, pp. 612-619, 2014.

[28] R. P. D. Redondo, A. F. Vilas, J. J. P. Arias, A. G. Solla, M. R. Cabrer, and J. G. Duque, "SCORM and Social Recommendation: A Web 2.0 Approach to E-learning," in *Recommender Systems for the Social Web*, ed: Springer, 2012, pp. 195-207.

[29] K. Pireva, P. Kefalas, D. Dranidis, T. Hatziapostolou, and A. Cowling, "Cloud e-Learning: A New Challenge for Multi-Agent Systems," in *Agent and Multi-Agent Systems: Technologies and Applications*, ed: Springer, 2014, pp. 277-287.

[30] S. Boudebza, L. Berkani, F. Azouaou, and O. Nouali, "Using an ontological and Rule-based approach for contextual semantic annotations in online comMunities," in *E-Learning Systems, Environments and Approaches*, ed: Springer, 2015, pp. 95-115.

[31] J. Broer and A. Breiter, "Potentials of Gamification in Learning Management Systems: A Qualitative Evaluation," in *Design for Teaching and Learning in a Networked World*, ed: Springer, 2015, pp. 389-394.

[32] R. Abel, M. Brown, and J. Suess, "A New Architecture for Learning," *Educause Review,* vol. 48, p. 88, 2013.

[33] H. Jerkovic, "Prototype framework for integration of digital repository systems with e-learning platforms."

[34] A. Fernández, M. Erdt, I. Dackiewicz, and C. Rensing, "Recommendations from heterogeneous sources in a technology enhanced learning ecosystem," in *Recommender Systems for Technology Enhanced Learning*, ed: Springer, 2014, pp. 251-265.

[35] A. Isaac and B. Haslhofer, "Europeana linked open data–data. europeana. eu," *Semantic Web,* vol. 4, pp. 291-297, 2013.

[36] L. K. Soo, E.-T. Yeoh, and S. B. Ho, "Ontology Based Personalized Recommendation Model for Learning Objects in a Service Oriented E-learning Environment," in *ICIIM: International Conference on Innovation and Information Management*, 2012, pp. 87-92.

[37] P. Manghi, M. Artini, C. Atzori, A. Bardi, A. Mannocci, S. La Bruzzo*, et al.*, "The D-NET software toolkit: A framework for the realization, maintenance, and operation of aggregative infrastructures," *Program,* vol. 48, pp. 322-354, 2014.

[38] N. Ostashewski and D. Reid, "A History and Frameworks of Digital Badges in Education," in *Gamification in Education and Business*, ed: Springer, 2015, pp. 187-200.

MARKOV MODEL OF MATHEMATICAL COMPETENCES IN ELEMENTARY EDUCATION

Gordana Paić[1], Božidar Tepeš[2], Krešimir Pavlina[3]

1. Gordana Paić prof., Elementary School Dr. Ivan Merz, Zagreb, Croatia, Račkoga 4, gogica.paic@gmail.com
2. Prof. dr. sc. Božidar Tepeš, Faculty of Teacher Education, University of Zagreb, Croatia, Savska 77, tepesbozidar@gmail.hr
3. Prof. dr. sc. Krešimir Pavlina, Faculty of Humanities and Social Sciences, University of Zagreb, Department of Information and Communication Sciences, Croatia, Ivana Lučića 3, kpavlina@ffzg.hr

ABSTRACT

In this paper, the authors defined mathematical competences for the 7th grade of elementary school. The basic objective was to measure the mathematical competences or mathematical knowledge, skills and abilities in mathematical education. Mathematical competences were grouped in following areas: Algebra and functions, Space and shapes and Measurements and Data. Statistical set for the research consisted of 48 children from the Elementary school Dr. Ivan Merz in Zagreb, Croatia. Authors measured 15 variables together with the evaluated results of described tasks. Using statistical set of variables which measured mathematical competences the authors created representation of distribution by the Markov network (undirected graph with information about conditional independences between mathematical competences). Statistical analysis were conducted using software Tetrad and Statistica. In the results authors describe Markov model between the mathematical competences for 7th grade elementary school children. This paper is result of our previous research on causal modelling of mathematical competences in kindergarten and in elementary school.

***Keywords*:** mathematical competences, Markov network and conditional independence

I. MATHEMATICAL COMPETENCE IN ELEMENTARY SCHOOL EDUCATION

Mathematical competence as a term is beeing defined by knowledge, ability and skills of applying the mathematical way of thinking or concluding with the purpose of solving all kinds of mathematical or interdiciplinary problems and also understanding mathematics as a cultural value, with the aim of understanding and creating the perception about things that surround us in our everyday existence. It is important to emphasize that intuitive ways of solving tasks are supported by rules of logical opinion in the form of: thinking and making adequate conclusions, using arguments, modeling, formulating, and solving the problem.

Mathematical competences are defined by lesson plan or curriculum with the acceptance of historical development and can be divided into four basic areas: Algebra and functions, Form and shapes, Measurements and Data. Every area mentioned contains competences specific content about knowledge, ability and skills which should be presented.

Mathematical competences are defined by lesson plan or curriculum with the acceptance of historical development and can be divided into four basic areas: Algebra and functions, Form and shapes, Measurements and Data. Every area mentioned contains competences specific content about knowledge, ability and skills which should be presented.

A. Mathematical concepts in the 7th grade of the elementary School

In September - *Form and space*

Students will learn how to:
- use geographical maps and a simple two-dimensional coordinate system to determine the distance between two points on the numeric line,

- draw, using a two-dimensional rectangular coordinate system, a point defined by coordinates; students will also read point coordinates.

In October - *Numbers*

Students will learn how to:
- read, write, and compare decimals, and percentages, and present them using common notations, convert fractions into decimals with a given level of preciseness,

- select the appropriate method of calculation (in their heads, on paper, and using a pocket calculator),

- assess the significance and accuracy of their calculation results,

- use positive rational numbers, whole numbers, and mathematical operations to solve simple mathematical problems and problems in their everyday lives.

In November - *Algebra and functions*

Students will learn how to:
- apply percentages and percentage calculations in concrete situations,

- recognise proportional quantities and apply ratios and proportionality in simple everyday situations,

- substitute variables with numerical values in a formula, and calculate the value of the remaining variable,

- recognise and apply proportionality and reverse proportionality in simple situations.

In December - *Data*

Students will learn how to:
- collect, classify, and organise data, and present them using tables, frequency diagrams, pictograms, bar charts, pie charts, and systemic lists, tables, frequency diagrams, line charts, bar charts, pie charts, „box and whiskers" charts, and various graphs,

- read and interpret, and analyse data presented in various ways,

- determine the arithmetic mean, range, and median of a series of numerical data,

- determine the number of possible and, for a given event, favourable outcomes in simple situations,

- apply the terminology of probability (outcome, event, opposite event, possible, impossible, certain, coincidental, random, probable, chance, probability),

- assess and calculate the probability of an outcome in simple situations, and present it in the form of a fraction, decimal, and percentage,

- efficiently determine by argument the number and probability of possible and favourable outcomes in simple situations,

- assess the probability of particular random events, explaining them as relative frequencies.

In January - *Form and space*

Students will learn how to:
- recognise, identify by name, construct, and classify two- and three-dimensional geometric forms, and identify and precisely describe their geometrical properties including congruous triangles,

- sketch/draw simple two-dimensional forms, and reproduce them using geometrical tool sets and a simple computer drawing programme,

- understand basic relationships and patterns in two- and three-dimensional geometric forms, including congruence and similarity in triangles,

- recognise congruence, similarity, and symmetry in two- and three-dimensional forms, in their everyday surroundings, and in art, and use them when analysing and describing the world around them.

In February - *Measuring*

Students will learn how to:
- convert standard units of measure for length, surface, volume, mass, time, temperature, angle, and average velocity, and apply these conversion techniques in their everyday lives,

- indirectly measure length by applying proportionality and similarity,

- apply formulas to determine the perimeter and area, as well as the sum of internal angles, of a triangle and rectangle,

- determine the measurable features of objects or phenomena in everyday situations, select appropriate measuring units and devices, and apply measurement techniques in problem solving.

In March and April - *Algebra and functions*

Students will learn how to:
- solve linear equations, and simple two-variable systems of two linear equations, and use substitution to verify the accuracy of their solutions,

- convert a simple problem into algebraic notation (e.g., numerical sentences, linear equations, systems of two linear equations), plan and execute its solution, and understand the significance of their solutions.

In May - *Form and space and Algebra and functions*

Students will learn how to:
- draw, using a two-dimensional rectangular coordinate system, a point defined by coordinates and a straight line defined by equation; students will also read point coordinates.
- present simple numerical relationships (e.g., linear, pure quadratic, square root) using mutually interchangeable words, associated values tables, equations, and graphics.

II. STATISTICAL SET AND MEASURING VARIABLES

Elements of our statistical set were 48 children from the Elementary school Dr. Ivan Merz. Testing was performed as a part of an ordinary testing of the children's competences, which is part of the school education's curriculum. [2] The gender structure of children examined is shown in Table 1:

Table 1 Children's gender

Children's gender	Number of children
male	27
female	21
Total	**48**

The measuring variables for algebra and functions competences are:

Ratio and proportions (RATPRO)
Proportionality (PROPOR)
Inverse proportionality (INVPRO)
Method of substitution (MRTSUB)
Method of elimination (METELI)
Graphing linear function (GRLIFU)

The measuring variables for form and space competences are:

Coordinate system (COORSY)
Similarity of triangles (SIMTRI)
Polygons (POLYGO)

The measuring variables for measuring competences are:

Perimeter and area of similar triangles (ARSITR)
Peripheral and central angle (PECEAN)
Perimeter and area of circle (ARECIR)

The measuring variables for data competences are:

Percentage (PERCEN)
Statistics (STATIS)
Probability (PROBAB)

Every variable on this list was described separately through the tasks. Every task was evaluated with the following grades: beeing 1(need improvement), 2 (satisfactory), 3 (good), 4 (very good) and 5 (excellent). Tasks for measuring variables are:

Ratio and proportion (RATPRO).
Determining any unknown participant of the ratio and applying normal sizes in reinforcement of the task.
Example: In what ratio are turtles? (Figure 1)

Proportionality (PROPOR). Recognizing and graphically presenting proportionality and applying it in tasks concerning everyday existence.
Example: Using this table, determine how many kunas (HRK) will you get for 200 Euros? (Figure 2)

Inverse proportionality (INVPRO). Recognizing and graphically presenting reversed proportionality and determining the unknown value using practical tasks.
Example: It takes two workers 15 days to finish redecorating an apartment if they work 8 hours per day. How long would it take 3 workers to finish redecorating the apartment if they worked 10 hours per day?

Method of substitution (METSUB) and Method of elimination (METELI).
Example: Find the solution for the system of two linear equations with two variables using: a) Method of substitution, b) Method of elimination.

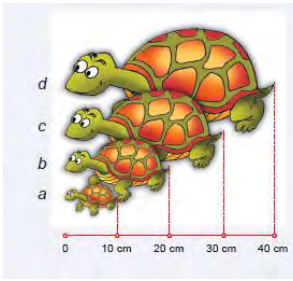


Figure 1

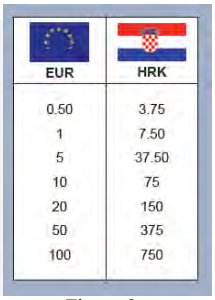


Figure 2

Graphing linear function (GRLIFU).
Recognizing a chart presenting a function. Drawing a chart of a linear function emphasizing the dependence between two values using the chart and various tables.
Example: Using the given chart determine which day in a week was the hottest.

Coordinate system (COORSY).
Pupils were asked to determine the position of a point in a plane using coordinates within the rectangle coordinate system. Each point was assigned to a rational number directed by rules.
Example: Find the coordinates of the marked points on the boat. (Figure 3)

Similarity of triangles (SIMTRI). Applying rules about triangle resemblances and determin the coefficient of similarity. Calculating the unknown lengths of triangle sides of different sorts of triangles.
Example: Help the painter measure the man's height in the right ratio. (Figure 4)

Polygons (POLYGO).
Drawing and constructing polygons. Determining the total number of diagonals within the polygon. Calculating the total summing of all inner polygon angles and all angles of the specific triangles within the regular polygon.
Example: Using the given data, find out the area of the polygon.

Perimeter and an area of similar triangles (ARSITR). Calculating an area of similar triangles.
Example: Using the given data, find out how wide is the lake? (Figure 5)

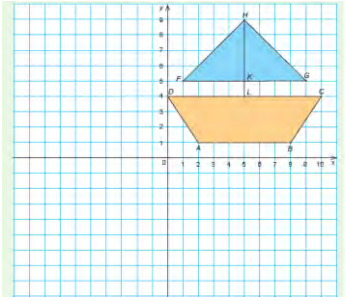


Figure 3



Figure 4

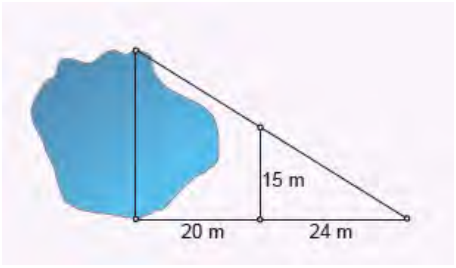


Figure 5

Peripheral and central angle (PECEAN). Determining the relation between the two sizes of two different angles and applying of Pythagoras rule in constructive tasks. Example: Determine the size of unknown angles.

Perimeter and an area of circle (ARECIR). Understanding how an area of circle is proportional to its diameter. Calculating the area and circumference of a circle with having the length of its radius.
Example: Determine the area of given circles. (Figure 6)

Percentage (PERCEN). Expressing the percentage using a rational number related to everyday situations.
Example: Write down the percentage of colored parts in these shapes. (Figure 7)

Statistics (STATIS). Recognizing the features of a certain information group, showing, sorting, and organizing the data from everyday situations. Presenting the information by using simple tables, charts or circled diagrams. Example: Which class has the most absent students on Monday, and which the fewest absent students on Wednesday? (Figure 8)

Probability (PROBAB). Recognizing the impossible, probable and safe event. Calculating the probability of the outcame of the event (result).
Example: Who do you think will win a 100 m race? (Figure 9)



Figure 6

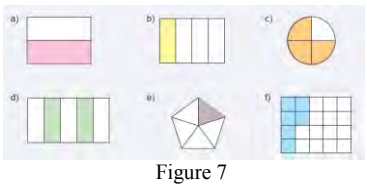


Figure 7

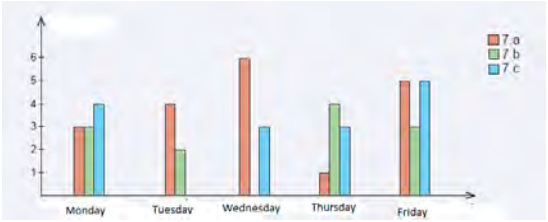


Figure 8

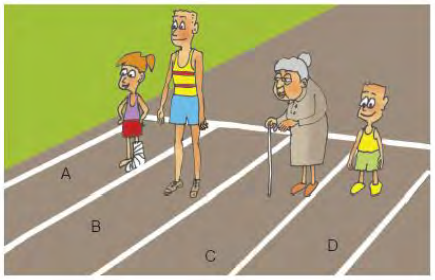


Figure 9

III. MARKOV MODEL

Markov model (MM) is Markov network or Markov random field. To describe MM we first have to describe Markov structure (MS). MS is undirected graph with random variables as vertices which connect some pairs of vertices. In our MS, vertices are measuring mathematical competences or $x_1, x_2,\ldots,x_{15} \in$ {RATPRO, PROPOR, INVPRO, METSUB, METELI, GRLIFU, COORSY, SIMTRI, POLYGO, ARSITR, PECEAN, ARECIR, PERCEN, STATIS, PROBAB}. Using structural equation model [1] we can describe MS with structural equations:

$$x_i = \sum_{j=1}^{17} b_{ij} x_j + e_i \qquad (i = 1,\ldots,17) \qquad (1)$$

where e_i are exogenous variables. In the matrix notation we can write equations.

$$x = Bx + e \qquad (2)$$

For undirected graph matrix B has to be symmetric. In this paper we use: free software Tetrad 5.2.1-3, [8] [13], namely PC Adjacency Search Algorithm or PC Algorithm [14]. PC Algorithm uses conditional independency in different orders. First order conditional independency is $i \perp j \mid k$, second order independency is $i \perp j \mid \{k, m\}$, third order independency is $i \perp j \mid \{k, m, n\}$ and so on. Conditional independency evaluate with partial correlation coefficients $\rho_{ij\,|k}$, $\rho_{ij\,|\{k,m\}}$, $\rho_{ij\,|\{k,m,n\}}$ and so on. If partial correlation coefficient is zero then we have conditional independency. In MS we have Markov property. Markov property for pair $\{i,j\}$ vertices is $i \perp j \mid E\backslash\{i, j\}$ where E is set of vertices in MS.

IV. RESULTS

The authors of this paper, Gordana Paić, teacher at Elementary School Dr. Ivan Merz in Zagreb, [3][4][5] together with professor Tepeš [9][10][11][12] and professor Pavlina at Faculty of Humanities and Social Sciences [6][7], chose the most appropriate Markov model on the Markov structure using Tetrad software [8][13]. Markov structure math competences for the 7th grade elementary school shown in Figure 10.
For Markov model math competences [15] we calculate conditional information. The conditional information is:

$$I(i \perp j \mid E\backslash\{i,j\}) = -0{,}5 \text{ lb } (1 - \rho^2_{ij\,|E\backslash\{i,j\}}) \qquad (3)$$

Where lb is logarithm with base 2.
The conditionals information are in Table 2.

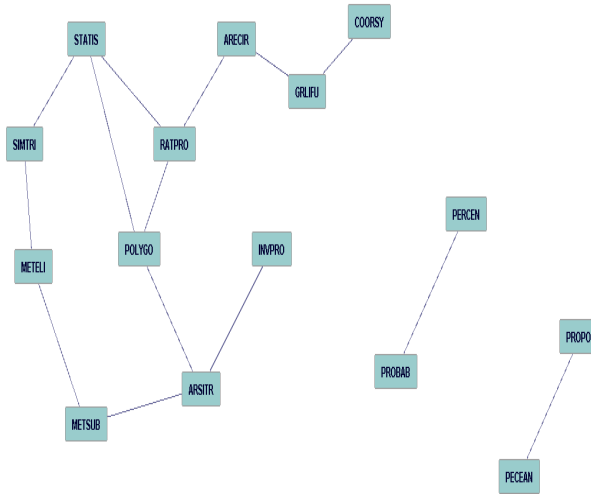


Figure 10

Table 2 Conditionals information

i	j	$\rho^2_{ij\,\|E\backslash\{i,j\}}$	$I(i \perp j \mid E\backslash\{i,j\})$
PROBAB	PERCEN	-0,0122	0,0001
ARSITR	INVPRO	-0,0269	0,0005
COORSY	GRLIFU	0,0291	0,0006
ARSITR	METSUB	0,0659	0,0031
POLYGO	RATPRO	0,0986	0,0070
STATIS	**RATPRO**	**-0,1175**	**0,0100**
METELI	**SIMTRI**	**0,1239**	**0,0112**
ARSITR	**POLYGO**	**-0,1400**	**0,0143**
STATIS	**POLYGO**	**0,1422**	**0,0147**
STATIS	**SIMITRI**	**-0,1457**	**0,0155**
ARECIR	**RATPRO**	**0,1604**	**0,0188**
PROPOR	**PECEAN**	**0,2413**	**0,0433**
ARECIR	**GRLIFU**	**0,2890**	**0,0629**
METELI	**METSUB**	**0,3042**	**0,0700**

In Table 2. we can see that conditional information greater then 0,01; in bold numbers, is in connection with STATIS. This result leads us to the need for greater role of learning about the data. On the other hand, METSUB, ARSITR and GRLIFU are very important. It means that we must have good competences in these math areas. Math competences COORSY, INVPRO, PERCEN and PROBAB as well as PROPOR and PECEAN have no connection with other math competences.
Statistics (STATIS) is associated with Similarity of triangles (SIMTRI), Polygons (POLYGO) and Ratio and proportions (RATPRO).

Statistics (STATIS) introduces students to the data and information (STATIS, SIMITRI), so (STATIS, POLYGO) is 50% higher than (STATIS, RATPRO), and STATIS has got a 50% greater effect on SIMITRI and POLYGO, than on RATPRO.

Perimeter and area of circle (ARECIR), Graphing linear function (GRLIFU), (ARECIR, GRLIFU) is three times higher than (ARECIR, RATPRO), that is, ARECIR represents geometry and has got a three times greater effect on GRLIFU, than RATPRO, which represents arithmetics.

Method of substitution (METSUB) and Method of elimination (METELI), (METELI, METSUB) is 7 times higher that (METELI, SIMITRI). This shows a high degree of connection between the equation solving methods.

Proportionality (PROPOR) and Peripheral and central angle (PECEAN), (PROPOR, PECEAN) = 0,0433 and shows a high degree of connection between PROPOR and PECEAN, since the first competention has got a presumption of the application in PECEAN competence.

V. CONCLUSION

This work demonstrated the Markov model of mathematical competences in school education. For the purposes of adopting mathematical competences, Markov model refers to the order of adopting of mathematical competences. For the purpose of further research, it is necessary to increase the statistical set of the number of children examined.

The curriculum for children in elementary schools in Croatia is based on social relations, and should have the educational structure of mathematical competences accustomed to the age and level of competences that children acquire by using information and communication technologies of contemporary society.

Based on the above, one could rearrange the competences and the research should be expanded to the cause and effect connections of the mathematical competences.

REFERENCES

[1] Boolen, K. A. (2007) *An Overview of Equation Models with Latent Variables*, Miami University Symposium on Computational Research, March 1-2, 1-51

[2] *National Curriculum Framework for Pre-school Education and General Compulsory and Secondary Education*, (2011), Ministry of Science, Education and Sport of the Republic of Croatia

[3] Paić G., Bošnjak Ž., Čulina B. (2014.): *Mathematical challenges 7 – textbook with workbook*, Alfa, Zagreb, (Croatian)

[4] Paić G., Bošnjak Ž., Čulina B. (2014.) : *Mathematical challenges 7 – worksheets*, Alfa, Zagreb, (Croatian)

[5] Paić G., Bošnjak Ž., Čulina B. (2014.): *Mathematical challenges 7 – tests*, Alfa, Zagreb

[6] Pavlina, Krešimir; Špiranec, Sonja, Pongrac Pavlina, Ana.Information Competences - Students' Perspective // 37th International Convention MIPRO 2014: Computers in Education

[7] Pavlina, Krešimir; Pongrac, Ana; Latas, Branka.Hardware equipment of computer classrooms in Croatian elementary schools // 35th International Convention MIPRO 2012: Computers in Education

[8] Scheines, R., Ramsey, J., Spirters, P. and Glymor, C. (2015), *Causal Model Search*, gentle introduction to Tetrad

[9] Tepeš, B. (2009), *Statistical Models on Graphs* (Croatian), Faculty of Humanities and Social Sciences

[10] Tepeš, B., Lešin, G & Hrkač,A. (2013), *Causal Modelling in Mathematical Education*, The 1st International Conference on "Research and Education - Challenges Towards the Future, Schroder, Albania

[11] Tepeš, B., Šimović V. and Tepeš, K.(2014), *Causal Model of Mathematical Competences in Kindergarten*, International

Teacher Education Conference 2014, Proceeding Book, pp. 102 – 107, UAE Dubai

[12] Tepeš, B., Šimović V. and Tepeš, K. (2015), *A Note on Modeling of Mathematical Competences*, 14th Hawaii International Conference on Education 2015, Proceeding Book, USA Hawaii

[13] The Tetrad Project (2015), *Launch Tetrad 5.2.1-3*

[14] Spirtes, P. Glymour, C. and Scheines, R. (2001): *Causation, Prediction, and Search*, The MIT Press

[15] Whittaker, J. (1990): *Graphical Models in Applied Multivariate Statistics*, John Willey & Sons

PYTHON as Pseudo Language for Formal Language Theory

Zdravko DOVEDAN HAN, Kristina KOCIJAN, Vjera LOPINA

Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb

Abstract
In this paper we use selected topics from formal language theory to show that programming language Python, with its syntax and semantics of primitive and structured statements (strings, tuples, lists and dictionaries) and functions, is suitable for use as a pseudo language and description of structures and algorithms in the formal language theory.

Keywords
formal language theory, pseudo language, pseudo code, Python, language defining, syntax analysis, translation.

I. INTRODUCTION

Formal language theory (FLT) appeared in the fifties of the twentieth century in the early works of Noam Chomsky. It was based on an appropriate natural language (e. g. English) model. Soon afterwards the use of context-free grammar and what is today called "Backus-Naur Form" (BNF) gave a major boost to the further development of FLT. It was used for the first time in the definition of ALGOL 60 language, in 1963.

Since then a strong feedback between formal language theory and programming languages was created: formal language theory found its application in defining programming languages and their translators, while programming languages have helped in the study of formal language theory. Many FLT books published from the end of the seventies to the present day have used Pascal and C for a description of certain structures and algorithms of formal language theory.

However, we argue that the right tool for use in FLT was obtained only by the emergence of Python language, which is now at the top of popularity, and combines all programming paradigms: structured, object-oriented, logical, dynamic and functional and which has a wide range of applications in many branches: mathematics, physics, chemistry, biology, natural language processing, databases, web applications, information technology, computer science, etc..

Using the examples of selected topics in formal language theory, this paper is structured as follows: In Section II. we introduce some basic definitions of formal language theory and show how the definition of grammar and generating of sentential form can be directly translated in Python. In Section III. we show how a pushdown recognizer for syntactic analysis of one simple language can be written in Python. We conclude by identifying some directions for future work.

II. LANGUAGE DEFINING

In order to define formal language, let's start from some basic definitions, [1] and[3].

Character is a unique (undivided or atomic) element. For example, upper- or lower-case Roman letters and digits are characters. *Alphabet* is a finite set of characters. *String* is a sequence of characters. *Empty string* is a string that has no symbols. It will be denoted by ε. The *length* of a string x, denoted $d(x)$ or $|x|$, is the number of characters in the string. The length of empty string is zero. If A is an alphabet, A^* denotes the set of all strings over A, including empty string ε. A^+ denotes a set of $A^*\setminus\{\varepsilon\}$. A *language* over an alphabet A is a set of strings over A, that is, $L \subseteq A^*$. It is often written $L(A)$ to show that some language L is defined over an alphabet A.

Sets of strings that make up elements of a language are called *sentences*. So, a language is a set of sentences. Strings of finite-length that can be seen as a unique, undivided whole are often observed. Such strings are called *symbols* or *words*. Set of all symbols defined over an alphabet A will be marked with V and called *vocabulary*. Since $V \subseteq A^*$, we conclude that V is a language. For example, vocabulary

V = {i,iv,v,ix,x,xl,l,xc,c,cd,d,cm,m}

can be defined over the alphabet

A = {i,v,x,l,c,d,m}.

According to the Chomsky hierarchy (Chomsky, 1957) languages are classified into four groups (or types), as follows:

type	language
0	*unrestricted*
1	*context-sensitive*
2	*context-free*
3	*linear*

There are several methods for specifying the set that makes language. One method uses the formalism of *regular sets* and *regular expressions*. It is applicable only to the description of the type 3 language.

Second method uses a system called generative grammar. Every sentence of a language can be derived using grammar rules (called "productions").

Third method belongs to a class of *automata*. *Generators* are automata that can generate sentences of languages, but automata are more often used in the role of *recognizers* (in syntax language analysis).

Grammars

In general, a language is a finite set of sentences where sentences are of finite length. However, for many languages it is not possible to put an upper bound on the length of the longest sentence in that language and/or to the number of sentences. Except for that, it would be very unpractical to list all the sentences of a language even if a language consists of a finite number of sentences (for example 100 or 5000!). There are two principal methods of defining languages: by grammar and by automaton.

Grammar is a formalism to generate all four types of languages. A grammar is a 4-tuple, $G = (N,T,P,S)$, where:

- N is a finite set of *nonterminal symbols*,
- T is a finite set of *terminal symbols*, different from N,
- P is a finite set of pairs (α,β),where:
  $\alpha = \alpha_1\gamma\alpha_2;\ \alpha_1, \alpha_2, \beta \in (N \cup T)^*,\ \gamma \in N$

  An element (α,β) in P will be written $\alpha \rightarrow \beta$ and called *production*.
- S is a special symbol in N, $S \in N$, called *start symbol*.

If in some grammar, P contains following productions

$\alpha \rightarrow \beta_1 \ldots \alpha \rightarrow \beta_n$

this is written

$\alpha \rightarrow \beta_1 \mid \beta_2 \mid \ldots \mid \beta_n.$

The sign ' | ' is read 'or'. β_i are alternatives for α.

A grammar defines a language in a recursive manner. If $G = (N,T,P,S)$ is a grammar, *sentential form* of G is defined recursively as follows:

(1) S is a sentential form.

(2) If $\alpha\delta\gamma$ is a sentential form, where $\alpha,\gamma \in (N \cup T)^*$, and $\delta \rightarrow \beta$ is in P, then $\alpha\beta\gamma$ is also a sentential form.

A sentential form of G containing no nonterminal symbols is called a *sentence* generated by G.

Let $G = (N,T,P,S)$ be a grammar. It is defined as a relation $\Rightarrow$, to be read as *directly derives*, on $(N \cup T)^*$ as follows: If $\alpha\delta\gamma$ is a string in $(N \cup T)^*$ and $\delta \rightarrow \beta$ is a production in P, then $\alpha\delta\gamma \Rightarrow \alpha\beta\gamma$.

If $\alpha_0, \alpha_1, \ldots, \alpha_n, \alpha_i \in (N \cup T)^*$, $n \geq 1$, such that

$\alpha_0 \Rightarrow \alpha_1 \Rightarrow \ldots \Rightarrow \alpha_n$

then $\alpha_0 \overset{n}{\Rightarrow} \alpha_n$ is a *derivation of length n*. Generally, it is written

$\alpha_0 \overset{*}{\Rightarrow} \alpha_n,\ n \geq 0,\ \alpha_0 \overset{+}{\Rightarrow} \alpha_n,\ n > 0$

and said that α_0 *derives* α_n. Thus, a language L generated by a grammar G is:

$L(G) = \{w \in T^* : S \overset{*}{\Rightarrow} w\}$

It is said that two grammars, G_1 and G_2, are *equivalent* if $L(G_1) = L(G_2)$.

Grammar classification

Grammars can be classified according to the format of their productions. If $G = (N,T,P,S)$ is a grammar, it is said that G is:

1) Right-linear or type 3 if each production in P is of the form

   $A \rightarrow xB$ or $A \rightarrow x$ $\quad A,B \in N,\ x \in T^*$

   Left-linear if each production in P is of the form

   $A \rightarrow Bx$ or $A \rightarrow x$ $\quad A,B \in N,\ x \in T^*$

   A right-linear grammar is called a regular grammar when

   (a) All productions, with the possible exception of $S \rightarrow \varepsilon$, are of the form $A \rightarrow aB$ or $A \rightarrow a$, where A, $B \in N, a \in T$.

   (b) If $S \rightarrow \varepsilon$ is in P, then S does not appear on the right side of any production.

2) Context-free or type 2 if each production in P is of the form

   $A \rightarrow \alpha$ $\quad A \in N,\ \alpha \in (N \cup T)^*$

3) Context-sensitive or type 1 if each production in P is of the form $\alpha \rightarrow \beta$
   where $|\alpha| \leq |\beta|$.

4) Unrestricted or type 0 if there are no restrictions as the ones above.

Grammar implementation in Python

The grammar $G = (N,T,P,S)$ can be directly translated in the Python 4-tuple in which the elements N and T can be implemented as lists, element P as dict and start symbol S as symbol from the list N. Initially we define grammar as text writing only the production in the form of:

```
alfa -> beta
```

Nonterminals are upper-case letters, `#` is empty string and symbol `->` is production sign. Here are two examples of productions, context-free grammar of the Roman numerals language and context-sensitive grammar of language $\{a^n b^n c^n d^n, n>0\}$:

```
Roman = """
R -> MA|CB|XD|I
M -> m|mm| mmm
A -> #|CB| XD I
C -> c|cc|ccc|cd|d|dc|dcc|dccc|cm
B -> #|XD|I
X -> x|xx|xxx|xl|l|lx|lxx|lxxx|xc
D -> #|I
I -> i|ii|iii|iv|v|vi|vii|viii|ix
"""
abcd = """
S  -> aBCSd| abcd
Ba -> aB
Bb -> bb
Ca -> aC
Cb -> bC
Cc -> cc """
```

Procedure `GRM()` from the productions of grammar, given as a text (or text file), returns the definition of grammar $G=(N,T,P,S)$ that can be displayed by calling the printing procedure, `Write_GRM()`:

```
from random import *
def GRM (X) : # Grammar definition
  N = T = ''
  A = (X.replace (' ', '')).split('\n')
  S = A[1][0];  Y = { 'alfa' : [], 'start' : S}
  for a in A :
     if not a : continue
     [x, y] = a.split('->')
     y      = tuple (y.split('|'))
     Y[x], Y['alfa'] = y, Y['alfa'] +[x]
     for c in y :
       T += c *(not c.isupper() and c not in T)
     for c in x :
       N += c *(c.isupper() and c not in N)
  return list(N), list(T), Y, S
def Write_GRM (G) :
  N, T, P, S = G
  print P['name'], '= (N, T, P, S)'
  print 'N = { ' +('%s, '*(len(N)-1) \
              % tuple(N[:-1])) +N[-1] +' }'
  print 'T = { ' +('%s, '*(len(T)-1) \
              % tuple(T[:-1])) +T[-1] +' }'
  print 'S =', P['start']
  print 'P :'
  for x in P['alfa'] :
      print x, '->', P[x][0],
      for y in P[x][1:] : print '|', y,
      print
  print
```

For example, the grammar of the Roman numerals language is:

```
>>> G = (N, T, P, S) = GRM (Roman); Write_GRM (G)
G = ( N, T, P, S )
N = { R, M, A, C, B, X, D, I }
T = { m, #, c, d, x, l, i, v }
S = R
P :
R -> MA | CB | XD | I
M -> m | mm | mmm
A -> # | CB | XD | I
C -> c | cc | ccc | cd | d | dc | dcc | dccc | cm
B -> # | XD | I
X -> x | xx | xxx | xl | l | lx | lxx | lxxx | xc
D -> # | I
I -> i | ii | iii | iv | v | vi | vii | viii | ix
```

and grammar of language $\{a^n b^n c^n d^n, n>0\}$ is:

```
>>> G = (N, T, P, S) = GRM (abcd);  Write_GRM (G)
G = ( N, T, P, S )
N = { S, B, C }
T = { a, d, b, c }
S = S
P :
S -> aBCSd | abcd
Ba -> aB
Bb -> bb
Ca -> aC
Cb -> bC
Cc -> cc
```

Procedure `DER(P)` generates sentences of the language defined by productions P of grammar G. If called by `DER(P,True)`, a sequence of sentential forms (SF) will be displayed.

```
def DER (P, DSP = False) :
  SF = P['start'];  print SF,
  if not DSP : print '*=>',
  while True :
    for a in P['alfa'] :
      if a in SF :
        x  = (''.join(sample (P[a], 1)))
        i  = SF.find(a);  x *= (x != '#')
        SF = SF[:i] +x +SF[i+len(a):]
        if DSP : print '=>', SF,
        break
    else :
      print SF if not DSP else ''; return SF
```

```
>>> G = (N, T, P, S) = GRM (Roman)
>>> for i in range(3) :
       sf = DER (P, True);  print

R => CB => cB => cXD => cxD => cxI => cxiii
R => XD => xxxD => xxxI => xxxi
R => MA => mA => mXD => mxcD => mxcI => mxciii

>>> G = (N, T, P, S) = GRM (abcd)
>>> for i in range(3) :
       sf = DER (P, True);  print

S => aBCSd => aBCabcdd => aBaCbcdd => aaBCbcdd
=> aaBbCcdd => aabbCcdd => aabbccdd
```

III. SYNTACTIC ANALYSIS

In practice we often encounter the problem that a grammar or a generator of the language is known and a character string is given, and the question is asked whether this is a sentence of the language generated by a given grammar or generator. This process is called *syntactic analysis*.

If the language is defined by a grammar, problem reduces to finding a sequence of derivations (sentential forms), starting from S, which would result in this string (sentence). Such a procedure of syntax analysis is called *parsing*. Parsing process structure on the computer (the program in some selected programming language) is called *parser*, [4].

If the language is defined by an automaton, we ask the question: can a given string be generated by a given generator? Then the automaton is in the role of language recognizer, which analyzes the input string, and after a finate number of changes in their configuration, starting from an initial state reaches a final state if the string is in the language and answers "yes", or the process interrupts and answers "no" if the input string is not in the language. Such a syntax analysis procedure is called *recognizing*, and an automaton that does it is called *recognizer*, [4].

Recognizing

Third language defining method is by an *automaton*. It is a device which consists of a combination of the following parts: an input tape with an input head (reader), an output tape with an output head (writer), an auxiliary memory, and a finite set of rules which controls or regulates the information flow.

Depending on the type of language that automaton recognizes, there are following types of recognizers:

Name	*Definition*	*Language*
final	$M = (Q,\Sigma,\delta,q_0,F)$	type 3
pushdown	$P = (Q, \Sigma, \Gamma, \delta, q_0, Z_0, \mathcal{F})$	type 2
double- pushdown	$Pt = (Q, \Sigma, \Gamma_1, \Gamma_2, \delta, q_0, \mathcal{F})$	type 1
Turing machine	$Tg = (Q, \Sigma, \Gamma, \delta, q_0, F)$	type 0

where:

Q	final set of states
$P(Q)$	power set of Q
Σ	alphabet
Γ	alphabet of stack
Γ_1, Γ_2	alphabet of first and second stack
Z_0	the initial character of stack, $Z_0 \in \Gamma$
δ	*transition function,*
q_0	*the initial state,* $q_0 \in Q$
F	*set of final states,* $F \subseteq Q$

Depending on the type of language, transition function is defined as:

δ: $Q \times \Sigma$	$\rightarrow P(Q)$	type 3
δ: $Q \times (\Sigma \cup \{\varepsilon\}) \times \Gamma$	$\rightarrow Q \times \Gamma^*$	type 2
δ: $Q \times (\Sigma \cup \{\varepsilon\}) \times \Gamma_1 \times \Gamma_2$	$\rightarrow Q \times \Gamma_1^* \times \Gamma_2^*$	type 1
δ: $Q \times \Sigma \cup \{\#\}$	$\rightarrow Q \times \{\Gamma \setminus \{\#\}\} \times \{-1,0,1\}$	type 0

To show the syntax analysis of context-free languages using pushdown recognizer P, first we introduce the definitions, [1]:

1) A *configuration* of P is a tuple (q, w, α) from $Q \times \Sigma^* \times \Gamma^*$, where:

q	current state
w	remaining portion of the input
α	content of the pushdown list; the leftmost symbol of α is the topmost pushdown symbol.

2) An *inital configuration* of P is (q_0, w, Z_0),
3) A *final configuration* of P is (q, ε, α), $q \in F$, $\alpha \in \Gamma^*$,
4) A *move* by P is binary relation $\vdash$. We write

   $(q, aw, Z\alpha) \vdash (q', w, \gamma\alpha)$

   if $\delta(q,a,Z)$ contains (q',γ) for any $q \in Q$, $a \in \Sigma \cup \{\varepsilon\}$, $w \in \Sigma^*$, $Z \in \Gamma$. We say that a input string w is *accepted* by P if

   $(q_0, w, Z_0) \vdash^* (q, \varepsilon, \alpha)$

The language defined by P, denoted $L(P)$, is set of strings w accepted by P. It is generally a context-free language:

$$L(P) = \{w: w \in \Sigma^* \wedge (q_0, w, Z_0) \vdash^* (q, \varepsilon, \alpha),\ q \in F,\ \alpha \in \Gamma^*\}$$

Python's *dict* is the most appropriate structure for the implementation of the transition function because it represents its copy. If D is the transition function of any type of recognizer, its elements will generally have the structure:

$$D = \{ x_0 : y_0,\ x_1 : y_1,\ \ldots,\ x_n : y_n \}$$

where x_i is domain tuple and y_i is codomain whose structure is dependent on automaton type. For example, pushdown recognizer of language `Exp` generated by a grammar:

```
E -> E+E| E*E| (E)| a| b
```

written in Python is given below:

```
# -PUSHDOWN RECOGNIZER
Exp = """
Q  = [0, 1];  A = ['a', 'b', '+', '*', '(', ')']
St = ['$', '('];  _1 = '$';  s = 0;  F = [1]
D  = {
  (0,'a','$'): (1, '$'),  (0,'a','('): (1, '('),
  (0,'b','$'): (1, '$'),  (0,'b','('): (1, '('),
  (0,'(','$'): (0, '($'), (0,'(','('): (0, '(('),
  (1,'+','$'): (0, '$'),  (1,'*','$'): (0, '$'),
  (1,'+','('): (0, '('),  (1,'*','('): (0, '('),
  (1,')','('): (1, ''),   (1,'','$') : (1,'')  }
Name = 'Exp'
DSP  = (Q, A, St, _1, D, s, F) """
NL = '\n'
def Input_W ():
  return (raw_input ('Enter input string: ')).\
         replace (' ', '')

def Write_SP (Name): # pushdown recognizer
  print Name
  print NL, 'SP = (Q, A, St, _1, D, s, F)', NL
  print 'Q  =', Q, NL, ' A =', A, NL, \
        'St =', St, NL, '_1 =', _1, NL, \
        ' s =', s, NL, ' F =', F
  print NL, 'D:'
  S = D.keys();  S.sort()
  for d in S: print ' ', d, '=', D[d]
  print

def Write_C (y, C): print y, C

def SP (x):
  global Q, A, St, _1, D, s, F

  Ok = True; End = False
  q = s; alfa = '$'
  C = (q, x, alfa);
  Write_C ('', C)
  while len(x)>=0 and Ok and not End:
    X = ''; a = ''
    if len(x) > 0    : X = x[0]; x = x[1:]
    if len(alfa) > 0 : a = alfa[0]
    Ok = False
    d = (q, X, a)
    if d in D :
      q, g = D[d]
      if g == '' and a != '': alfa = alfa[1:]
      if g != ''            : alfa = g +alfa[1:]
      Ok = True
    else : Ok = False

    if Ok:
      C = (q, x, alfa);
      Write_C ('   |--', C)
      if q in F and alfa == '' : End = True
      if End and x != ''       : Ok  = False

  Ok = Ok and End
  return Ok

exec Exp;  Write_SP (Name)

w = Input_W(); print
while len(w) > 0:
  Ok = SP (w)
  if Ok: Write_C ('   |--', 'accept')
  else : Write_C ('   |--', 'error')
  print
  w = Input_W(); print
```

```
>>>
Enter input string: a*(a+b)
 (0, 'a*(a+b)', '$')
   |-- (1, '*(a+b)', '$')
   |-- (0, '(a+b)', '$')
   |-- (0, 'a+b)', '($')
   |-- (1, '+b)', '($')
   |-- (0, 'b)', '($')
   |-- (1, ')', '($')
   |-- (1, '', '$')
   |-- (1, '', '')
   |-- accept
```

IV. TRANSLATION

If Σ is input alphabet and Δ is output alphabet, *translation* from the language L_1, $L_1 \subseteq \Sigma^*$, to the language L_2, $L_2 \subseteq \Delta^*$, is a relation T from $\Sigma^* \times \Delta^*$ so that L_1 is the domain and L_2 is the codomain of T. The sentence y, such that (x, y) is in T, is called *the output* of x.

Syntax-directed translation

One of formalisms for defining translation is *the syntax-directed translation scheme*. Intuitively, syntax-directed translation scheme is simply a grammar in which translation elements are related to each production.

Whenever a production was used in the derivation of an input sentence, the translation element is used to help compute a portion of the output sentence related to the portion of the input sentence generated by that production. *Translational form* of *T* is defined as follows:

1) (S, S) is a translational form and the first S is related to the second S.
2) If $(\alpha A\beta, \alpha' A\beta')$ is a translational form and if $A \to \gamma, \gamma'$ is a rule in R, then $(\alpha\gamma\beta, \alpha'\gamma'\beta')$ is new translational form. Nonterminals from γ and γ' are exactly related, same as in the rule. Nonterminals from α and β are related to such nonterminals from α' and β' in the new translational form exactly as in the old. We write:

   $(\alpha A\beta, \alpha' A\beta') \Rightarrow (\alpha\gamma\beta, \alpha'\gamma'\beta')$

   and read (αAβ, α'Aβ ') "directly derives" ($\alpha\gamma\beta$, $\alpha'\gamma'\beta'$). Similarly to the derivation of sentential form, a series of deriving *k* translational form, where k≥0, will be denoted by $*\Rightarrow$, so the translation defined by *T*, denoted as τ (T), is a set of pairs:

$$\tau(T) = \{ (x, y) \mid (S, S) *\Rightarrow (x, y), x \in \Sigma^*, y \in \Delta^* \}$$

The implementation of SDT in Python is given in the procedure SDT ():

```
def SDT (X) : # Syntax-directed translation
              grammars
  A = (X.replace (' ', '')).split('\n')
  Y = {'start' : A[1][0]}
  for a in A :
     if not a : continue
     b = a.split('->'); N = b[0]
     b = b[1].split(',');
     Y [N] = (tuple (b[0].split('|')),
              tuple (b[1].split('|')))
  return Y
```

where X is an input-output grammar of languages to be translated, with productions of form

```
A -> I, O
```

where *I* are the alternatives for input and *O* for output language. For example, for translating Roman into Arabic numerals, a grammar RA can be defined:

```
#    Input grammar,  Output grammar
RA = """
R -> MA| CB| XD| I,  MA| CB| XD| I
M -> m| mm| mmm,     1| 2| 3
A -> #| CB| XD| I,   000| CB| 0XD| 00I
C -> c| cc| ccc| cd| d| dc| dcc| dccc| cm, \
     1| 2| 3| 4| 5| 6| 7| 8| 9
B -> #| XD| I,       00| XD| 0I
X -> x| xx| xxx| xl| l| lx| lxx| lxxx| xc, \
     1| 2| 3| 4| 5| 6| 7| 8| 9
D -> #| I,           0| I
I -> i| ii| iii| iv| v| vi| vii| viii| ix, \
     1| 2| 3| 4| 5| 6| 7| 8| 9 """
```

To translate the input string, the Roman numeral, into Arabic, it is necessary to derive syntax analysis tree (sequence of derivations) by some parsing procedure. It is left out here, so let's show the scheme of translating Roman numerals into Arabic using the examples of three generated sentential forms:

```
T = SDT (RA);  frm  = "(%s, %s)"
for i in range (3) :
  x = y = T['start']
  print frm % (x, y),
  while not x.islower() :
    for s in x :
      if s.isupper() :
        a, b = T[s]
        z = ''.join(sample (a, 1))
        i = a.index(z);  z = z *(z != '#')
        x = x.replace (s, z)
        a = b[i];  y = y.replace (s, a)
        print '\t-->', frm % (x, y)
        break
  print
```

```
>>>
(R, R) --> (XD, XD)        (R, R) --> (CB, CB)
       --> (xD, 1D)               --> (dB, 5B)
       --> (x, 10)                --> (dXD, 5XD)
(R, R) --> (MA, MA)               --> (dxcD, 59D)
       --> (mmA, 2A)              --> (dxc, 590)
       --> (mmI, 200I)
       --> (mmvi, 2006)
```

Finite transducer

Finite transducer is defined as a 6-tuple $M = (Q,\Sigma,\Delta,\delta,q_0,F)$. It is a finite automaton where Δ is *output alphabet* and δ is a mapping from $Q\times(\Sigma\cup\{\varepsilon\})$ to finite subsets of $Q\times\Delta^*$. A *configuration* of finite transcudes M is a tuple (q, x, y), where

- $q \in Q$ is the current state,
- $x \in \Sigma^*$ is the input string remaining on the input tape, with the leftmost symbol of x under the input head,
- $y \in \Delta^*$ is the output string emitted up to this point.

The *initial configuration* is (q_0, x, ε) and the *final configuration* is (q, ε, y), $q \in F$, $y \in \Delta^*$. A *move* by M is the binary relation ⊢. We write

$$(q, ax, y) \vdash (r, x, yz)$$

if $\delta(q,a)$ contains (r,z) for any $q \in Q$, $a \in \Sigma\cup\{\varepsilon\}$, $z \in \Delta^*$. We say that y is an *output* for x if

$$(q_0, x, \varepsilon) \vdash^* (q, \varepsilon, y)$$

The *translation defined by M*, denoted $\tau(M)$, is

$$\tau(M) = \{(x,y) \mid (q_0, x, \varepsilon) \vdash^* (q,\varepsilon,y),\ x \in \Sigma^*,\ y \in \Delta^*\}$$

Here is an example of transducing Roman numerals into Arabic. D is transition function with structure of elements

```
q : (Tm, Td, Tc, Tl, Tx, Tv, Ti, Tf)
```

where q is state, T*a*=D[q]['mdclxvi$'.find(*a*)].

```
# TRANSDUCER OF ROMAN NUMERALS INTO ARABIC
# Transition function ("table")
e = '';
# q    m      d      c      l       x
#         v        i        $
D = { 'q0' : 0, 'tr' : 'mdclxvi$',
       'Q' : range (31), 'F' : range (1, 31),
  0: ((1,e), (8,e), (4,e), (17, e),(13, e),
         (26,e),  (22, e),  e      ),
  1: ((2,e), (8,1), (4,1), (17,10),(13,10),
         (26,100),(22,100),(e,1000)),    # m
  2: ((3,e), (8,2), (4,2), (17,20),(13,20),
         (26,200),(22,200),(e,2000)),    # mm
```

```
  3: (    e, (8,3), (4,3), (17,30),(13,30),
         (26,300),(22,300),(e,3000)),    # mmm
  4: ((12,e),(7,e), (5,e), (17, 1),(13, 1),
         (26, 10),(22, 10),(e, 100)),    # c
  5: (    e,     e, (6,e), (17, 2),(13, 2),
         (26, 20),(22, 20),(e, 200)),    # cc
  6: (    e,     e,     e, (17, 3),(13, 3),
         (26, 30),(22, 30),(e, 300)),    # ccc
  7: (    e,     e,     e, (17, 4),(13, 4),
         (26, 40),(22, 40),(e, 400)),    # cd
  8: (    e,     e, (9,e), (17, 5),(13, 5),
         (26, 50),(22, 50),(e, 500)),    # d
  9: (    e,     e,(10,e), (17, 6),(13, 6),
         (26, 60),(22, 60),(e, 600)),    # dc
 10: (    e,     e,(11,e), (17, 7),(13, 7),
         (26, 70),(22, 70),(e, 700)),    # dcc
 11: (    e,     e,     e, (17, 8),(13, 8),
         (26, 80),(22, 80),(e, 800)),    # dccc
 12: (    e,     e,     e, (17, 9),(13, 9),
         (26, 90),(22, 90),(e, 900)),    # cm
 13: (    e,     e,(21,e), (16, e),(14, e),
         (26,  1),(22,  1),(e,  10)),    # x
 14: (    e,     e,     e,       (),(15, e),
         (26,  2),(22,  2),(e,  20)),    # xx
 15: (    e,     e,     e,       (),     (),
         (26,  3),(22,  3),(e,  30)),    # xxx
 16: (    e,     e,     e,       (),     (),
         (26,  4),(22,  4),(e,  40)),    # xl
 17: (    e,     e,     e,       (),(18, e),
         (26,  5),(22,  5),(e,  50)),    # l
 18: (    e,     e,     e,       (),(19, e),
         (26,  6),(22,  6),(e,  60)),    # lx
 19: (    e,     e,     e,       (),(20, e),
         (26,  7),(22,  7),(e,  70)),    # lxx
 20: (    e,     e,     e,       (),     (),
         (26,  8),(22,  8),(e,  80)),    # lxxx
 21: (    e,     e,     e,       (),     (),
         (26,  9),(22,  9),(e,  90)),    # xc
 22: (    e,     e,     e,       (),(30, e),
         (25,  e),(23,  e),(e,   1)),    # i
 23: (    e,     e,     e,        e,      e,
          e,(24,  e),(e,   2)),          # ii
 24: (    e,     e,     e,        e,      e,
          e,       e,(e,   3)),          # iii
 25: (    e,     e,     e,        e,      e,
          e,       e,(e,   4)),          # iv
 26: (    e,     e,     e,        e,      e,
          e,(27,  e),(e,   5)),          # v
 27: (    e,     e,     e,        e,      e,
          e,(28,  e),(e,   6)),          # vi
 28: (    e,     e,     e,        e,      e,
          e,(29,  e),(e,   7)),          # vii
 29: (    e,     e,     e,        e,      e,
          e,       e,(e,   8)),          # viii
 30: (    e,     e,     e,        e,      e,
          e,       e,(e,   9)) }         # ix
```

```
p = lambda x : x if str(x) else '#'
TR = D['tr']
def FT (w, a=1):
  A = a;  w += '$';  q = 0;  i = 1;  y = ''
  C = (q, p(w), p(y));  print "(%2s,%s,%s)" % C
  while 'Ok' :
    a = w[0]; j = TR.find(a)
    if j >= 0:
      Q = D[q][j]
      if Q != () :
        q, z = Q; y += str(z); w = w[1:]
        C = (p(q), p(w), p(y))
        print " |-- (%2s,%s,%s)" % C
        if not w : return True
      else : print 'syntax error';  return False
    else:
      print 'illegal character';  return False
w  = raw_input ("input Roman ").lower()
Ok = FT (w)
```

```
>>>
input Roman X
( 0,x$,#)
 |-- (13 $,#)
 |-- ( #,#,10)
```

```
>>>
input Roman mmvi
( 0,mmvi$,#)
 |-- ( 1,mvi$,#)
 |-- ( 2,vi$,#)
 |-- (26,i$,200)
 |-- (27,$,200)
 |-- ( #,#,2006)
```

```
>>>
input Roman dxc
( 0,dxc$,#)
 |-- ( 8,xc$,#)
 |-- (13,c$,5)
 |-- (21,$,5)
 |-- ( #,#,590)
```

V. CONCLUSION

In this paper we demonstrated how Python programming language can be employed as a pseudo-language using the example of selected structures and algorithms of formal language theory for their description and implementation. In [3], [4] and [5] there are many examples of applying Python for implementing procedures of the context-free languages syntax analysis and for recognizing the languages of all types.

Python is also suitable for use in creating the interpreters and the preprocessors, both as an encoding language and as a target language. Examples of such applications are given in [5] where it is implemented as an interpreter for language PL/0, according to [8], and preprocessor of mini language defined in [2].

It has been shown that Python is suitable for use in natural language processing (NLP). In [6] and [7] Python has been successfully applied in the implementation of the lexical and syntax analysis of English language and its translation in a special form.

Finally, we can conclude with satisfaction that the implementation of Python can significantly improve the study of the formal language theory and its applications.

References

[1] AHO, V. A.; ULLMAN, D. J.: *The Theory of Parsing, Translation, and Compiling*, vol. I: *Parsing*, Prentice-Hall, 1972.

[2] DIJKSTRA, E.W.: *A Discipline of Programming*, Prentice-Hall, 1976.

[3] DOVEDAN HAN, Z.: *FORMALNI J EZICI I PREVODIOCI*
- *regularni izrazi, gramatike, automati*, Element, Zagreb, 2012.

[4] DOVEDAN HAN, Z.: *FORMALNI JEZICI I PREVODIOCI*
- *sintaksna analiza i primjene*, Element, Zagreb, 2012.

[5] DOVEDAN HAN, Z.: *FORMALNI JEZICI I PREVODIOCI*
- *prevođenje i primjene*, Element, Zagreb, 2013.

[6] JAKUPOVIĆ, A., PAVLIĆ, M., & DOVEDAN, H. Z.: *Formalisation method for the text expressed knowledge*. *Expert systems with applications*. 41 (11). 5308-5322, 2014.

[7] PAVLIĆ, M., & DOVEDAN, H. Z., JAKUPOVIĆ, A.: *Question answering with a conceptual framework for knowledge-based system development ''Node of Knowledge''*. Expert systems with applications. 42 (2015) 5264–5286, 2015.

[8] WIRTH, N.: *Algorithms + Data Structures = Programs*, Prentice-Hall, 1976.

Croatian Students' Attitudes towards Technology Usage in Teaching Asian languages - a Field Research

Marijana Janjić, Sara Librenjak and Kristina Kocijan
Department of Information and Communication Sciences
Faculty of Social Sciences and Humanities Zagreb
marijanajanjic@yahoo.com, sara.librenjak@gmail.com, krkocijan@ffzg.hr

Abstract - Methodology of foreign language learning in the past 70 years tends to favor the acquisition of skills and know-how tips for the successful application in various social situations. In order to be a well-versed user of a foreign language, it is necessary to learn both grammar and syntax and on the other side, vocabulary. Today's technology enables the usage of different applications in order to achieve the objectives of teaching foreign languages. Thus students have the opportunity to use different applications either (or both) in the classroom in the presence of a teacher, or alone at home as a supplement to the classroom work.

As a part of a larger project, supported by European Social Fund, we researched Croatian students' attitudes towards technology usage in learning Asian languages. The research of attitudes is one of several goals important for our project that also includes the development of materials for e-learning, implementation of the e-tools in the classroom, and testing student's improvement in knowledge every three months. This paper deals with the first phase: a field research of students' attitudes. Our goal was to present the state-of-art of the students' self-described knowledge, issues in learning and willingness to use technology in learning. We plan to compare these results to empirical results in later phases of the project.

Croatian students (N=203) studding in Zagreb, Rijeka, Varaždin and Split participated in the survey, answering a number of questions on the use of modern technology in Japanese (N = 104), Korean (N = 46), Hindi (N = 32) and Sanskrit (N = 21). They expressed their views on questions related to specific difficulties in learning vocabulary, grammar or writing system of a particular language. They also answered questions about their attitude towards teaching materials such as course books and general attitudes about technology in learning. We used students' feedback into account during the process of developing e-materials for each of the mentioned Asian languages.

I. INTRODUCTION

It was in the 18th century that Alexander von Humboldt suggested that the language could not be taught but that one can only create conditions for learning to take place. Three centuries later, deep into the digital era, it seems that we are still not recognizing the power of technology and its usage for creating such language-learning oasis.

Language teaching techniques (have your students repeat after you, use a limited vocabulary initially, help students practice reading and speaking, teach language through pictures to make it meaningful) proposed by Jan Comenius, the famous Czech language teacher, in the 17th century are still valid techniques that may easily be embedded into e-tools. And although many teachers are trying their best to make their language classrooms innovative and implement all the possibilities technology offers such as CALL, MALL, GALL, etc. [1], there are still many of them who underestimate not just the power of technology but also their students' needs, learning habits and attitudes on e-learning in general.

We believe that student's learning needs and learning styles have to be taken into account when preparing learning materials since the main purpose of teaching is to help someone learn. That is why we asked Croatian students of Asian languages to tell us how they perceive e-learning and which aspects of learning one or several Asian languages do they find difficult and challenging. After we analyzed the anonymous student survey results, we created e-learning materials that could be implemented in the classroom and/or as a side learning material for students' individual use to help them cope with some of the issues they stated in the survey. We expect our students to easier acquire the language of their choice, and subsequently perform better on their tests, when exposed to materials that facilitate their learning. Their progress is monitored weekly and tested every three months.

We considered two main factors while developing both survey and learning materials: 1. learning new language requires learning and memorizing new vocabulary, 2. it also requires learning and memorizing new grammatical and syntactic patterns. It is relevant to note that both students and teachers share these goals. Another goal that is shared by probably everyone in the foreign language discourse is the need to be practical: be practical in teaching, be practical in learning, and focus on practical skills and knowledge. Practical here is taken as a synonym for application of skills and knowledge in language classroom, that are both from teacher's and student's point of view rated as relevant and useful outside of the classroom for purposes for which student

requires them. Our idea is to see if and how technology can be useful in attaining these goals in order to help both students and teachers.

In the following sections, we will give more details about the survey we conducted and explain the results we got from the survey dividing them into two main topics (issues in language learning and attitudes towards e-learning technologies). We will conclude the paper with some final remarks.

II. STUDENT SURVEY

Motivation and attitudes towards learning have often been quoted as relevant for student's success in education in general [2], [3], and thus their relevance in CALL or other technology enhanced learning environments should not come as a surprise. For that reason we were very keen to learn about the reasons that attract Croatian students toward the study of one (or more) Asian languages, their aims and goals as well as the relevant learning experiences that they have had thus far in language classroom with their teachers and technology input.

Two hundred and three students in total, studying Hindi, Japanese, Korean or Sanskrit or several of these languages simultaneously (see Table 1), have participated in a survey. The survey had been anonymous in order to protect their privacy and not make them vulnerable in their later contacts with their teachers, since teachers have been given the opportunity to see the survey results. Most of the students are young adults, 45% of them aged 19-23, and 20% aged 24-28. Although majority of students are taking regular college classes at the Faculty of Humanities and Social Sciences in Zagreb, many of them also invest time and resources in the continuing education and thus enroll in language schools, take private classes with tutors etc. A number of students, attending Japanese course in Varaždin, Rijeka or Split have also took part in the survey as well as the first generation of students enrolled in Japanese course at the University of Pula.

Table 1. Total number of students in the survey.

Gender	Number	%
Male	48	24%
Female	152	75%
Undeclared	3	1%
Total	203	100%

One of the first things we wanted to learn about students of Asian languages is how long have they been learning a particular language. The candidates were asked to give an estimate of learning hours they have invested in the learning process prior to their participation in our project. We wanted to see how long does the average learner spend learning the language, and to see if there are differences in the speed of progress after we implement the second phase of the project. The average result shows that 242.8 learning hours (at the high school or university) are behind the majority of them, which means that they have been learning a particular language for more than two years. This estimation is based on the classroom hours and self study. Thus, their motivation cannot be described as lacking, especially since many of them are aware of a particular goal they want to achieve with the language course they are taking. The three most important students' goals can be summed up as: a) to achieve a working proficiency in the particular language, b) to communicate with native speakers, c) to visit or live in a country where a particular language is spoken. Some students have been determined enough to try their luck with several teachers, various course books and various resources. Taken altogether, students included in the survey can be described as well motivated individuals to learn a particular language, as an average individual had over 240 learning hours in at least one of the Asian languages of his/her choice.

III. ISSUES IN LEARNING

There are several categories of learning issues that we were interested in including script, speech production, grammar, memorizing the vocabulary, speech understanding, pronunciation, and writing and composition. We will show here how these issues differ across languages and explain possible reasons for them.

All of our survey participants have a Croatian language as a mother tongue and thus they use Latin script for written communication. On the other side, this is not the case for any of the Asian languages that we are interested in this project (Japanese, Korean, or Indian languages). Of course, as it would be quite natural to assume, this presents the first obstacle in learning the language. This is especially true for Japanese language, which uses three different scripts – an obstacle that can be challenging to overcome to many learners.

Table 2. Level of difficulty in learning a script (student survey).

Language	Script	Average difficulty rating
Japanese	Hiragana	1,91 out of 5
Japanese	Katakana	2,68 out of 5
Japanese	Kanji	4,29 out of 5
Korean	Hangeul	2,0 out of 5
Hindi / Sanskrit	Devanagari	2,57 out of 5

However, according to our survey results, all but one script (*kanji*) has scored the difficulty level lower than 3 out of 5 (where 1 is not difficult at all and 5 is extremely difficult) (Table 2.). Korean script *Hangeul* is not rated as very difficult to learn, with the average difficulty rating of 2 out of 5. *Devanagari*, Indian script, seems to be somewhat more difficult to learn and is rated 2.57 out of 5. In the case of the Japanese, *hiragana* script is considered relatively simple (1.9/5), while katakana, the script used for foreign words, is considered a bit more difficult (2.7/5). The third Japanese script, ideographic

kanji script that consists of around 2000 characters, is rated as the most difficult script to learn (4.3/5). To overcome this difficulty in learning kanji, we prepared some multimedia solutions that are presented in more details in [4].

Representations of the difficulty rating of different scripts are given in Table 2 while the most common issues in learning are presented in Figures 1, 2, 3 and 4 for Japanese, Korean, Hindi and Sanskrit, respectively.

Contrary to Croatian, which is highly flective language with seven cases, three genders and complex verb systems, East-Asian languages generally do not possess the case and gender system or many complex verb tenses. Japanese and Korean are agglutinative languages with SOV word order, and the syntactic relations between words are expressed with particles. What can be problematic in language learning of these languages is the difference in language logic and pragmatics that influence the grammar. For example, students of Japanese consider honorific speech system as the most difficult element in grammar learning, because Croatian language (besides the *ti-Vi* difference in politeness[1]) does not possess such a system.

Most students also report having difficulties in using and understanding particles, different verb endings, and some tenses which are expressed with one tense in Croatian, while in Japanese there are several ways of expressing different nuances of the tense (e.g. conditional).

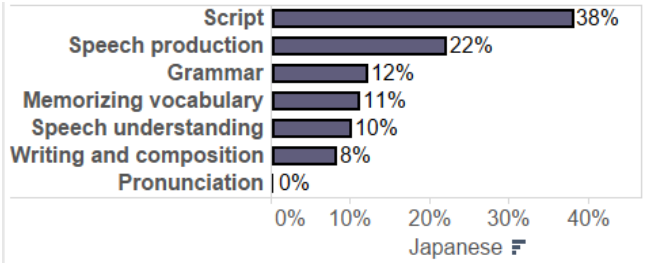


Figure 1. Japanese students' issues with the acquisition of grammar and syntax

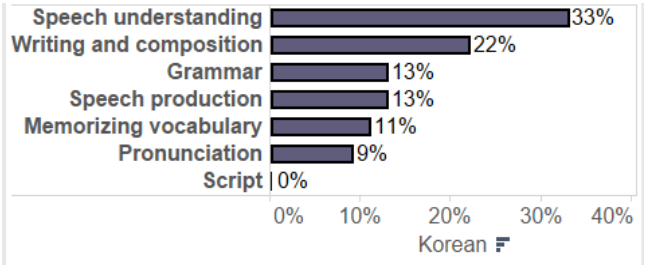


Figure 2. Korean students' issues with the acquisition of grammar and syntax

Although Hindi and Sanskrit are Indo-Aryan languages, students experience issues in learning these languages just as in learning Korean and Japanese. For these learners, speech production, writing and composition seem to present the biggest issues (see Figures 3 and 4). It is very interesting that Sanskrit students have marked writing and composition as an issue at all, as students are generally not asked to produce written compositions in Sanskrit. Their study of Sanskrit is rather the opposite of it, as they translate numerous texts from Sanskrit into Croatian.

Speech understanding is also highly problematic to Hindi learners but not so much for Sanskrit learners. Contrary again to Sanskrit grammar, Hindi grammar does not appear to be so difficult to learn. Students learning Sanskrit are facing more difficulties with the pronunciation than their fellow students of Hindi who do not find this area of language highly problematic.

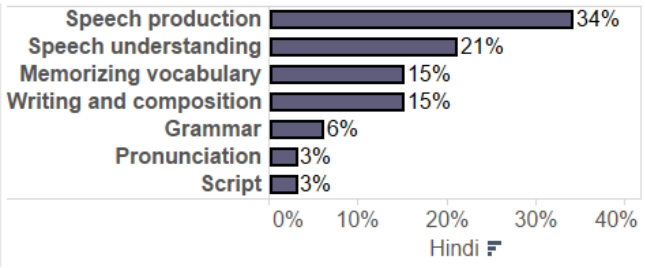


Figure 3. Hindi students' issues with the acquisition of grammar and syntax

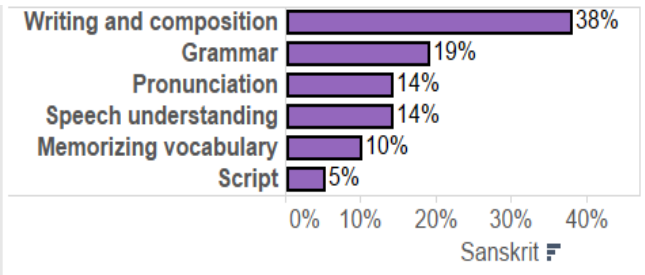


Figure 4. Sanskrit students' issues with the acquisition of grammar and syntax

IV. ATTITUDES TOWARDS E-LEARNING

The following set of questions was intended for finding about student's e-learning attitudes. We asked them about the amount of time they spend on Internet in order to learn something, not just the language. The results are ordered by the score obtained for each of the categories (Figure 5). Their answers were chosen on the scale from 1 to 5, where 1 is "Not relevant at all" and 5 is "I find it very relevant for myself".

Results for each question are above 3.60, which shows positive attitude towards their study and different elements of that study (teacher, course book, self-satisfaction). Use of Internet for educational purposes is also very high, 4.20 on average. Similarly, when we asked about relevance they would give to learning on mobile devices, such as smartphones, their estimation

[1] In Croatian language, there is a difference in grammatical forms while addressing older people or strangers, made by using different pronouns and verb endings. In those situations instead of using forms to denote second person singular, one uses markers of second person plural.

was almost identical 4.1. Their attitude towards use of such devices in language learning was very positive, with an average score of 4.5. Again, this information should not come as a surprise since most of them spend a lot of time in the proximity of mobile devices and over 60% of students engage with their mobile devices at least three to four times an hour.

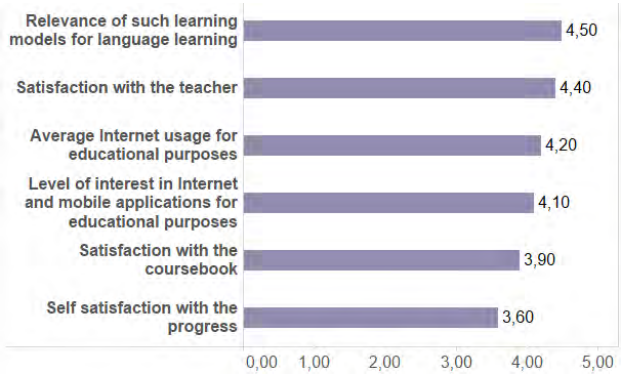


Figure 5. Students' attitudes on learning process and use of Internet and other tools.

On the other hand, only 48 out of 203 students, mostly students of Japanese language, had had a chance to engage with e-tools in their language classrooms. The results seem to be disparate, and even more so, since more than few students are aware of the potential resource that could be put in use for language learning: Duolingo, Memrise, Quizlet, Anki, Lang8, WaniKani. Resources such as Busuu, Cram, Pleco, Lingvist.io, FluentU, etc. also made it to the list of resources, when students were asked about them. It should be noted that in the second phase of the project, we used Memrise and Quizlet to develop our own materials for learning, using Croatian language and a relevant language pair. We choose these two environments due to their high quality, previous exposure with students, and adaptability in developing materials.

Apart from surveys, we also conducted testing of students' skills prior to the exposure to e-materials developed by us. They were developed in the already created environment for e-learning (Memrise and Quizlet).

By the time of writing this paper, the pre-exposure testing had been followed by only one after-exposure testing. It is thus too early to say how much the existence of e-learning materials has influenced students' results. We hope that the two additional after-exposure tests, which will be conducted in April and July 2016, will give us insight that is more relevant.

V. CONCLUSION

A number of studies anticipate positive results from the usage of technology in classroom for various purposes. Taylor (1980) and Stevenson (2008) have shown that a computer can be implemented in the classroom as either a teacher, tutee (environment), tool or a resource [5]. Several empirical studies in Croatia have shown that e-technology can be implemented well in the language teaching and teaching in general [4], [6], [7], [8], [9] and [10].

Still, there are teachers who do not use these tools for their classes, in spite of all the development of e-learning tools and materials, which already assumes the confident use of technology by both teachers and students. There are many reasons why technology should be used in the classroom and we will name here three that we find the most important: 1) it enhances teaching and learning process; 2) students learn how to use technology with the tools they are going to need in their work place; 3) students' needs are better cared for. This can further be supported with the fact that the technology is already omnipresent and it has an effect on the way people perceive information, look for information, use it, create it or learn it [11].

The question is then why the classrooms are already not swarming with materials developed for e-learning as an enhancement for the learning process. Could it be that the answer mainly lies with the teacher as several studies have shown that the interest and attitudes of teachers are also relevant [12], [13], and [14]. Namely, if teachers do not feel comfortable with technology and do not support its usage, there is not a lot that students can do even if they would like to use e-materials.

Due to the fact that we are living in the 21st century, it is valid to ask if we, as educators, are depriving our students of needed knowledge. Are we failing to educate a 21st century student? In other words, although we would like to think we are better skilled in the use of technology than our students are, they probably have a better inside information on it than us, and use it more proficiently than many of us will ever be able to.

Now, imagine a 21st century student, accustomed to learning from everything in his/her surrounding with the help of technology (from newest song or video on YouTube to a late-night chat with someone on the other side of the globe), being made to learn a language with the same type of resources our parents used, before technology explosion. There is nothing inviting or exciting in such an approach for many of them, and as the survey shows, they are aware of how technology can be implemented for educational purposes, as they are accustomed to using it on their own accord. The question is thus not so much whether students are ready to accept e-tools in language learning, but whether teachers are. Since the studies suggest that some teachers avoid technology because of their lack of skills, the good way to solve the situation would be to invest in teachers' training. There are several papers that suggest the same solution, and we as a team have created an experimental workshop for teachers at the Faculty, involved in teaching various languages, to teach them how to use Memrise and Quizlet as tools in their teaching. The anonymous survey at the end of the workshop showed that 100% of the attendants were interested in learning more about similar topics in an organized manner.

It is important to note that students' willingness to employ a certain tool in the learning process does exist

and the survey we have conducted confirms that. It has to be met by educators' attitudes towards it [15] in order to turn it into a profitable (for both student and teacher) learning-teaching tool. We are aware of issues related to the successful implementation of e-tools and e-learning materials into the classroom, but new beginnings have never been easy in any field of work. And that has never before stopped the humankind to learn and prosper.

Acknowledgment

This work was supported by the European Social Fund; grant agreement no HR.3.2.01.-0040.

References

[1] N. Saqlain. "Technology and Foreign Language Pedagogy: what the literature says". http://er.educause.edu/articles/2012/6/technology-and-foreign-language-pedagogy-what-the-literature-says

[2] E. Ushida. "The Role of Students' Attitudes and Motivation in Second Language Learning in Online Language Courses." *CALICO Journal 23(1), pp. 49-78, 2005. https://journals.equinoxpub.com/index.php/CALICO/article/view/23165/19170*

[3] B. Bellés-Fortuño, N. O. Ramírez. "Motivation: A key to success in the foreign language classroom? A case study on vocational training and higher education English courses." *1st International Conference on Higher Education Advances, HEAd'15*, 2015. DOI:http://dx.doi.org/10.4995/HEAd15.2015.431

[4] S. Librenjak, K. Vučković, Z. Dovedan Han. Multimedia assisted learning of Japanese kanji characters. *Proceedings of the 35th International Convention MIPRO,* Rijeka, 2012, pp. 1284-1289.

[5] J. Simpson, A. Walker. "New Technologies for English Language Learning and Teaching." Routledge Companion to English Studies. Eds. C. Leung, B.V. Street. 475-489, 2014.

[6] Z. Dovedan, S. Seljan, K. Vučković, "Multimedia in Foreign Language Learning", Proceedings of the 25th International Convention MIPRO 2002: MEET + MHS, Rijeka, 2002, pp.72-75.

[7] E. Družijanić Hajdarević, K. Vučković., Z. Dovedan. Računalo ili raćunalo uz pomoć računala. *Proceedings of the 29th International Convention MIPRO*, Rijeka, 2006, pp. 283-287.

[8] T. Lauc., S. Matić, N. Mikelić. Educational multimedia software for English language vocabulary. *Proceedings of the 1st International Conference on Multidisciplinary Information Sciences and Technologies: InSciT2006, Vol. I: Current Research in Information Sciences and Technologies Multidisciplinary approaches to global information systems*, Merida, 2006, pp. 117-121.

[9] M. Leko, S. Kišiček, J. Knežević, I. Martinović, P. K. Marić, E. Vusić. Interactive Application for Learning the Latin Language. *InFuture 1*, 2013, pp. 237-248.

[10] S. Sendra, J. M. Jiménez, L. Parra, J. Lloret. Blended Learning in a Postgraduate ICT course. *1st International Conference on Higher Education Advances, HEAd'15*, 2015. DOI:http://dx.doi.org/10.4995/HEAd15.2015.491

[11] G. Bautista, A Escofet. "Competences for Teaching and Learning in an e-Learning Setting." International Handbook of e-Learning, Vol. 1 : Theoretical Perspectives and Research. Eds. B.H. Khan, M. Ally, 171-179, 2015.

[12] C. Baubeng-Andoh. "Factors influencing teachers' adoption and integration of information and communication technology into teaching: A review of the literature." International Journal of Education and Development using Information and Communication Technology, Vol. 8, Issue 1, 136-155, 2012.

[13] F. Bordbar. "English teachers' attitudes toward computer assisted language learning." International Journal of Language Studies, Vol. 4 (3), 27-54, 2010.

[14] H. Peralta, F.A. Costa. "Teachers' competence and confidence regarding the use of ICT." Educational Sciences Journal, Vol. 3, 75-84, 2007.

[15] G. Wiebe, K. Kabata. "Students' and instructors' attitudes toward the use of CALL in foreign language teaching and learning." *Computer Assisted Language Learning 23(3), 221-234, 2010.*

Adaptive E-learning System for Language Learning: Architecture Overview

V. Slavuj*, B. Kovačić* and I. Jugo*
* Department of Informatics, University of Rijeka, Rijeka, Croatia
vslavuj@uniri.hr, bkovacic@uniri.hr, ijugo@uniri.hr

Abstract - **Computer–based language learning involves employing computer technology for the purpose of learning natural languages. In such educational setups, instructional software becomes a tutor responsible for the construction of learners' linguistic knowledge and development of their language competences and skills. This is possible if the software has the ability to deliver appropriate pedagogical and expert competences a language teacher possesses, such as presenting learning materials at an appropriate level, setting up and adapting tasks and learning activities, identifying deficiencies in learners' knowledge or offering constructive feedback on their work and achievements. Introducing adaptive and intelligent features to a system is thus inevitable. The paper at hand describes the architecture of one such e-learning system currently being developed at our institution. The system follows the Common European Framework of Reference for Languages (CEFR) and has a twofold purpose: a) identifying the CEFR level of language competence by employing an adaptive knowledge validation procedure, and b) systematically supporting learning by guiding learners through the learning domain towards higher proficiency levels of the CEFR. In addition, its design is based on the socio-cognitive theory of second language acquisition and the notions of activity and communicativeness.**

I. INTRODUCTION

Computer technology has long been regarded as a viable support for meaningful educational experiences [1], and the field of teaching/learning natural languages is no exception. However, over the years, the development of dedicated software has lagged behind the advances in hardware [2]. The trend has since been discontinued as the breakthroughs in related disciplines, namely Artificial Intelligence (AI) and other areas of Computer Science, enabled the development of software more suited to the needs of both teachers and learners.

Nowadays, computers are generally employed for language learning at two distinct levels which can be referred to as (1) computer-enhanced and (2) computer-based language learning. We may broadly define the former as using computer software (such as spell checkers, digital dictionaries and glossaries, or Web 2.0 tools) as a teaching aide for supporting and enriching day-to-day classroom activities and tasks so that they are carried out with more efficacy. The latter on the other hand includes the use of advanced language learning software, whether web-based or standalone, which is independent in teaching tasks and responsible for the construction of learners' linguistic knowledge and development of their language competences and skills. Moreover, computer-based language learning environments act as tutors rather than teachers in the traditional sense. A tutoring role puts special requirements on the teaching process: a piece of software needs to adapt its behaviour to suit the learners and deliver personalised instruction using various types of interaction [3]. Adaptivity feature is thus crucial in successful tutoring contexts and incorporates monitoring the learner so as to infer his or her characteristics and make appropriate adjustments to the teaching/tutoring process [4]. This is of particular importance in online environments where a wider variety of students, with a wider variety of characteristics, have access to the learning system [5].

In this paper we narrow our focus to computer-based language learning environments and characteristics of software used for tutoring purposes. The research and development of such systems is the responsibility of intelligent computer assisted language learning (ICALL), field which deals with the systematic study and application of intelligent computer technologies (adaptivity included) for the purpose of learning and teaching natural languages [6]. As pointed out in [7], the term itself is often synonymously used for systems offering any form of intelligent natural language processing, a dominant AI technique found in numerous language learning systems, most commonly employed for evaluating learner's productive language skills. However, the main goal of ICALL should be a wider one and include the availability of pedagogic and expert competencies of a teacher notwithstanding time or place. This requires significantly more than just checking written or spoken learner output, e.g.: leading the learner through appropriate learning materials, identifying shortcomings in learner's linguistic knowledge, giving feedback regarding achievement, and deciding on the order of delivery of learning activities. Introducing adaptive and intelligent features to a tutoring system is thus inevitable.

The aim of the paper is to present the architecture of one such e-learning system for English currently being developed at our institution. The rest of the paper is structured as follows: firstly, we describe and discuss the theoretical background of tutoring systems and draw attention to some of the intelligent and adaptive language learning systems currently in use; secondly, we describe the architecture of our system under development; finally, we suggest future work and conclude the paper.

II. THEORETICAL BACKGROUND AND MOTIVATION FOR SYSTEM DEVELOPMENT

A. Intelligent Tutoring Systems Technology

Technology which seems able to integrate all the above suggested requirements concerning efficient and effective language learning is intelligent tutoring systems (ITSs). ITSs are representative of the micro-adaptive approach to adaptivity [8] and are usually defined as instructional systems able to evaluate actions learners make during system use and develop appropriate models of their characteristics [9]. Information contained within such a learner model is subsequently used as the basis for making sound educational decisions that suit the needs of individual learners. Additionally, there are several other models in an ITS responsible for its behavior, namely: domain model, used for the definition and organization of the subject matter instructed by the system; instructional model, aimed at defining various learning scenarios and offering answers to the question *how* instruction is adapted; and user interface model, dealing with the delivery and presentation of learning materials to the learner.

These four models are inextricably linked and share data amongst themselves during the entire tutoring process. Moreover, the data they keep has to be accurate and up-to-date: as the tutoring process advances, learners' characteristics, needs and goals change. That is why instead of evaluating learner characteristics just once (at the beginning of instruction, upon accessing the system for the first time) and proscribing an optimal way of instruction to be applied consistently until the end, ITSs infer learner characteristics multiple times, taking into consideration both qualitative and quantitative features of learning [4][8].

B. Characteristics of Existing Intelligent Tutoring Systems

When considering examples of ITSs for language learning in existence, those developed for commercial purposes are most often encountered. This is probably due to significant costs, complexity of design, implementation and maintenance of such systems, as well as the multidisciplinary context in which they are built [10]. However, the volume of scientific publication on the topic, combined with the number of implemented systems in the last decade or so, reveals a growing trend of developing intelligent and adaptive systems for research purposes.

Among those, there are two systems that particularly stand out in overview papers and other relevant literature: E-Tutor [11], used within several university-level introductory courses of German, enabling the delivery of topical content through activities related to listening, reading and writing skills, grammar and vocabulary; and TAGARELLA [12], based on the idea of intelligent and interactive workbook for learning Portuguese, used to supplement the existing pedagogical materials with a variety of activities such as reading, listening, rephrasing, giving descriptions, doing close tasks, etc. Both systems are considered adaptive, which is evidenced in individualized feedback messages they issue to learners while trying to mend their misconceptions and gaps in linguistic knowledge. Adaptive feature in both cases is enabled by the construction and maintenance of an elaborate student model, helped by a natural language processing unit to deal with evaluating learners' written production.

Besides various properties of feedback, other elements of a system may be targeted for adaptation. Thus certain language learning systems focus on transforming learning content as is the case with the AL-TESL-e-learning system [13] and UoLmP [14]. Other systems adjust the sequence in which learning activities are presented to a particular learner, e.g., VocabTutor [15] or PIMS [16]. Nevertheless, one common feature remains: the existence of a source of adaptivity (i.e., to what is instruction adapted) implemented as a model of learner characteristics.

A recent paper [17] dealing with issues in intelligent and adaptive language learning systems revealed that there is a large number of systems oriented towards a single aspect of language learning, i.e., horizontally restricted system with regards to a single language skill (e.g., [18]) or aspect of grammar (e.g., [15]), exclusive focus on vocabulary learning (e.g., [19]), or syntactic accuracy (e.g., [20]), while systems like E-Tutor or TAGARELLA, which integrate several language skills and elements, are rather rare. Moreover, most of the existing systems are tied to a specific curriculum. This puts the emphasis on language learning achievement, rather than on language proficiency, so such systems cannot be used for general purposes. Today however there exists a tendency towards the integrative approach to language teaching and learning which brings together language skills, general and language strategies, and linguistic knowledge for the purpose of enabling communication and developing new skills.

The system described in the section to follow takes into consideration the integrative approach in the design of its learning domain, and has the aim of overcoming several of the issues and limitations mentioned above.

III. THE ARCHITECTURE OF THE PROPOSED ADAPTIVE E-LEARNING SYSTEM

The system we describe in this section, and which we are currently developing, is an adaptive, web-based e-learning system for learning English. It has a twofold purpose: 1) identifying the level of language competence by employing an adaptive knowledge validation procedure, and 2) systematically supporting learning by guiding learners through the learning domain towards higher proficiency levels.

The schematic view of the proposed system is given in Fig. 1 below. The design follows the traditional architecture of ITSs and includes all four previously discussed models. The architecture is further enriched by several modules which serve the goal of increasing the efficiency of the system during tutoring, namely evaluation module, controller module, and learner model initialization module. The rule of thumb for

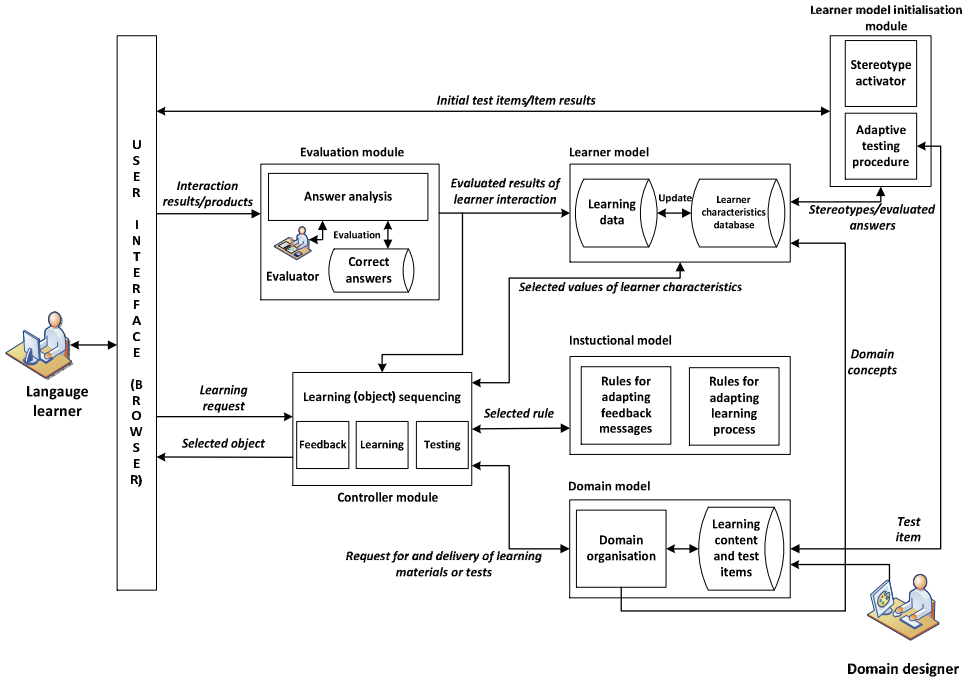


Figure 1. A schematic view of the proposed adaptive language e-learning system

differentiating between the two is that models keep latest versions of data about certain aspects of the system, while modules perform different tasks and serve as means to an end. The subsections to follow discuss each of the architectural elements and describe their functionality.

A. *Domain Model*

The system's domain model is based on the Common European Framework of Reference for Languages (CEFR), separately for several language skills (reading, listening, and writing) so as to achieve domain granularity and increase the precision of language knowledge assessment. The CEFR defines in detail language skills, strategies and linguistic knowledge of learners at each of the six predetermined levels [21]. It does so in a language-neutral form, so it is applicable in a variety of contexts, including learning, teaching, testing, policy making, etc. [22]. By choosing CEFR as the grounds for developing the domain, we are able to deal with the issue of strict focus on achievement, and move towards putting emphasis on general proficiency in the language.

The domain itself was designed by an English language expert and consists of two parts: the organization portion which establishes the relationships between domain concepts (i.e., *descriptors* developed based on the CEFR), and a database storing all learning content, including learning materials and test questions. The concepts/descriptors that make up the domain are organized as a hybrid hierarchical tree in which elements are connected by "*part of*" and "*is a*" types of relations characteristic of hierarchy domain models, but also by "*is prerequisite*" type of relations often found in network models. Enriching the hierarchy model allows us to control the dynamics of completing the domain and to simulate the usual sequence of language learning in which certain concepts are necessarily learned before others. What is more, the model establishes the relationship between learning concepts/descriptors and specially prepared authentic materials intended to accompany them. The materials follow CEFR recommendations regarding grammar and vocabulary appropriate to each level.

B. *Learner Model*

The learner model is a database with up-to-date data concerning each learner, implemented as a version of the overlay model: for each descriptor of the domain model, there is an entry in the learning model. Besides a general learner profile, which includes information such as username and password, year of starting studying English, or type of finished secondary school, there is additional data concerning the quality and quantity of domain mastery and learner activity with the system.

In case there is insufficient knowledge about a learner, a common occurrence upon first access to the system, initialization of data within the model is done via stereotypes[1], which are, in turn, formed following data provided by the module for learner model initialization. Conversely, in case there is new knowledge about a learner due to his or her interaction with the system, a timely update is performed after the quality of interaction has been assessed by the evaluation module. The rules that describe how user interaction is mapped onto the learner model are also defined within the learner model.

[1] Stereotypes can be defined as groups of characteristics shared among a certain number of people, and their corresponding values. This we consider well suited to a research context in which users of the same proficiency level need to have very similar values of particular characteristics.

C. Evaluation Module

Evaluation module of the system has the task of analyzing user's interaction with the system. Besides rating the correctness of activities, the module is concerned with monitoring other elements of interaction such as time spent on an activity or number of attempts at a question or test item, which contributes to the higher accuracy and efficiency of the learner model. The results produced by the evaluation module represent evidence on which updates to the learner model are made and constitute the basis of the evidence-oriented approach we take [23].

Besides influencing the learner model, data obtained through evaluation is sent to the controller module so that a feedback message about the completed activities may be delivered to the learner and reinforce his or her linguistic knowledge.

Due to the nature of learning domain and intricacies of free written output, some of the higher CEFR levels/bands of proficiency for writing are not included in the system. If they are to be included, one solution to the problem of assessment accuracy may be to employ a human rater in addition to a natural language processing unit.

D. Learner Model Initialization Module

In order to solve the problem of initial lack of knowledge about a new learner in the system, we have developed a procedure for determining learner's characteristics and initializing the values in the learner model. The procedure is done in two distinct phases and is presented in Fig. 2. In the first phase, the user is offered a specially designed questionnaire intended to gather initial data about the user. The gathered data is then used as input for the next phase, namely, a computerized adaptive test (CAT).

Employing an adaptive testing procedure contributes to the efficiency of the e-learning system as each learner is not presented with all the items constructed for the test (which is the case in traditional tests), but receives only a subset of items [24].

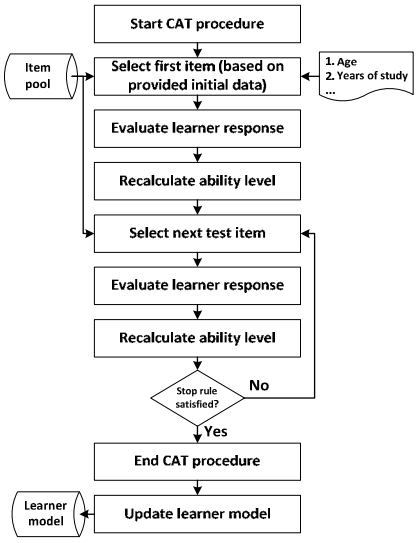


Figure 2. A diagram of the adaptive testing procedure

Test items are generated one by one and the choice of the next item to be presented depends on the correctness of the response to the previous item. Items are presented to the learner until the finishing condition is met, i.e., until the proficiency level for each language skill has been determined with particular certainty. In that way, learners are provided with their own version of the test, each with its own number of items.

The test is grounded in well known statistical methods and models, namely item response theory, and follows guidelines for good design and implementation produced by the Language Policy Division of the Council of Europe [25].

E. Instructional Model

Instructional model represents the central point of the system and the overall educational process as it defines the logic behind the system's behavior. The model is a repository of heuristic rules about teaching practice so as to approximate system's behavior to that of real teachers and their responsibilities in the language classroom. The rules are activated according to the values stored in the learner model. Upon activation of a certain rule, the controller module is able to choose the type of activity that follows (e.g., is it time for giving feedback, presenting new content, doing a remedial activity or repeating what was already introduced), as well as concrete materials from the domain model database. As noticeable from the above, we employ the approach referred to by [5] as curriculum sequencing, one of three core ITS technologies.

F. Controller Module

The controller module is the operational heart of the system. Its main task is to coordinate the work of the entire adaptive system: checking the values stored in the above discussed models, taking responsibility for their selection and application of rules, requesting appropriate teaching/learning materials from the database and sending them for presentation to the learner via user interface. Moreover, the module deals with possible learner requests for learning (when learners express their wish/need for learning particular content). In sum, the module analyses the current state of learner knowledge and makes decisions about subsequent activities with the aim of reaching higher language proficiency levels.

IV. Conclusion and Future Work

In this paper we gave a general description of the architecture of an adaptive e-learning system for language learning. The development of the system is aimed at dealing with two main issues: (1) devising a way of determining the current level of general language proficiency for each learner; and (2) offering learners structured guidance through the learning domain and helping them reach higher proficiency levels. Proficiency levels are defined using the CEFR and awarded to learners following an adaptive testing procedure, while the curriculum sequencing portion of the system is based

on defining and applying heuristic rules that relate to teaching practice.

As the system is still in its developmental phase, future work includes implementing several remaining components, namely learning materials, instructional model, controller module and user interface. After the system has been fully implemented and a test run successfully completed, its efficiency in tutoring will be empirically tested.

ACKNOWLEDGMENT

"This work has been fully supported by the University of Rijeka under the project number 13.13.1.2.02."

REFERENCES

[1] G. Blattner and M. Fiori, "Facebook in the Language Classroom: Promises and Possibilities", International Journal of Instructional Technology & Distance Learning, vol. 6(1), pp. 17-28, 2009.

[2] M. Bush, "Computer-Assisted Language Learning: From Vision to Reality?", CALICO Journal, vol. 25(3), pp. 443-470, 2008.

[3] J. Bourdeau and M. Grandbastien, "Modeling Tutoring Knowledge", in Advances in Intelligent Tutoring Systems, R. Nkambou, R. Mizoguchi, and J. Bourdeau, Eds., Springer, 2010. pp. 123-143.

[4] V. Shute and D. Zapata-Rivera, "Adaptive Educational systems", in Adaptive Technologies for Training and Education, P. J. Durlach and A. M. Lesgold, Eds., Cambridge University Press, 2012, pp. 46-66.

[5] P. Brusilovsky, "Adaptive and Intelligent Technologies for Web-based Education", Künstliche Intelligenz, vol. 13(4), pp. 19-25, 1999.

[6] M. Schulze, "Ai in CALL – Artificially Inflated or Almost Imminent?", CALICO Journal, vol. 25(3), pp. 510-527, 2008.

[7] Ö. Esit, "Your Verbal Zone: An Intelligent Computer-Assisted Language learning Program in Support of Turkish Learners' Vocabulary Learning", Computer Assisted Language learning, vol. 24(3), pp. 211-232, 2011.

[8] J. Lee and O.-C. Park, "Adaptive Instructional Systems", in Handbook of Research on Educational Communications and Technology, J. Spector, M. Driscoll, J. van Merrienboer and M. Merril, Eds., Lawrence Erlbaum Associates Inc., 2007, pp. 469-484.

[9] F. Gutierrez and J. Atkinson, "Adaptive Feedback Selection for Intelligent Tutoring Systems", Expert Systems with Applications, vol. 38, pp. 6146-6152, 2011.

[10] M. Shaughnessy, "CALL, Commercialism and Culture: Inherent Software Design Conflicts and their Results", ReCALL, vol. 15(2), pp. 251-268, 2003.

[11] T. Heift, "Developing an Intelligent Language Tutor", CALICO Journal, vol. 27(3), pp. 443-459, 2010.

[12] L. Amaral and D. Meurers, "From Recording Linguistic Competence to Supporting Inferences about Language Acquisition in Context", Computer Assisted Language Learning, vol. 21(4), pp. 323-338, 2008.

[13] Y. H. Wang and H. C. Liao, "Data Mining for Adaptive Learning in a TESL-based E-learning System", Expert Systems with Applications, vol. 38, pp. 6480-6485, 2011.

[14] S. Gómez, P. Zervas, D. G. Sampson, and R. Fabregat, "Context-aware Adaptive and Personalized Mobile Learning Delivery Supported by UoLmP", Journal of King Saud University – Computer and Information Sciences, vol. 26, pp. 47-61, 2014.

[15] G. Stockwell, "Vocabulary on the Move: Investigating an Intelligent Mobile Phone-based Vocabulary Tutor", Computer Assisted Language Learning, vol. 20(4), pp. 365-383, 2007.

[16] C. M. Chen and S. H. Hsu, "Personalized Intelligent Mobile Learning System for Supporting Effective English Learning", Educational Technology & Society, vol. 11(3), pp. 153-180, 2008.

[17] V. Slavuj, B. Kovačić, and I. Jugo, "Intelligent Tutoring Systems for Language Learning", Proceedings of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics – MIPRO, pp. 915-920, 2015.

[18] T. C. Hsieh, T. I. Wang, S. Y. Su, and M. C. Lee, "A Fuzzy Logic-based Personalized Learning System for Supporting Adaptive English Learning", Educational Technology & Society, vol. 15(1), pp. 273-288, 2012.

[19] H. Wahl and W. Winiwarter, "A Prototypical Implementation of the Intelligent Integrated Computer-assisted Language Learning (iiCALL) Environment", Proceedings of the 11th International Conference Advances in Web-Based Learning, pp. 328-333, 2012.

[20] N. Nagata, "Robo-Sensei's NLP-based Error Detection and Feedback Generation", CALICO Journal, vol. 26(3), pp. 562-579, 2009.

[21] Council of Europe, "Common European Framework of Reference for Languages: Learning, Teaching, Assessment", Cambridge: Cambridge University Press, 2001.

[22] Introductory Guide to the Common European Framework of Reference (CEFR) for English Language Teachers, Cambridge: Cambridge University Press, 2013.

[23] R. J. Mislevy, R. G. Almond, and J. F. Lukas, "A Brief Introduction to Evidence-centered Design", Research report, Princeton, NJ: Educational Testing Service, 2003.

[24] E. Guzmán, R. Conejo, and J. L. Pérez-de-la-Cruz, "Adaptive Testing for Hierarchical Student Models", User Modeling and User-Adapted Interaction, vol. 17, pp. 119-157, 2007.

[25] Association of Language Testers in Europe, "Manual for language test development and examining", Strasbourg: Council of Europe Language Policy Division, 2011.

L2L – Learn to Learn: Teach to Learn: CARTOON ENGLISH

(A constructivist approach to teaching and learning)

Krunoslav Bedi
School of Civil Engineering, Art and Craft Čakovec
Športska 1, HR-40000 Čakovec, Croatia
bedi.krunoslav@gmail.com

***Abstract* - The purpose of this research is to demonstrate the applicability of constructivist teaching / learning to secondary education with reference to constructivism as the most suitable approach to modern and quality education.**

In accordance with theoretical and practical guidelines of constructivism we designed and conducted a pilot project named "Cartoon English". Constructivist teaching methods were institutionally integrated as a part of a secondary-school subject "Multimedia Technologies" attended by students of respective ages. To assess content adoption, the students had to create an interactive-educational multimedia CD / DVD regarding the English language and containing visual material which had been drawn by pre-school children as a part of their English classes.

The research results confirm that constructivist teaching is both applicable in the classroom as well as complementary to other subjects.

Keywords: constructivism, constructivist teaching, constructivist methods, secondary education, multimedia interactive-educational CD / DVD

* Project trailer can be accessed via URL:
https://www.youtube.com/watch?v=woJno92Coso

I. INTRODUCTION

The language is one of the most basic instruments in using the basic language skills and thinking strategies that the individual will use lifelong to establish a healthy communication with surroundings during his/her self-development process. The skills of individuals to generate ideas and express these ideas develop as their language skills develop. As a result the researches conducted confirm these facts, it is started that the new approaches for language education in new programs have been adopted [5].

Constructivism is believed to take its place among the most recent approaches to foreign language teaching [2].

According to Somerset Maugham, „it is a great nuisance that knowledge can be acquired only by hard work.“ However, the process of acquiring general knowledge and various skills does not necessarily need to be difficult. On the contrary, it may in fact be challenging and innovative. Therefore, as teachers of the future, they will possibly organize classroom environments on the base of their perception [6].

In Bloom's taxonomy, the first level of knowledge is evaluated by the teacher while the students are required to retrieve previously learned information. In traditional education, teacher is the center of the knowledge and the authority to transfer knowledge to students verbally. Students mostly memorize what they are trying to learn [8].

However, memorization and repetition are not the only skills undergoing teacher evaluation. At the second level of knowledge the students are expected not only to understand and apply the knowledge but to analyse, synthesize and evaluate the educational materials.

The constructivist definition of the instruction means mainly the pupil’s inner motivation to learn, the pupil’s active constructing of the knowledge, the learning with thinking, realizing the thought and educational processes, the individualisation and differentiation in the instruction, and a social context of the learning – the interaction and the cooperation. In the constructivist definition of the instruction, the role of the teacher is changing – he/she

becomes a warranter of the method, not the warranter of the truth. The teacher becomes a facilitator of the pupils' learning, he/she helps the pupil to find the effective approaches to the learning, i.a. with the use of scale of the activating educational methods, which activate mainly the thought processes of the pupil, and the cooperative strategies of the instruction [9]. Therefore, in order to assess the third level of knowledge the teacher will test the students' ability to use their knowledge in a particular situation by demonstrating a specific skill or solving a problem.

Therefore, the process of learning is active construction, recursive and continuous. In other words, student is building their own knowledge. This means that knowledge cannot be transferred in their mental as someone pours the water into empty bottles. Instead, students construct their knowledge through the selective experience to create a conceptual structure that forms the basis of their knowledge [7].

II. CONSTRUCTIVIST TEACHING AND LEARNING

Social-cognitive theories focus on the social framework in which learning occurs (through processes of observation, modeling), the individual's interaction with the environment, the significance which he attributes to reality, highlighting, in fact, the importance of situations and social context for learning [4].

Constructivism could be defined as a theory of learning or, rather, a theory of knowledge acquisition and access to education. It places emphasis on cognitive ways and mechanisms which allow each individual to create an image of the world and determine meaning. These images could be described as different types of filters through which an individual observes reality while trying to transform the state of chaos into order. The constructivist approach to teaching and learning strives to create appropriate conditions enabling the student to gain real-life experience in a real environment. This is the only possible way of gaining real, reliable and proven knowledge while at the same time providing conditions for evaluating these skills [1].

Constructivist model assumes that knowledge is constructed in the mind of the learner. This difference in perception towards knowledge led to a change in teaching strategies in classroom [12].

There are three personality domains included in Bloom's taxonomy - cognitive, affective and psychomotor. It is not advisable to address a student while discarding any of the three aforementioned domains due to the indisputable fact that all of them are connected and mutually affect each other. Thus, the work of every teacher should be based on activities which could simultaneously be monitored / assessed while taking all three domains of a student's personality into account. Creativity lies at the heart of the 21st century education. Learner' creativity or learner's creative thinking skills are among the most important skills they need to be prepared for the knowledge society [10].

The schooling system has always had two major goals: conveying knowledge and gaining skills. Unfortunately, most teachers in an average classroom focus their activities on achieving the simplest educational goal.

Therefore, we should change the traditional teaching method, which is the mandatory classroom teaching mode, teacher inculcated the knowledge upon students. It is very important for us to seek for a new teaching method to culture student's initiative and cooperation spirit [3].

The ideology of constructivist pedagogy places learning in first plan [11]. As an approach to teaching and learning constructivism is based on the assumption that knowledge emerges from "mental construction", that is, students learn by connecting new pieces of information with the previously acquired ones. In result, experience-based learning is the most effective one. Student-student and student-teacher interactions are important ingredients of learning from a constructivist perspective [13].

Several different learning methods appropriate for children of various ages have been incorporated into the learning process itself. According to Piaget, if children lack the exposure to a stimulating experience, they cannot reach the highest possible level of cognitive development [14]. Children's individual differences affect their learning. For example, some children learn through a particular experience gained from the direct interaction with the teacher, some prefer oral presentations while others opt for theoretical literature benefiting critical thinking. Finally, the fourth group of students engages in experiments actively, learns through practical work, exercises, tasks and co-operation. Teaching is not and cannot be a one-way street. Interactive teaching encourages the necessary dialogue and exchange of opinion.

Constructive learning environment holds an assumption that learners actively construct their own set of meaning and understanding; knowledge is not a mere copy of the external world, nor is knowledge acquired by passive absorption or by simple transference from educator to another learner or knower [15].

The modern teacher should give each student the opportunity to actively participate in the learning process, conduct independent research and use it to perfect one's own cognitive abilities. Furthermore, the student should be given the opportunity to gain as much experience as possible through learning tasks causing cognitive "imbalance". It is necessary to provide the students with the construction of knowledge that leads to personality development. Teachers should have the understanding of constructivist theory, principles and pedagogy in order to provide effective teaching and learning in the technology classroom [16].

III. „L2L“ PROJECT - CARTOON ENGLISH

"L2L" (Learn to Learn) Project was designed with the purpose of encouraging learning diversities in the process of mastering English as a foreign language. Secondary-school students and pre-school children attending a language school (English learners) were made part of the project.

“L2L” project had multiple benefits. On the one hand, high-school students were performing tasks as a part of their mandatory educational programme in the field of interactive multimedia (and were learning English ☺ simultaneously). On the other hand, pre-school children were learning English actively by making drawings as a part of their English learning programme. The project activities resulted in a multimedia CD / DVD which could be used repeatedly for teaching English to pre-school children (and others) after the project itself was completed (project sustainability).

"L2L" Project has been conducted since the beginning of the academic year 2014/2015 as a part of the mandatory course "Multimedia Technologies" (3rd grade, WEB Designers). Secondary-school students (three students of Web Designers Course) were involved in the making of a multimedia interactive CD/DVD about learning English as a foreign language. Children – language school students of English – helped them to complete the project by drawing pictures which would later be implemented as an equally important part of the multimedia CD/DVD.

Project goals:

General:

- promoting various interactive forms of learning English through an educational content - multimedia interactive CD / DVD

Specific:

- developing a multimedia educational content based on real, practical and empirical examples of pre-school children engaging in a foreign language learning process
- establishing cooperation between project participants, sharing knowledge and experience (English, multimedia)
- involving children and youth in the project observation and work (drawing, workshops) resulting in improvement and deepening of their knowledge
- making an educational content about the English language which would be widely available and upgrading educational content through fun

In order to simultaneously fulfill the curriculums of both the English language (pre-school children) and multimedia application (secondary-school students), the project needed to be properly coordinated and structured. While cooperating with the teacher, the students’ project team studied each language lesson carefully. The students then digitalized and saved all materials obtained after language exercises with pre-school children had been completed. The materials used were a part of the conventional process of compulsory pre-school English learning which is based on a verified and approved textbook.

Project activities:

1. Creating a timetable and preparing project documentation - writing an elaborate
2. Drafting multimedia solutions
3. Data collection and teaching process analysis
4. Processing set design elements
5. Developing interactions (ideas)
6. Developing, customizing, processing and optimizing elements (interactivity)
7. Demonstration

Project activities were designed to support and help develop language knowledge and skills of students and teachers, increase teaching effectiveness by integrating language learning and teaching content ("Content and Language Integrated Learning", CLIL) and promote computer-assisted language learning ("Computer Aided Language Learning", CALL).

The project has had two benefits; it resulted in a new, more diverse and educational multimedia content suitable for two user groups (children and teachers). It also created a group of both direct and indirect users of educational content. The results were evaluated through media coverage of activities and materials designed as a part of the project. Alongside self-evaluation, the project objectives will be compared to the final results after project completion.

Independent assessment of implemented project activities should also be provided. In the process of activity implementation, we have noticed that simple educational content and spontaneous presentation have a beneficial effect on activity, creativity and participation. Children improve their knowledge and learn English as a foreign language through drawing, listening and observation.



Picture 1. Drawing pictures in an English language class

These skills are not lost but re-integrated each time as a (new) "input" which is later exchanged through virtual and direct transfer formin knowledge as a whole.

Because we live in a time of computer literacy, an audio-visual medium is ideal for presenting various contents. The child adapts to new media quickly (digital natives). A CD / DVD is a durable media tool which can be easily copied and distributed while its content complements English skills in non-formal education (kindergartens, language schools, etc.).

Educational material is explanatory and tailored to suit the children of pre-school age. It includes listening files as well. The project is ideal for monitoring the teaching process which includes gathering practical experience. In addition, it offers an appropriate framework for transfering knowledge and skills to anyone interested.

Available videos and photographs facilitate content understanding. The project offers the opportunity of participation and visibility to all interested kindergartens, schools and non-governmental organizations through direct implementation.

Students / users respond to this form of teaching / learning well because it represents something new and interesting.

Through activity implementation, we have noticed that content simplicity and spontaneity spark activity, creativity and participation. Different foreign language learning materials can be more easily used, improved and exchanged. Methods and tools used in this project have proven to be of positive effect on language knowledge as well as improving skills of both teachers and students.

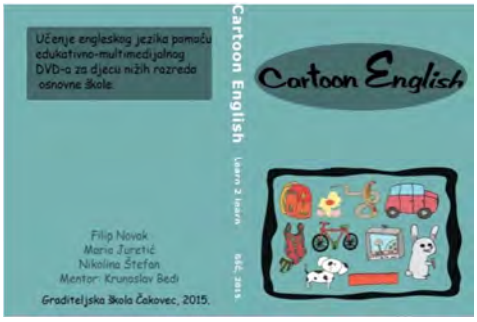
Other EU languages have not yet been made part of the project but could be included easily in the future. Languages and cultures could be connected synergically. Project implementation is very simple with results above satisfactory.

The project involves children and youth from pre-school to high-school age. Other participants from kindergartens and lower grades of primary school could also be included. Parents, kindergarten and school teachers, interested NGOs and civil initiatives as well as citizens themselves (through content promotion) could also be regarded as possible participants..

Interested citizens not participating in the project activities could benefit from the multimedia educational content as well. Digitalized content can be uploaded on the Internet very easily creating a website for a wider audience. Such content can become a part of a digital e-learning database accessible to numerous interested users.

In this way, the available contents and activities can be modernized and made recognizable.

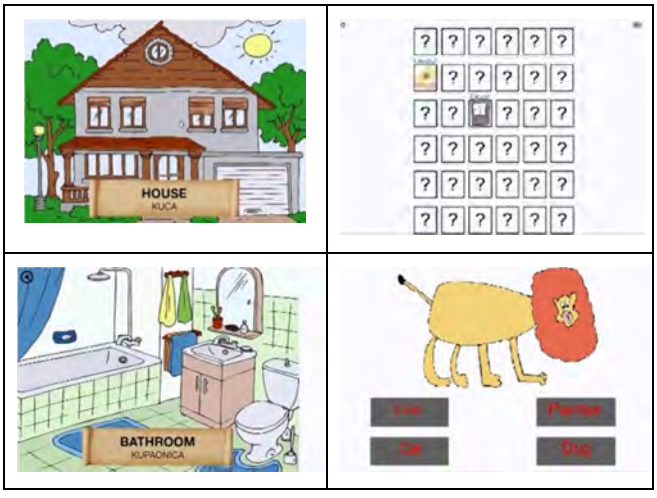
Picture 2 shows the DVD cover.



Picture 2. DVD cover

The project itself includes a greater number of children and youth, allowing them to benefit from innovative approaches to teaching foreign languages in non-formal organizations, as well as kindergartens, schools and various other institutions.

Involving kindergarten and school teachers as well as parents contributes to the feeling of integration. All participants gain the experience of becoming a part of the educational and cultural system of their community as part of the wider EU surroundings.



Picture 3. Application screenshots

IV. RESEARCH METHODS

The purpose of this research is to show the applicability of constructivist teaching / learning in secondary-school classes. The focus is to identify the activities which have a beneficiary effect on teaching as a part of a future work assignment / projects.

The survey containing seven questions was conducted on a sample of 36 secondary-school students (from first to fourth grade respectively) attending the four-year vocational school for the profession web designer in the field of multimedia. The Likert scale offering four possible choices (1 - strongly disagree, 2 - disagree, 3 - agree, 4 - strongly agree) was used for assessing the attitudes of participants who expressed their subjective agreement or disagreement with the statements.

V. RESEARCH RESULTS

The opinions of students regarding the connection between classes and daily life:

Table 1. The opinions of students

I think that….	
… we have to learn how to resolve daily-life problems outside the classroom.	
Mean	Std. Dev
4.31	0.65
… all classes have to be based on resolving daily-life problems.	
4.90	0.23
… we have to know how to apply our knowledge in daily life.	
3.89	1.13
… the knowledge gained in class improved my understanding of reality.	
4.52	0.43
… the knowledge gained in this particular class has helped me to understand reality better.	
4.19	0.91
.. in class I have learned interesting things about daily life outside school.	
3.99	0.44
... the knowledge gained in class has no connection to daily life.	
1.13	1.55

The results show high student averages. The highest answer score refers to addressing the challenges of daily life (Mean = 4.90, Std.Dev = 0.23) and knowledge which improves the understanding of reality (Mean = 4.52, Std.Dev = 0.43). Almost identical results are related to the issue of resolving daily-life problems (Mean = 4.31 , Std.Dev = 0.65). Note the shift in responses regarding the classes that would encourage students' personal interest in daily life (Mean = 3.99, Std.Dev 0.44) and the application of learned skills in daily life (Mean = 3.89 , Std.Dev = 1.13). A major shift is visible in the responses regarding daily life (Mean = 1.13, Std.Dev = 1.55) showing that students participating in these particular activities learn how to appreciate specific tasks and associate them with daily life.

VI. CONCLUSION

While taking research limitations into consideration, the sample size and the complexity of the concepts used, we were able to demonstrate that constructivist teaching is applicable in secondary school and can become a part of planning and programming in the current teaching practice. Students have maintained a positive attitude towards the learning process during the entire course of the project. Therefore, we should not question the new role of the teacher, regardless of possible fundamental changes undergoing in the classroom. Constructivist approach to teaching clearly emphasizes students' personal involvement, collaboration and informational diversity. The teacher becomes the moderator of the teaching process. Bearing in mind the limitations of the research it is necessary to conduct further surveys with a larger number of students and compile a richer statistical analysis.

Furthermore, steps of qualitative progress not to be taken in order to ensure a higher teacher engagement regardless of professional competencies. Pilot projects such as this are applicable in practice and can be beneficial for teaching.

REFERENCES

[1] http://www.pmfst.unist.hr/IEPPTZ/ Kontekstualni_pristup_seminarI.pdf, uploaded November 8, 2015

[2] Arif Sarıçoban; Students' Opinions of Foreign Language Education on Constructivist Learning Environment, Procedia - Social and Behavioral Sciences, Volume 116, 21 February 2014, Pages 2770-2773

[3] Quanyou Zhang, Qiongjie Kou; The Course Research for the Software Program Based on the Constructivism Teaching Theories, Physics Procedia, Volume 25, 2012, Pages 2294-2297

[4] Letiția Trif; Training Models of Social Constructivism. Teaching Based on Developing A Scaffold, Procedia - Social and Behavioral Sciences, Volume 180, 5 May 2015, Pages 978-983

[5] Cahit Epçaçan; Examination of Texts of Secondary School Turkish Schoolbooks in Terms of Constructivism Approach, Procedia - Social and Behavioral Sciences, Volume 116, 21 February 2014, Pages 5105-5114

[6] Yücel Kabapınar; Turkish prospective teachers' images of the constructivist view of teaching: are they compatible with the

epistemological foundations of constructivism?, Procedia - Social and Behavioral Sciences, Volume 2, Issue 2, 2010, Pages 1223-1228

[7] Ovelyn Matanluk, Baharom Mohammad, Dg. Norizah Ag. Kiflee, Molod Imbug; The Effectiveness of Using Teaching Module based on Radical Constructivism toward Students Learning Process, Procedia - Social and Behavioral Sciences, Volume 90, 10 October 2013, Pages 607-615

[8] Özge Yiğitcan Nayir, Binnur Yildirim, Hakkı İlker Koştur; Pre-service teachers' opinions about constructivism, Procedia - Social and Behavioral Sciences, Volume 1, Issue 1, 2009, Pages 848-851

[9] Čestmír Serafín, Jiří Dostál, Martin Havelka, Inquiry-Based Instruction in the Context of Constructivism, Procedia - Social and Behavioral Sciences, Volume 186, 13 May 2015, Pages 592-599

[10] Mahnaz Fatemi Aqda, Farideh Hamidi, Farhad Ghorbandordinejad; The impact of constructivist and cognitive distance instructional design on the learner's creativity, Procedia Computer Science, Volume 3, 2011, Pages 260-265

[11] Sorin Cristea; The Fundaments of Constructivist Pedagogy, Procedia - Social and Behavioral Sciences, Volume 180, 5 May 2015, Pages 759-764

[12] Khalid Alsharif; How do Teachers Interpret the Term 'Constructivism' as a Teaching Approach in the Riyadh Primary SchoolsContext?, Procedia - Social and Behavioral Sciences, Volume 141, 25 August 2014, Pages 1009-1018

[13] Cavide Demirci; Constructing a philosophy: prospective teachers' opinions about constructivism, Procedia - Social and Behavioral Sciences, Volume 9, 2010, Pages 278-285

[14] https://marul.ffst.hr/centri/circo/Nastava/Kognitivni_razvoj.pdf, uploaded November 8, 2015

[15]Fadzilah Abd Rahman, Jon Scaife; Sustaining Constructive Learning Environment: The Role of Multi-sources Regulation, Procedia - Social and Behavioral Sciences, Volume 35, 2012, Pages 180-186

[16] Moses Makgato; Identifying Constructivist Methodologies and Pedagogic Content Knowledge in the Teaching and Learning of Technology, Procedia - Social and Behavioral Sciences, Volume 47, 2012, Pages 1398-1402

Facilitating mobile learning by use of open access information resources

Radovan Vrana
Department for information and communication sciences,
Faculty of humanities and social sciences, University of Zagreb, Zagreb, Croatia
rvrana@ffzg.hr

Abstract – Global transformation of educational via ICT has resulted in development of several types of distance education including mobile learning or m-learning. Review of the current literature and results of research in this topic suggest that m-learning is currently type of learning which draws most public attention aiming to improve learning experience in general and to offer more democratic access to human knowledge. The m-learning concept consists of at least three important elements: learners (at the heart of the concept), mobile devices (means for transfer of knowledge) and learning materials. The m-learning is still not universally recognized despite its growing popularity mostly because of inability of learners to own a mobile device and lack of learning materials prepared for use on mobile devices (small screens) or inaccessibility of the learning materials in digital format. The purpose of this paper is to present and emphasize advantages of m-learning by explaining each of the essential three elements and the role of open educational resources in the development of m-learning.

I. INTRODUCTION

M-learning or mobile learning is a subcategory of e-learning, a type of distance education during which m-learners can use mobile devices as an educational technology in many locations and at their time convenience. [1] According to UNESCO, "mobile learning involves the use of mobile technology, either alone or in combination with other information and communication technology (ICT), to enable learning anytime and anywhere. Learning can unfold in a variety of ways: people can use mobile devices to access educational resources, connect with others, or create content, both inside and outside classrooms.". [2, 6] Based on analysis of the two selected definitions, we can conclude that mobile learning is a concept consisting of at least three major elements: (mobile) learning (and learners), mobile devices (technology) and content (learning materials). Each of these elements is equally significant because mobile learning as a concept cannot exist without any of them. Each of these elements is equally significant because mobile learning as a concept cannot exist without any of them. Consequently, this paper will explain each of these three elements with a special focus on content i.e. learning materials in form of open educational resources.

II. MOBILE LEARNING

Considering changes that happened so far in the educational sector since the introduction of ICT decades ago, Park [3] suggests the existence of a framework consisting of three types of distance learning. The first type of distance learning is e-learning which is usually perceived as "the effort to transform educational processes through application of different up-to-date electronic media and to customize learning to students' needs in terms of study style, culture, time, and space". [4, 1] E-learning is further expanded by m-learning, which is often seen as a subset or a subcategory of the e-learning which promotes "use of mobile technologies that increase access to education for increasingly mobile learners seeking just-in-time and just-for-me education outcomes". [4, 2] Finally, there is u-learning or ubiquitous learning which is "an environment where the computer is integral but embedded into the background of daily life". [3, 29] Depending on the level of development of educational infrastructure and access to ICT [5], we are still far from u-learning and are closest to a transitional phase between e-learning and m-learning.

Table 1. Comparison of electronic learning (e-learning), mobile learning (m-learning) and ubiquitous learning (u-learning) [3]

	E-learning	M-learning	U-learning
Physical device	Wired	Wireless	Disappeared
Computation and communication	Distinctive	->	Blurry
Learning	Confined to the single desk	->	Dynamic / flexible

Basic idea behind mobile learning is to bring "education to the learner rather than bringing the learner to a specific location to learn". [6] By doing this, mobile learning extends classroom interaction to other locations outside classroom via communication networks. [7] It brings learner to the center of learning by use of a mobile device no matter where he or she is situated at the moment of learning. As a result of these facts, mobile learning is generally understood as being learner centric instead of technology centric.

Mobile learning offers many benefits to learners: it provides equity of education by facilitating ownership of mobile devices; facilitates personalized learning; provides feedback to learners and teachers about their learning progress; supports anytime and anywhere learning; supports productive use of time in classrooms; creates communities of learning; provides support to situated learning; offers seamless learning regardless of the hardware learners use to access the content; bridges formal and informal learning; minimizes educational disruption in conflict and disaster areas; assists learners with disabilities; improves communication between participants of the learning process and administration of that process and maximizes cost-efficiency. [2] Scanlon, Jones and Waycott [8] identified another three important benefits of mobile learning: learners move around physically over time and between devices; much learning takes place outside formal learning settings and situations; and much learning is ubiquitous in nature. Evidently, mobile learning has many benefits which are acknowledged both by learners and teachers during their daily activities. Mobile learning has become popular on the global scale because of the convergence of computers and telecommunications, the internet and digitalization which allowed the creation of environments "in which all forms of information, whether visual, graphic, moving, or auditory, can be stored digitally in reusable and portable form". [4, 3] In addition to the convergence, mobile learning has also been characterized by the creation of new information environments for education. These environments will be used by mobile devices owned by students (so ownership of a mobile device became a basic prerequisite for participation in mobile learning).

III. MOBILE DEVICES

In spite of the fast technology development including development of new and advanced generations of mobile devices, their popularity and ubiquity, one should bear in mind that "mobile learning is not about the technology, it is about the learner. The learner has become mobile and is at the centre of the learning, and the technology allows the learner to learn in any context" [9, 145].

Since the early days, mobile devices have affected daily lives of their users [7] and society significantly. Mobile devices have become the primary means for people to get online; they are part of digital convergence concept; there are many applications for mobile devices for almost any purpose; many of these applications deliver location based services to users' mobile devices; they use push and pull concepts for delivering content; they change the existing telecommunication tariffs directing them more and more towards flat rate or at least telecommunication tariffs that make more sense to today's mobile device user (i.e. they are driving demand for mobile broadband and Wi-Fi networks). [11]

Students are frequent users of mobile devices and their use of mobile devices is required because "mobile learning projects tend to assume that students have uninterrupted and largely unregulated access to technology" [2, 7]. Unfortunately, the inequality in ownership of a mobile device or access to fast internet is still present in many world countries. Even if more people owned a mobile device they still wouldn't have access to the fast internet. [5]

Mobile devices are particularly important in education because they can bridge a divide between users and available online interactive services. Mobile devices are "removing existing barriers and are empowering citizens to connect to governments to access a wide range of information and services in a number of policy areas, including education". [12] In education, mobile devices are also supported by the growing number of applications which have made their use more dynamic and pervasive and they also promise more educational potential. [3]

Mobile devices have also some disadvantages: small screen, size, sometimes heavy weight, inadequate memory, and short battery life; content and software application limitations; network speed and reliability etc. [13]

The future of mobile devices in education seems to be bright, as Ally and Prieto-Blázquez [9] suggest: they will make possible more ubiquitous mobile learning; their smart systems will be available everywhere that learners can learn from; learners themselves will be mobile more than today; learners will learn from multiple sources rather than using one device; the next generation of mobile technology will be virtual, with virtual input and output capabilities. Mobile devices are constantly evolving shaping mobile learning in yet unseen ways. While mobile devices are getting more advanced and more available, there is still a constant shortage of the third element in the concept of mobile learning - content or learning materials adapted for use on mobile devices (small screens).

IV. OPEN EDUCATIONAL RESOURCES

Learning materials represent another key component of the concept of mobile learning. According to Ally and Samaka [18], many technology based education initiatives around the world fail primary because of a shortage of quality and accessible learning materials. This problem is labeled – the learning divide. The learning divide is a gap "between learners who have access to the quality learning materials and learners who do not have access to the quality learning materials because they cannot either afford to travel to other locations where the learning materials are located or pay for the high cost of textbooks and other learning materials". [14, 15] Many countries are putting a lot of effort to improve their educational systems in order to meet requirements of their respective national educational strategies and / or national labor markets. During that process, countries are coping with the growing costs of the creation and distribution of the learning materials in digital format. In their effort to make the learning materials available to every student, countries are supporting different initiatives which are oriented towards aggregation of freely available learning materials stored in open access digital repositories. The process of aggregation of the open access information resources i.e. open access learning materials advances rather slowly as the concept of open access develops. The role of open access information resources in education is important as

both financially poor and financially rich educational environments tend to increase their use instead of fee based information resources. [15],[16]

One possible solution to this problem is creation and use of open educational resources (OER). Open educational resources are "teaching, learning or research materials that are in the public domain or released with an intellectual property license that allows for free use, adaptation, and distribution". [17] Open Educational Resources "can be used for teaching, learning, research and more, are made available for free to be used with few or no restrictions". [18]

They are easy to rebuild, remix and recombine into new online courses. Open educational resources aren't still universally accepted in educational sector and are subjects to constant debate and assessment in academic community. [19]

Open educational resources include full courses, course materials, modules, textbooks, streaming videos, tests, software and any other tools, materials or techniques used to support access to knowledge. [12]

A search made in February 16h 2016 in the Directory of Open Access Repositories [20] showed 181 active repositories related to education. Further refined search showed only 50 active digital repositories related to education offering learning objects. The remaining number of digital repositories offer other types of materials such as articles, conference papers, unpublished material, books, references, multimedia, theses etc. Much of this content is not prepared for use on mobile devices (small screens) and some of the materials are not available freely or without substantial restrictions.

To understand the potential of open educational resources in education, we should look at their main characteristics: (1) OERs are free and freely available, (2) OERs are suitable for all levels of education, (3) OERs are modular, (4) reusable, and (5) online. [21] Kerres and Heinen [22] offered another view on key characteristics of open educational resources: (1) reuse: right to reuse content in its unaltered form, (2) revise: right to alter content, (3) remix: right to combine the (altered) content or parts hereof with other content to create something new, (4) redistribute: right to share copies of (revised) content or remixes to others. Open educational resources offer significant cost savings in the educational sector. [23] In general, open educational resources are widely applicable in different educational settings because they can be reused in similar courses and even repurposed for use in different courses [12]. For students, open education resources provide free access to the learning materials and cost savings, especially in case of poor students for whom access to the learning materials is limited. For teachers they provide free and legal access to some of the world's best learning materials applicable in classrooms [23]. Open education resources help institutions in exploring their potential to contribute to improved delivery of education. [24] This type of educational resources helps in achieving the goal of enabling "education for all, and the consequent goal to make information and knowledge more relevant, accessible, and useable for the good of the public who want to consume this knowledge". [25] All of these characteristics are important for wide recognition of open educational resources in educational institutions around the world and their implementation in the learning process.

To bridge the learning divide, different world initiatives facilitate access to the open access learning materials. The next part of this chapter will present (in short version) some of the most prominent Web sites offering open educational resources

OER Commons

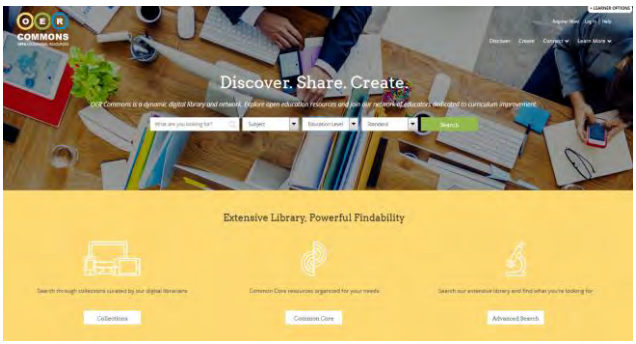


Figure 1. OER Commons (https://www.oercommons.org/)

OER Commons represents a knowledge base which provides a single point of access to the quality learning materials needed in education. OER Commons helps educators, students, and lifelong learners in searching the right educational materials offering access to over 50 000 high-quality OER. [26]

Merlot



Figure 2. Merlot (https://www.merlot.org/merlot/index.htm)

MERLOT is one of the oldest projects on the internet related to the open education resources. It "a community of staff, volunteers, and members who work together in various ways to provide users of open educational resources teaching and learning materials with a wealth of services and functions that can enhance their instructional experience". [27] It consists of 40 000 materials in 19 different material type categories inclueding discipline-specific learning materials, learning exercises, together with associated comments, and bookmark collections. [28]

Content Without Borders

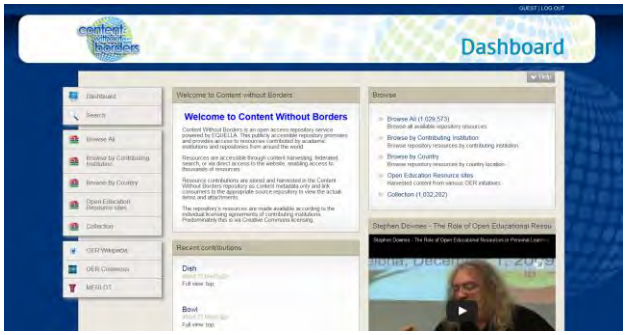


Figure 3. Content Without Borders (http://oer.equella.com/access/home.do)

Content Without Borders is an open access repository which content is contributed by academic institutions and repositories from around the world. Resources in this repository can be accessed through content harvesting, federated search, or via direct access to the website. [29]

Open textbooks

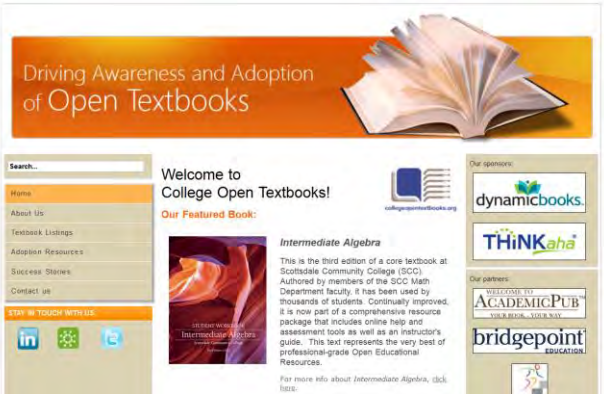


Figure 4. Open textbooks (http://collegeopentextbooks.org/)

Open textbooks is another example of open educational resources access point. In this case the Web site offers free textbooks. It is "a collection of twenty-nine educational non-profit and for-profit organizations, affiliated with more than 200 colleges, is focused on driving awareness and adoptions of open textbooks". [30]

Wikimedia Commons



Figure 5. Wikimedia Commons (https://commons.wikimedia.org/wiki/Main_Page)

Wikimedia Commons is a Web site and a media file repository "making available public domain and freely-licensed educational media content (images, sound and video clips) to everyone". [31]

Share my lesson



Figure 6. Share my lesson (http://www.sharemylesson.com/)

Share my lesson is a Web site which helps learners to share user-generated resources free of charge. The content on this Web site is developed by teachers for teachers providing access to high-quality teaching materials. This Website also offers a help from an online community where teachers can collaborate with each other. [32]

V. CONCLUSION

ICT has entered education on many levels and has been creating new educational opportunities and learning environments in which learners can access knowledge and create their personal learning style more easily than ever before. Yet, in modern education supported by ICT there are issues which still remain, to a certain degree, unresolved. The focus of this paper was on mobile learning and open educational resources as support to mobile learning. Mobile learning has become possible since the number of owners of mobile devices has grown. However, mobile learning still can't replace its predecessor e-learning (relying mostly on PCs) because mobile devices are still not universally available to all students due to the financial reasons. Mobile devices are hardly usable in education without proper digital content which would support mobile learning. There are two major problems related to the learning materials in digital format. The first one is related to fee based or commercial learning materials which are inaccessible to students and universities that can't afford them. This problem can be partially solved by use of open access information resources and their variant in education called open educational resources. Open educational resources are still being analyzed for appropriateness for general use in education on the level of content as well as on the level of pedagogy needed for use of this type of content on mobile devices. The second problem is related to the format of open educational resources as they are often unsuitable for use on mobile devices (small screens) and which are

prepared for use on devices with larger screens such as personal computers and laptops. Consequently, the lack of learning materials accessible and usable on mobile devices such as smartphones, phablets or tablets remains to be a challenge.

The development of education supported by ICT is progressing at a steady pace and is becoming more interesting as different new electronic devices appear on the market and are included in education. Their use in education is still followed by many uncertainties but this area of education offers a lot of potential for research and development.

REFERNCES

[1] M.L. Crescente, and D. Lee. "Critical Issues of M-Learning. Design Models, Adoption Processes, and Future Trends", Journal of the Chinese Institute of Industrial Engineers. vol. 28, pp. 111–123 2011.

[2] UNESCO policy guidelines for mobile learning. Paris : UNESCO, 2013.

[3] Y. Park, A Pedagogical Framework for Mobile Learning: Categorising Educational Applications of Mobile Technologies into Four Types. In: Ally, Mohamed and Tsinakos, Avgoustos eds. Increasing Access through Mobile Learning. Perspectives on Open and Distance Learning. Vancouver: Commonwealth of Learning, 2014, pp. 27-48.

[4] L. Rajasingham. „WillMobile Learning Bring a ParadigmShift in Higher Education?", Education Research International, vol. 2011, pp. 1-10. 2011.

[5] ICT Facts and Figures – The world in 2015 Retrieved January 20, 2016. from http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx

[6] M. Ally, „Mobile Learning: The Equalizer in Education" Retrieved January 20, 2016. from http://educoas.org/portal/la_educacion_digital/147/pdf/ART_Athabasca_EN.pdf

[7] E. Baran. "A Review of Research on Mobile Learning in Teacher Education". Educational Technology & Society, vol. 17, pp. 17–32. 2014.

[8] E. Scanlon, A. Jones, and J. Waycott. „Mobile technologies: Prospects for their use in learning in informal science settings", Journal of Interactive Media in Education, vol. 21, pp. 1–17. 2005.

[9] M. Ally and J. Prieto-Blázquez, "What is the future of mobile learning in education? Mobile Learning Applications in Higher Education [Special Section]", Revista de Universidad y Sociedad del Conocimiento (RUSC), vol. 11, pp. 142-151. 2014.

[10] Mobile devices on the rise: Their impact on our lives and on networks, ITU News. vol. 3. 2011. Retrieved January 20, 2016. from https://itunews.itu.int/en/533-Mobile-devices-on-the-rise.note.aspx

[11] Mobile devices on the rise: Their impact on our lives and on networks, ITU News. vol. 3. 2011. Retrieved January 20, 2016. from https://itunews.itu.int/en/533-Mobile-devices-on-the-rise.note.aspx

[12] R. McGreal. Open Educational Resources & Mobile Learning at Athabasca University. Retrieved January 20, 2016. from http://www.unisa.ac.za/blogs/wp-content/uploads/2012/05/OER-Mobile-learning-at-Athabasca-University.pdf

[13] A. Kukulska-Hulme, „Mobile usability in educational context: What have we learnt?". International Review of Research in Open and Distance Learning, vol. 8, pp. 1–16. 2007.

[14] M. Ally and M. Samaka, „Open Educational Resources and Mobile Technology to Narrow the Learning Divide". The international review of research in open and distributed learning. vol. 14, pp. 14-27. 2014.

[15] M. Talyor. Open access means a bright future for scientific research, The Guardian, 2012. Retrieved January 20, 2016. from https://www.theguardian.com/science/blog/2012/jul/17/open-access-scientific-research

[16] I. Sample. Harvard University says it can't afford journal publishers' prices, The Guardian, 2012. Retrieved January 20, 2016. from https://www.theguardian.com/science/2012/apr/24/harvard-university-journal-publishers-prices

[17] Open Educational Resources. Retrieved January 20, 2016. from http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/open-educational-resources/

[18] Helping Educators Determine the Quality of Open Education Resources. Retrieved January 20, 2016. from http://www.achieve.org/helping-educators-determine-quality-open-education-resources

[19] S. Weiland. Open Educational Resources: American Ideals, Global Questions, Global education review, vol. 2, pp. 1-23. 2015.

[20] Directory of Open Access Repositories. Retrieved January 20, 2016. from http://www.opendoar.org/

[21] A. Kanwar, B. Kodhandaraman, and A. Umar. „Toward sustainable open education resources: A perspective from the global south", The American Journal of Distance Education, vol. 24, pp. 65–80. 2010.

[22] M. Kerres and R. Heinen. „Open Informational Ecosystems: The Missing Link for Sharing Educational Resources", International Review of Research in Open and Distributed Learning, vol.16, pp. 24-39. 2015.

[23] Why should I care about OERs? Retrieved January 20, 2016. from http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/open-educational-resources/why-should-i-care-about-oers/

[24] N. Butcher and S. Hoosen. Exploring the business case for open educational resources. Vancouver: The Commonwealth of learning, 2012.

[25] N. Butcher. A basic guide to open educational resources (OER). Vancouver and Paris: Commonwealth of learning and UNESCO, 2015.

[26] About OER Commons. Retrieved January 20, 2016. from https://www.oercommons.org/about

[27] About Merlot Retrieved January 20, 2016. from http://info.merlot.org/merlothelp/index.htm#who_we_are.htm

[28] MERLOT Collection. Retrieved January 20, 2016. from http://info.merlot.org/merlothelp/merlot_collection.htm

[29] Welcome to Content Without Borders. Retrieved January 20, 2016. from http://oer.equella.com/access/home.do

[30] About College Open Textbooks. Retrieved January 20, 2016. from http://collegeopentextbooks.org/about-us/who-are-we

[31] What is Wikimedia Commons? Retrieved January 20, 2016. from https://commons.wikimedia.org/wiki/Commons:Welcome

[32] About us Retrieved January 20, 2016. from http://www.sharemylesson.com/article.aspx?storyCode=6000208&navCode=285

Work-based Learning: New Skills for New Technologies

M. Lamza Maronić, Ph.D.*, I. Ivančić, Ph.D.**
* Faculty of Economics, Osijek, Republic of Croatia
** Economic and Trade School Ivan Domac, Vinkovci, Republic of Croatia
maja@efos.hr
ivanai753@gmail.com

Abstract - Vocational education and training is designed to enable participants to develop practical skills and understanding needed to find employment in a particular occupation. Learning at the workplace is a powerful tool for integrating young people into the labour market because work-based learning is directly linked to the mission of vocational education to help learners acquire knowledge, skills and competencies which are essential in working life. The paper starts with a theoretical framework explaining the role of vocational education in the European educational environment. Subsequently, it emphasizes the role of work-based learning within vocational education in acquiring new skills for the 21st century technology. Finally, a case study of a successful application of work-based learning in a Croatian company is presented at the end of the paper. A case study of the work-based learning in Croatian current practice indicates that no matter what model of vocational education dominates in the national area, the supply and demand in the job market can be adjusted through close cooperation of educational and economic sector.

I. INTRODUCTION

Vocational education and training system is based on acquiring knowledge, skills and competences needed for the labor market. Moreover, it gives the opportunity of personal development in the higher education system as well as contributing to gaining knowledge and innovativeness of a society in general. A majority of scientists and political and economic stakeholders agree that flexibility of work and working careers, the dissemination of information and communication technologies in virtually all areas and the growing role of knowledge are increasingly calling for flexible, broadly-based and transferable skills [1]. Therefore, early identification of future skill requirements is the most important for political decisions on the design of vocational education and training systems.

Maclean and Lai [2] refer that the global economic system requires urgent and innovative responses in the field of technical and vocational education and training services, so the demand for skills is now higher than ever before. The response for that challenge is in the work-based learning concept. The work-based learning refers to learning that occurs through undertaking real work, through the enterprise-based training workshops and training classrooms. It is used in vocational education and training to develop basic work habits and specific occupational competences. The Organisation for Economic Cooperation and Development (OECD) study confirmed the existence of a strong correlation between combining work and learning and the probability of finding work after leaving education [3]. It provides confirmation of the important role of work-based learning in the development of the economy and society in general.

This paper represents a contribution to the development of this phenomenon by analysing the role of work-based learning within vocational education and training in acquiring new skills for the 21st century technology. The goal of the research is to determine the benefits of work-based learning for all stakeholders: students, employers, vocational education and training institutions, economy and society in general. It uses qualitative methodology which includes the analysis of relevant literature and publications and a case study as a useful method for in-depth research process and structure. The paper starts with a theoretical framework explaining the role of vocational education in the European educational environment. Subsequently, it emphasizes the role of work-based learning within vocational education and training in acquiring new knowledge and skills. Also, it describes learning at the workplace as a powerful tool for integrating young people into the labour market. Finally, a case study of a successful application of work-based learning in a Croatian company is presented at the end of the paper.

II. THEORETICAL FRAMEWORK

A. *Vocational education and training in European environment*

European Centre for the Development of Vocational Training (CEDEFOP) defined vocational education and training as "education and training which aims to equip people with knowledge, know-how, skills and/or competences required in particular occupations or more broadly on the labor market" [4]. It is divided into:

- mainly school-based vocational education and training where at least 75% of the vocational education or training hours are spent in school, college or training center and the balance in a work environment,

- mainly workplace-based vocational education and training where at least 75% of the vocational education or training hours are spent in a working environment and the balance in a school, college or training center and
- combined school and workplace-based vocational education and training where less than 75% of the vocational education or training hours are spent in a school, college or a training centre with the rest spent in a work environment [5]

Vocational education and qualifications have the main part in preparing young adults for the demands of labor market and its various positive outcomes have been confirmed [6]. Out of all positive outcomes, the most important is the one emphasizing that individual who graduate from vocational educational programs, in relation to those from general education programs, have better employment opportunities as well as possibilities of keeping a job [7]. Early recognition of future knowledge and skills needed for the labor market is necessary for successful creating of vocational curricula and the usefulness of education depends not only on the extent of acquiring regulatory syllabi, but on the extent of qualifying attenders for the actual labor market demands.

Ensuring employment conditions is a fundamental mechanism used by modern society in order to protect their citizens from the risks of poverty and social exclusion, whereas for each individual education is the first and the main step in achieving such employment.

These theses are prominent in key document of European development; in Europe 2020. The Member States and social partners established a new framework for European vocational education and training policy for 2010-20, which included qualitative priorities to support the Europe 2020 strategy for smart, sustainable and inclusive growth [8]. Europe 2020 cites these themes as priority ones: “Young adults on the move”, with the aim of improving outcomes of educational systems and facilitating young adults entering the labor market, and “A Programme for New Skills and Working Places” with the goal of modernizing labor market and empowering people by developing their skills during a life’s course, as well as better ratio between offer and demand of labor force, including its mobility.

A relation between education, employment and well-being is noticeable in European Employment Strategy directions, whereas employability and relevance of educational outcomes for the European Labor market make the central determinants of Bologna and Copenhagen process and the establishment of The European Qualification Framework (EQF).

Contemporary European system of vocational education and qualification responds to the demands of individuals by means of analyzing present-day and predicting future trends on the labor market, and adjusting their educational offer accordingly. Educational policies are directed towards promoting attractiveness of this system through modernizing of the curriculum and cooperation of participants interested in its functioning. The perception of the public and of the future attendants is directed mostly towards the awareness that vocational education offers the possibility of quick job attainment, this factor being critical for its improvement as well.

B. Work-based learning models

Work-based learning is directly linked to the mission of vocational education to help learners acquire knowledge, skills and competences which are essential in working life [9]. There are three main models of WBL:

(1) Apprenticeship is known in Austria and Germany as the "dual system". This model is fundamentally based on the integration of companies as training providers together with vocational education and training schools or other education and training institutes. In these programs, students spend a significant time on training in companies. In parallel, or in "alternating" periods, they acquire general and occupation-related knowledge and often complementary practical skills and key competences in vocational education and training schools or other education and training institutes. Countries with strong apprenticeship systems report very good results in terms of young people's transition to employment. Learners acquire the skills they need for a first step into the labor market, while employers train their workforce with a view of supplying the knowledge, skills and competences which they need to stay competitive,

(2) School-based vocational education and training which includes on-the-job training periods in companies. On-the-job training periods typically cover internships, work placements or traineeships that are incorporated as a compulsory or optional element of vocational education and training programs leading to formal qualifications. They can be of varying duration but typically represent less than 50% of the training program duration. They are primarily intended as effective school-to-work transition mechanisms that allow young people to familiarize themselves with the world of work and thus facilitate their transition from education to employment,

(3) WBL is integrated in a school-based program, through on-site labs, workshops, kitchens, restaurants, junior or practice firms, simulations or real business or industry project assignments. The aim is to create "real life" work environments, establish contacts and cooperation with real companies or clients, and develop entrepreneurship competencies. In this model, schools or training centers have the main responsibility for creating close to real life or real life working environments. Vocational education and training schools are equipped with school workshops, labs, kitchens and restaurants, or cooperate with business and industry to use their facilities. The mandatory share of learning in these working environments varies, depending on the type of vocational education and training. Work in practice firms are frequently used and often form mandatory parts of curricula. Teachers design learning activities in cooperation with companies. They need to develop the skills to work in multidisciplinary teams and focus on work process orientation, innovation and creation processes.

There are benefits from WBL for all parties involved, beginning with the student and the company which hosts the student and cooperates with the vocational education school, through vocational education providers themselves and up to the broader level of society [9].

WBL models that are based on intensive periods of training in the company, such as apprenticeships, offer young people the possibility to stay with the employer on completion of the apprenticeship, or reduce the time it takes to find the first job. Positive early employment outcomes are a clear benefit for students. Youth unemployment rates in European countries with a strong tradition of WBL such as Austria, Denmark and Germany are lower than in countries with less developed WBL systems. The period of time spent by graduates to find employment is also shorter. Furthermore, the more workplace contents in the education, the better the labor market outcomes.

WBL represents an important tool to react in a flexible, efficient and concrete way to the needs of the company. It also offers the opportunity to train young people according to the needs of the company. This can be of particular interest to employers in those sectors where there are skills shortages.

WBL provides opportunities for teachers to follow developments in workplace practices, processes, equipment and technology. Furthermore, good links and networking between vocational education and training providers and employers facilitate direct access for learners and teachers to the latest technology. In the case of industrial sectors, the investment required for vocational education and training schools to purchase such equipment may be too high and the possibility of cooperating with businesses makes it more cost-effective for them to train learners. On the other hand, the vocational education and training institution may be the regional hub which offers the up-to-date technology which companies can access. School and business collaboration on the provision of WBL can also enhance cooperation in areas such as curriculum design, career guidance and mentoring. Employer engagement sends a signal to the student that the vocational education and training program has value on the labor market. Such signals positively affect the image of vocational education and training. The perception of the public and of the future attendants is directed mostly towards the awareness that vocational education offers the possibility of quick job attainment, this factor being critical for its improvement as well.

III. WORK-BASED LEARNING IN CROATIAN CURRENT PRACTICE: A CASE STUDY

This section describes a case study of the company Nexus Ltd. Cerna, Republic of Croatia, which implements a work-based learning, where students have a significant time of traineeship during their secondary education.

A. Objective and research hypotheses

The purpose of this research is to confirm the benefits of implementing a work-based learning in vocational education system. The main hypothesis of the research is: the supply and demand in the job market can be enhanced through cooperation of educational and economic sector by implementation of work-based learning in the vocational education system.

B. The research methodology

The study used methods of qualitative analysis of documents and public information and implementation and qualitative analysis of an interview. The basic information about the company Nexus Ltd. Cerna was obtained by searching the publicly disclosed information on the company's website and data from internal company records.

A semi-structured interview with Mr. Alen Magdić, Director of Nexus Ltd. Cerna was carried out, with a purpose of researching the positive or negative employer's experiences in association with vocational educational sector of finding quality and competent workforce. The conversation was held on March 19th 2015 on the premises of Nexus Ltd. Cerna.

The topics of conversation were predetermined: a) the quality of vocational education system, b) cooperation with vocational educational institutions, c) recruiting students/trainees after completion of education, d) advantages of employing persons who were former trainees during their education.

The participation was voluntary and the interview lasted for 25 minutes. The answers were recorded and it was written down. The record was transcribed and minimally linguistically arranged. The interview was conducted individually, in a time and place that was convenient to the respondent and the purpose of research was explained before the interview. The respondent authorized the publication of the research results without concealing identity. The method of the semi-structured interview was chosen for the reasons of flexibility in data collection.

C. Results and discussion

1) Basic information about Nexus Ltd. Cerna

The company Nexus Ltd. Cerna started its work in April 2000 with the activities of metal machining, the production of parts for welding machines, the production of tools and devices that are used in mechanical engineering and the automotive industry. The company founders and members of society are Strama-MPS Maschinenbau GmbH & Co.KG Germany, Herman Georg Wagner and Alen Magdić (Fig. 1).

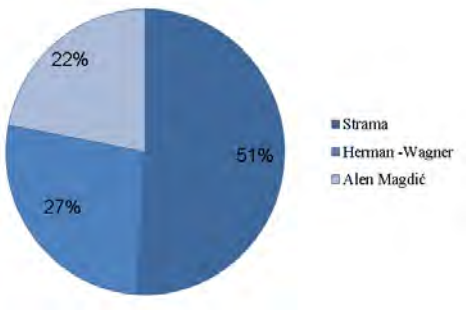


Figure 1. Nexus Cerna founders and members of society

On 17th March 2015, Nexus Ltd. Cerna had 131 employees, 9 of them were unskilled workers, 24 were skilled workers with three-year secondary education qualifications, 93 with four-year secondary education qualifications, 4 with a higher education and one employee with a university degree. The two other users were employed for professional training without employment (Fig. 2).

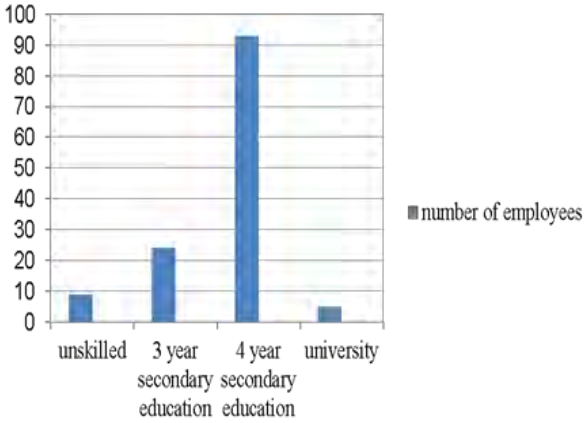


Figure 2. Number of employees in Nexus Ltd. Cerna on 17th March 2015

The company receives students for traineeship and successfully cooperates with Secondary Vocational School from Vinkovci and Trade and Industry School from Županja. In school year 2013/2014, there were 41 students on traineeship for apprentice locksmith, CNC operator and electrician (Fig. 3).

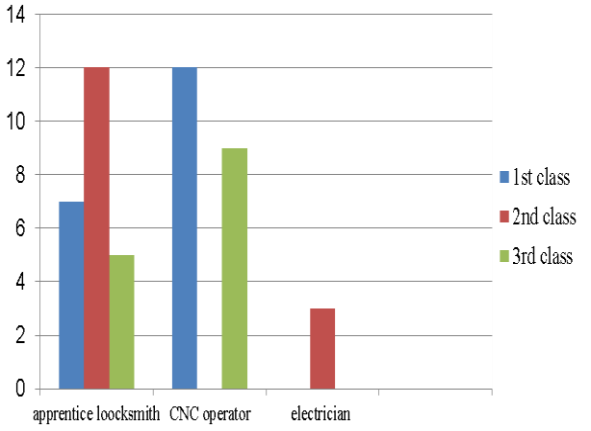


Figure 3. Students at traineeship in school year 2013/2014

During summer holidays of 2014 the traineeship was also taken by students of four-year schools for the technical professions: computer technician for mechanical engineering and technician in mechatronics.

Two of the best trainees were employed by the end of 2014, using measures to encourage employment of the Croatian employment service. In school year 2014/2015, there were 48 students on traineeship.

2) Results of the interview with the employer

Findings from the interview with Mr. Alen Magdić, Director of Nexus Ltd. Cerna have been categorized into four broad areas: the quality of vocational education system, cooperation with vocational educational institutions, recruiting students/trainees after completion of education and advantages of employing persons who were former trainees during their education. Some statements from the interview were included to illustrate the points discussed.

Employer was asked whether vocational schools prepare well enough for the labour market. He answered that vocational schools today do not prepare students well in order to get hired after graduation. He thinks there should be a better cooperation between schools and employers.

"We are trying to connect with parents and schools to get students which suit us. We did not invent anything but copied the German dual system of education."

He thinks students need to understand what is expected of them during the performance of traineeship, and the employer must have good communication with the parents and with the school.

Employer was asked whether the company accomplishes cooperation with vocational schools.

"We are mostly cooperating with the Trades and Industry School in Županja and the Technical School in Vinkovci. We accept students who are studying for CNC operators and locksmiths for our traineeships."

He emphasizes that students who had finished their school in the last year were offered positions on permanent basis. With the Technical School in Vinkovci the company cooperates through donations of computer equipment so that students can learn what is necessary for the employer. The employer is aware of the lack of funds so the schools cannot get computers and programmes, therefore the company connected the school with their software suppliers. The software manufacturers are very interested in this kind of cooperation. In fact, if students in schools learn to work in their programme, it means that the industry will eventually buy more of their software because they will be able to use it.

The employer thinks that schools do not need to change the curriculum, but teachers need to be educated and enable school learning with new equipment.

"They do not have to change those programs. It is important to learn how to apply the software that we give them."

So, he thinks that employers do not have a role in shaping the curriculum but they should be more involved in solving the problems of financing the modern teaching equipment. Also, by accepting students for traineeship, employers are able to meet quality future employees more easily.

When employer was asked whether the company employs students who have performed an internship after completing their education, he answers that the company

uses funds to encourage the employment of the youth through programs of the Croatian Employment Service.

"*But even if we do not have them, we would work like this because we have to. We do it out of self-interest, but it also helps us to invest the money which we receive into the school workshop where we pay people to work with students."*

He emphasizes that the company is unable to offer permanent employment to all trainees after completion of their education.

"We will hire the top five and the rest will be enabled to get hired anywhere in Europe, not only in Croatia. Our goal is to train quality workforce so that all employers could benefit, and if every company would do this, there would be enough high-quality workforce in the labour market."

When he was asked what the advantages of hiring people who finished their traineeship at the employers' are, the employer indicates the problem of finding workers with the competencies in the necessary profession of mechanical engineering, because there are no unemployed people in this profession. Experience has shown that when employing persons with inadequate interest, in which there are invested significant financial resources, after a certain time the workers give up and try to find a job in the profession for which they were educated.

"The advantage is when we take one student, we pay him to train and he is motivated. I have the opportunities to follow his strengths and weaknesses and this way we direct him. After three years we get a well-trained and specialized employee for a position appropriate for him. Another thing is that he was motivated to work for three years because he was paid, and the third is that I can offer him a starting salary as a master and I already know what kind of a worker he is."

During the three years of apprenticeship it is possible to estimate what the student is like, so a wrong selection of workers can be minimized. The employer thinks that accepting students to traineeship in order to be trained and eventually recognized as the best workers has many advantages.

"We accomplished it for the first time last year to cover almost all of our needs for workforce by our trainees. If someone leaves us or if there is a new position, we hire our trainees."

IV. CONCLUSION

The relevant authors and documents indicate the important role of work-based learning, being a powerful tool for integrating young people into the labour market. According to that, a work-based learning is directly linked to the mission of vocational education to help learners acquire knowledge, skills and competencies which are essential in working life. Moreover, countries with strong apprenticeship systems as Germany and Austria report very good results in terms of young people's transition to employment. In this WBL model ("dual model") learners acquire the skills they need for a first step into the labour market, while employers train their workforce with a view of supplying the knowledge, skills and competences.

The method of a semi-structured interview with an employer from Republic of Croatia is used to confirm the benefits from work-based learning for all parties involved, beginning with the company which hosts the student and cooperates with the vocational schools. An analysis of a case study is in contribution to the knowledge of the problems of work-based learning implementation in the Republic of Croatia. The answers of the director of Nexus Ltd. Cerna indicate the willingness of the employer to cooperate with the institutions of vocational education which educate students in the occupations for employer's need. The employer believes that Croatian vocational schools do not prepare students well for the labour market, but at the same time shows understanding for the lack of financial resources in the education system which results in poor school conditions and reduces the possibility of modernization and procurement of modern teaching aids. After completion of their vocational training he is ready to employ persons who performed traineeship in his company, and in such a way of employment he recognizes the benefits for employers and for young unemployed people.

A case study of the work-based learning in Croatian current practice indicates that no matter what model of vocational education dominates in the national area, the supply and demand in the job market can be adjusted through close cooperation of educational and economic sector. According to this, a research confirms that the analysed case could be considered relevant for similar cases in the implementation of work-based learning approach to address the problem of youth unemployment. Therefore, the results confirm the hypothesis that the supply and demand in the job market can be enhanced through cooperation of educational and economic sector by implementation of work-based learning in the vocational education system.

Each study has specific limitations which must be taken into account when evaluating the results. Although the advantage of the interview is that through such a method one can get an insight into the experience of the survey participants, the drawback is that this experience is subjective and there is a possibility that not enough importance was attached to some aspects of the integration of vocational education and the labour market.

REFERENCES

[1] M. Tessaring, "Identification of future skill requirements, Activities and approaches for European cooperation," Cedefop Reference series; Early identification of skill needs in Europe, Cedefop, 2003.

[2] Maclean, R. and A.Lai, "Future of Technical and Vocational Education and Training: Global challenges and possibilities," International Journal of Training Research, vol. 9(1–2), 2011, pp. 2–15.

[3] OECD, "Initial Education to Working Life: Making the Transition Work," Paris, 2000, p. 95.

[4] CEDEFOP, "Terminology of european education and training policy," Publications Office of the European Union, Luxembourg, 2014, p. 292.

[5] CEDEFOP, "From education to working life: The labour market outcomes of vocational education and training," Publications Office of the European Union, Luxembourg, 2012, p.15.

[6] P. Sahlberg, "Secondary education in OECD countries; Common challenges, differing solutions," European Training Foundation, 2007.

[7] M. Dickson and C. Harmon, “Economic returns to education: what we know, what we don’t know, and where we are going – some brief pointers,” Economics of education review, 30 (6), 2011, pp. 1118-1122.

[8] European Commission, “Europe 2020: a strategy for smart, sustainable and inclusive growth,” Communication from the Commission, 2010.

[9] European Commission, ”Work-Based Learning in Europe,” Practices and Policy Pointers, 2013.

Creating Assets as a Part of Tertiary Education of Technical Domains

Josef Brozek*, Dan Hamernik* and Zbynek Kopecky*
*Department of Information Technologies, University of Pardubice, Pardubice, Czech Republic
mail@jobro.cz

Abstract*—Game engines are becoming the forefront interest of modern-game developers. Budget savings caused by simplification of work with 3D graphics, physics, and event oriented scripting are benefits that may provide game studios with a competitive advantage. Because university education should reflect technological advances in the IT field, a concept was created on how to include Asset creation into the curriculum of tertiary education. The introductory part of the article is dedicated to Assets; it sums their principle, base benefits and practical use. The article focuses on the specifics of implementation of Asset-based education into the education process of key subjects, and monitors the results of these changes.

***Keywords*— asset, education, Unity3D, 3D graphics, scripts, information technologies, education process**

I. INTRODUCTION

Trends in the game developing industry are constantly evolving and higher computer performance allows for ever more complex projects, be it in terms of computational complexity, UI, or from the game graphics perspective. Assets and their key advantage, reusability, are for this reason becoming more and more important in agile game development.

Modern education must react to these trends, among others. Apart from emphasising interactive education, e-learning and alternative approaches to studies, it is important to watch out for practical use, and try to predict its development. Creating assets, and their subsequent application, allows for advances in classwork directed more towards the desired practical use. Apart from the combined benefits in various subjects, actuation methods are used, and at the same time, some of the soft skills are practiced – teamwork for example. The new concept also allows for more self-realisation, as students feel more fulfilled by this interactive form of education.

Assets work as packages with various content, and are created and used by students. They are in principle, similar to libraries, which programmers use – and their key feature is reusability. Assets, however, differ in deployment – instead of offering only code accessible through an interface, they offer other content as well. The content of assets is above all:

1. Elements and solutions of 2D and 3D graphics
2. Complete 3D models
3. Fragments (empty models, textures, sculpting brushes, code fragments),
4. Scripts
5. Libraries
6. Complex elements (animations, complete or partial solutions).

The essence of an Asset is, above all, maximal reusability, with minimal need for modifications. The difference between Assets and libraries lies mainly in the facts that:

- Assets allow for changes, unlike libraries, which are usually available only for use (through library function calls). For an Asset, the typical utilisation method is code extraction.
- A completed Asset may contain unfinished solutions, fragments or a solution outline as well.
- Assets may be mutually complementary, or contain complex systems created from many other assets.
- Assets do not have to be independent, and may edit some of the global items (physics settings for example), which is a disadvantage in comparison to libraries.

More information about Assets can be found at [1] and [2].

II. STATE OF THE ART

The Asset issue and the issue of Asset management is complex in terms of design quality and utilisation. Assets and asset structure and profiling were covered by the introductory chapter.

Concerning asset management, most of the AAA studios (for example Warhorse studios [3], Blizzard Entertainment [4] and Ubisoft Entertainment [5]) have their own internal asset stores, used during development of projects. The general concept is based on a database of completed assets, which are then used further inside the company (or, when the concept is generalised and outputs of the database are monetised, we get a regular asset store, as understood by layman public). These asset stores serve mainly one purpose – to ease the work of developer teams, so that they do not have to repeatedly code identical segments of code, and instead allow them to reuse code fragments from other projects. On the other hand, there are teams tasked with optimisation of existing assets and their modernisation and extension. An example of practical

application of these ideas might be reutilisation of the animations gained by the motion capture method in different game titles.

The use of one's own optimised assets during education allows the students to understand various problems with more ease, through demonstration on already completed examples. The principle might be described as a rule, that when using assets, it is not necessary to "reinvent the wheel"; instead we can proceed directly to utilisation. If we transfer this simple idea to practical use: students are, through the asset utilisation, allowed to go from elementary works to working with bigger units quickly, and thus focus more on the logic of the issue itself instead of on the lowest level problem solving.

A. Creating and providing assets

The creation of assets itself is more complex than it seems. It is necessary to keep in mind their complexity, universality and authenticity. Assets in various public commercial asset stores are created mostly by freelancers (a big community of freelancers with this focus is for example Freelancer.com [6]), professionals without a long-term contract. A freelancer is responsible for the quality of the asset, and sets its price as well. These assets may then be used by anyone who paid for them. That means some of the assets may be used independently in various games. That is unacceptable for AAA studios, and is the reason why public asset stores are almost never used by such studios. It is simpler to hire freelancers externally, or utilise the company's employees. There also are companies who are oriented on asset making, such as the OrenVideo [7] company, which created its product as a native plugin for Unity, and made it available in the asset store [8].

Similarly, various plugins by the Unikron Software Ltd [9] company are available. The issue of utilisation of the products of these studios is the accentuation on uniqueness, required by major game and application developers. The dominant players in the game development branch, also create their own asset stores. The management, recovery and stocking of these asset stores is a very interesting complementary activity for employees not currently working on a game title. This work of otherwise unutilised experts, saves a lot of money and time during later stages of the development process. This model is utilised by Valve [10], for example. Copyright laws are usually very specific in these asset stores, and data release is usually divided into several levels of access privileges, and also is subject to post release protection.

Plugins for game engines are usually licensed differently from game engines themselves, so each game engine plugin has a different license. Some examples of the licenses are for example: CreativeCommons [11] and End User License Agreement (EULA) [12]. The greatest difference in internal (private) and external (public) asset stores, lies in 3D model rights. Companies often have specific requirements and rules for their own internal models, which they uphold. In public asset stores we can see models of different quality and price. The cheapest are provided free, for commercial use, with the

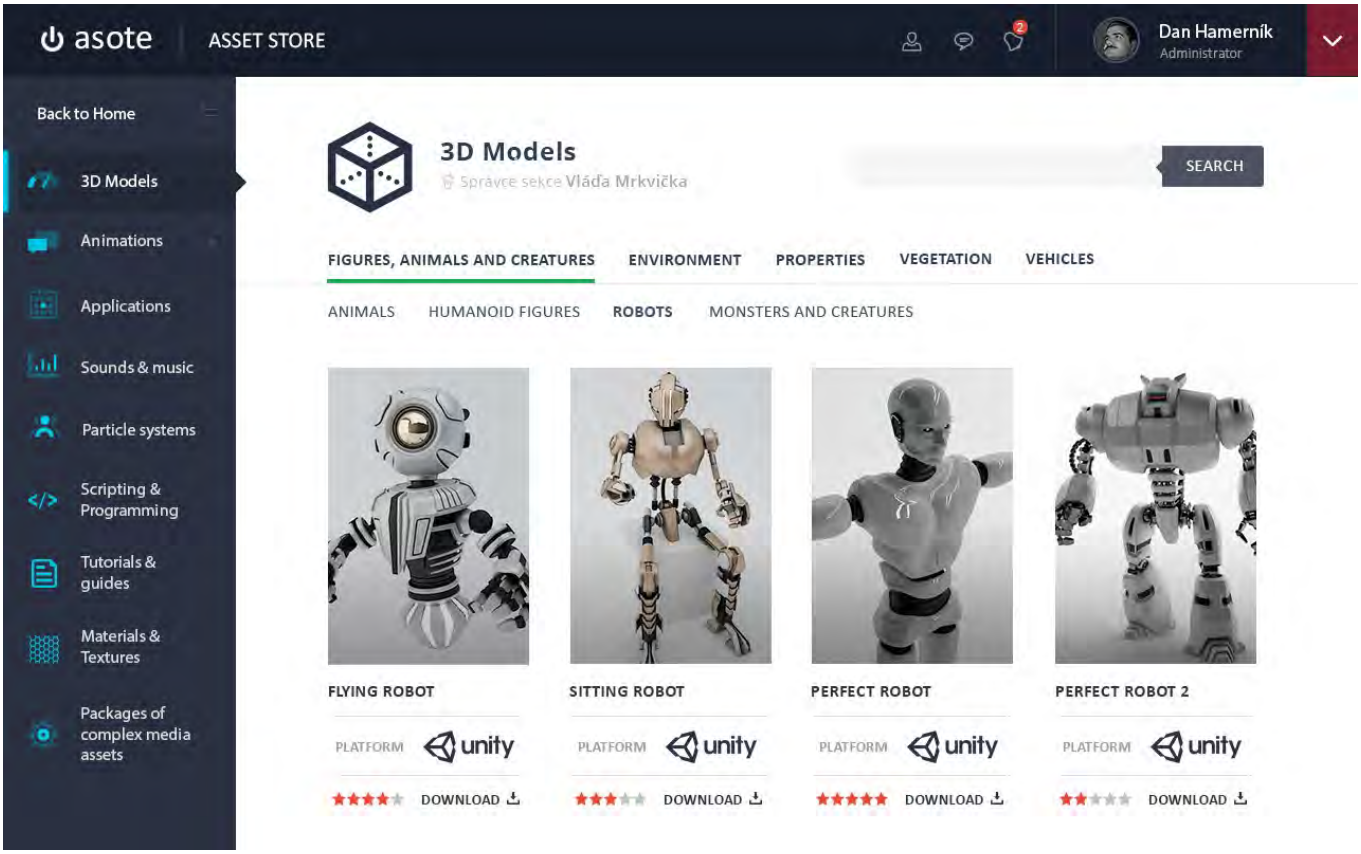


Figure 1: University Asset store

only requirement being the asset's name being listed in credits. The most expensive costing as much as 500$ [13]. One of the biggest stores offering 3D models is turbosquit [14] – but the assets sold here are notably more expensive than the assets in the Unity Asset store.

B. *Application of assets in education*

Application of assets in education presents an important innovative step forward. This trend is a novelty in the Czech Republic, although it needs to be said that some faculties, such as the Philosophical faculty of the Mendel University in Brno, and the Faculty of Science of the South Bohemian University in České Budějovice, specifically the 3D modelling subject. Abroad, assets in education are used for example in the Worcester Polytechnic Institute.

The trend is visible especially at western universities, such as the Middlesex University London [15]. Some of the other representatives are the University of Hertfordshire [16] in the UK, and RMIT [17] Melbourne, Australia.

An interesting fact is that the general practice of using assets and its potential has been known for a long time. The first vision of a similar principle being deployed comes from the year 1995, in a thesis [18] from the Delft University of Technology in the Netherlands. An example of an interesting application for visualisation, needed in health care, was described in an *unimpacted* thesis [19], which was created in cooperation with the Harvard University and Women's Hospital in Birmingham.

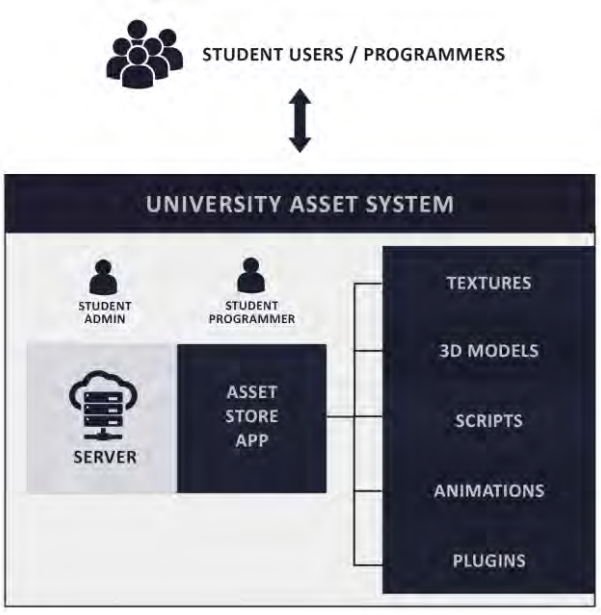


Figure 2: Diagram of asset management app implementation

III. IMPLEMENTATION OF ASSET MANAGEMENT APPLICATION

The solution of asset implementation into education originates partly from the game studios themselves. Practically speaking, if there is a process that is commercially successful and viable in the long term, it is desirable to assess its potential. The solution for education is structured into three individual parts:

1. Creative activities
2. Application activities
3. Supplementary activities

Every part is very loosely dependent on the other parts. The whole solution is illustrated in Figure 2. In the left inner area, we see the key software components – that is the Asset store app, running on a server, and service personnel. The Asset store app is interesting, as it was developed and tested by students. Students also take care of the maintenance of the whole system (as we can see in Figure 2). The system contains a database of all the individual works of students. Students can either add their works, or request works of their colleagues.

A. *Creative activities*

Creative activities represent the first part of the innovative teaching principles. Students create the content themselves.

When creating assets, the students choose one of the predefined branches, which interests them most, and then complete their own asset, which is then added to the application for asset management. The branches are among others: programming, 3D graphics, databases and database systems, game development and algorithm development. During asset creation, the students have the possibility to consult with teachers and experts in the respective fields, or students of higher years, and also authors of the most popular assets, visualization in figure 1.

The created asset is assessed before release, by a committee, primarily composed of students, and is then rated by students in terms of universality and quality. The best assets and their authors are then displayed in the asset management app in a special category, which evokes competitiveness in students and raising the quantity and quality of the created assets. A specific example might be the need to create 3D models of vehicles. A student creates a model, enters it in the application, and it may be used by other students if it passes the committee. A student who downloads such model can rate it afterwards with stars, or enter a note directly to this asset, they can also contact the author, thus creating a feedback automatically. This can be seen in figure 1.

Through the project education, students have the opportunity to experience what it is like to work on a real project, just like in practice. A student can choose a team and activity, or a student can start his own. Different members of the team then create and use different assets, consult their work with the team leader or teachers, who then rate their activity, thus granting the students feedback.

1) *Creation of complex assets*

The creation process is fully in the competence of students, and is incorporated through project education (KST/INVS1). If the student wants to participate, all he needs to do is enrol in the project education, create an asset, and enter it in the asset system. In case a team need a specific asset, a request can be made. Completed assets can be used freely by other students of

the project education to simplify the work on team projects, and individual projects alike.

The solutions created in the frame of the project education are usually the most complex, as the allotted time is used completely for practice. However, project education is interesting for students from a certain level of know-how, as the concept of the education is designed as to fixate already gained knowledge.

2) Creation of scripts

Scripts are partial program sequences, called by, for example, an event, or sequence bound to certain objects. A set of scripts can match a library of a common programming language with the exception of the black box principle – the code is open, ready for modifications.

It is beneficial to apply script creation in subjects focused on programming. Elementary use of physical models and simple logics allows for creation of assets by complete beginners. However, advanced scripts with several dependency levels are challenge even for students in the last years of the masters study programmes.

3) Creation of textures

A texture, in terms of computer graphics, is a designation for most usually a raster image, which is "glued" to a 3D model, or a 2D plane. There are several kinds of textures, with specific uses. Generally, we can split textures into raster textures and procedural textures, which contain a mathematical function, and thus do not possess a resolution.

In computer graphics, 2D textures are used along with displacement maps to create a realistic look of 3D models. When lit correctly, and with high-quality displacement maps, a 3D surface can be simulated only by a 2D plane, which reduces draw calls and thus increases FPS in computer games. The most commonly used textures are diffuse, normal, reflection, environmental textures and light refraction and absorption textures.

4) Creation of 3D models

In modern game development, the focus has shifted to realistic graphics thanks to the increasing computing power, and thus, higher requirements are made on 3D models. In spite of the increasing computational power of graphic cards, it is still necessary to create a high-quality, multi-levelled LOD (Level of Detail), as the formerly low polygon and high polygon models are not sufficient today.

When creating assets, the focus is mainly on the extreme level of detail, which is removed or lowered with the increasing LOD. Models in the asset store contain a preview, LOD and some contain textures as well. In the case of animal or human models, the models contain bones as well, as high-quality rigging is a valued attribute of a model. [20,21,22].

At the University of Pardubice, the use of assets was reflected strongly in the subject of 3D Graphic software. Students used created 3D models and then various simple animations – they were afterwards useless, and were thrown away. Now, a student can download a complex 3D model instead of modelling, and begin animating it right away, thus allowing the students to create more complex and longer animations. In case a student decides to take the other path, i.e. the model creation path, he can save them in the asset management application, and the models will not be created in vain.

5) Fragments

The inner implementation of the asset management application allows students to save fragments of assets in a special category. These fragments can be later edited and versioned, even by other users. Fragments are not rated, as they are "work in progress", and rating them would be pointless. This system allows the less experienced users to work on an idea and provide the more experienced students a possibility to realise it.

6) Re-development

Finished assets are rated in the application, and in case an asset receives a low rating, the author receives an automatically generated message. The author can edit the asset at any time, and upload a new version, just like other students. This system ensures that the assets are up to date according to the newest trends, and maintains the database and ratings up to date. [20,21,22]

7) Creation of the asset management system

During the implementation of the asset management system, the idea of functional asset stores was analysed, modified and then recreated. The system is a standard application shell over a database system. Design, categorisation and logic was implemented fully by students.

B. Application activities

Application activity is a simplified term marking the access of individual assets to the students, and the possibility of further reutilisation, or even modification.

When designing an innovative solution of an internal asset management application, the focus was set on the fact that students will be able to use a wide variety of assets covering various subjects. The application contains various tutorials and learning material, optimized and clean source codes, modern models and animations with a high-quality level of detail, sounds and music, finished applications, textures and many more.

In order to safely provide data to students and to minimise the risk of abuse, several key countermeasures had to be made. Business process model of the whole cycle of picking one specific asset is displayed in Figure 3, with the function described in the following paragraphs.

First of the several countermeasures is the need to authenticate to enter the system. After authentication, authorization is needed, which provides the user with access rights. At the same time, the activity is logged in the database of the asset store, separately after each step. All attempts to

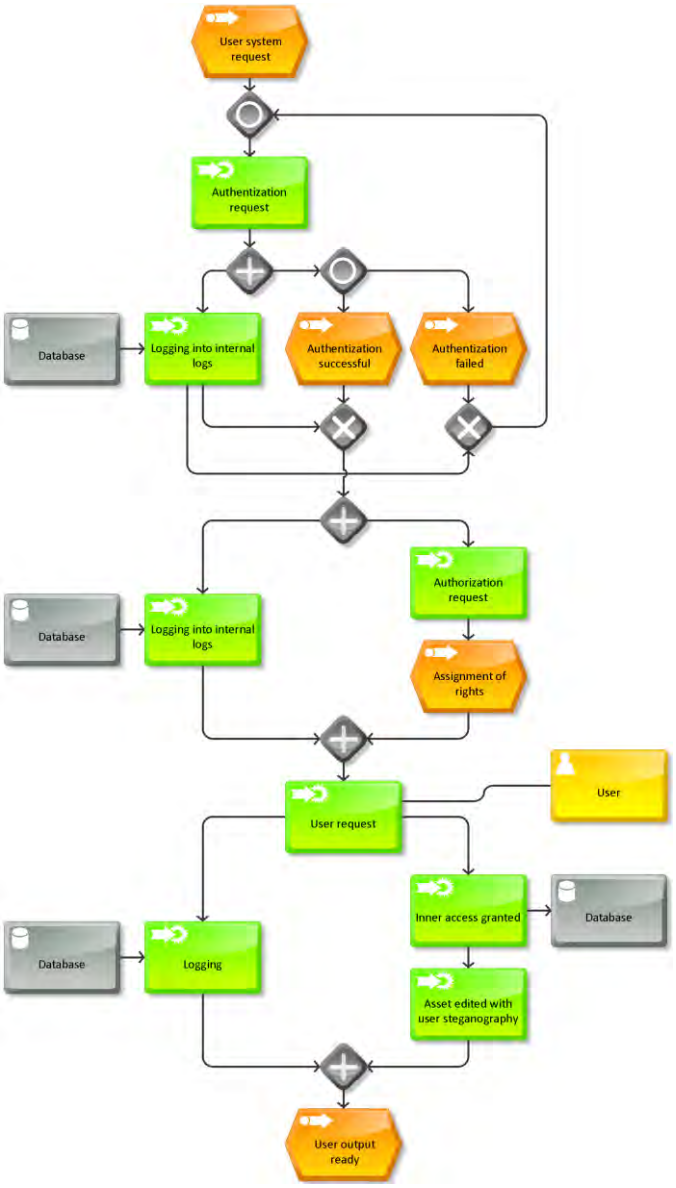


Figure 3: Diagram of obtaining an asset from the application

break into the system using brute force are very easily discovered.

After successful authorisation and authentication, the user is provided with previews of all available assets. According to the user rights, specific assets are available for download. In case download is not allowed, a student may ask for download rights (after authorisation by an administrator of the specific section, the student receives an email with a direct download link).

In case the user request is completed successfully and download was allowed, the asset (if possible) is protected through steganography. A simple principle is used, even by big Hollywood studios. Invisible information is embedded into the image / model / sound, uniquely identifying the user's download. In case of leak / unauthorised commercial use / etc. the user responsible is identified unambiguously.

C. *Supplementary activities*

A supplementary activity in terms of asset creation in education means servicing, connected with the management of the solution. Basic maintenance is ensured by the administrators of the server, UAS (University Asset System) and database.

However, other students may find application as well, as the solution requires periodical assessments, section management, review management and testing. Students may self-realise also during the commercialisation of successful assets, as it is a way of funding the server and other activities.

D. *Economic problematic*

From the economic point of view, creating and use of assets brings significant time savings, and also ensures technical foundation for projects and education. It might seem that creating assets is inconvenient because of the time requirements. After closer inspection, however, it was proven that the initial time investment into asset creation is paid, many times over, in the future. It is so because a single asset might be used several times in different subjects, and that means time saved (which would be wasted creating the asset several times). [20,21,22]

IV. Conclusion

Use of assets in the environment of tertiary education proved itself worthy in the first few case studies. Although the solution is complex in the beginning, and a complex evaluation from the didactic standpoint is possible only after deployment in a wider variety of subjects, a complex evaluation from the professional standpoint is simpler. Students are familiarised with all the original skills and knowledge through asset deployment, and also, they gain experience with actual practical development. Both motivation and results got better after deploying assets, as individual works (seminary works, half-year works) were no longer purposeless, and as they became a part of the work database. The last benefit that must be mentioned is the fact, that assets develop analytic thinking and the thought process leads to synthesis from partial elements.

Asset creation thus became a practical benefit supporting classwork with focus on algorithm development, programming and graphics. It is necessary to note that assets may also work, as for example database connectors, in various programming languages, or recordings of environment sounds, or students' auctorial music. High-quality assets enable the less experienced students, as well, to study the work of their preceding colleagues and interactively engage in the education process of subjects that interest them. Cooperation on various projects is very engaging, and opens a complex overview on modern software development with the option of specific specialisation. It is, de facto, an extension of the usual teaching

methods, such as experimentation, lecturing, practical exercises and dialog.

ACKNOWLEDGMENT

All work presented in this article was created thanks to the realisation team ASOTE: Application of Software Technologies. The team is being built without external financing from funds, state budget or University sources, and its members are primarily students, allowing for self-realisation.

This articles publication is financed through the Student grant competition of University of Pardubice.

REFERENCES

[1] O. Bouanani, "How to Fund Your Games By Creating and Selling Game Assets - Envato Tuts+ Game Development Article", Game Development Envato Tuts+, 2015. [Online]. Available: http://gamedevelopment.tutsplus.com/articles/how-to-fund-your-games-by-creating-and-selling-game-assets--cms-24380. [Accessed: 22- Feb- 2016].

[2] Gamasutra.com, "Creating Character Animation Assets", 2016. [Online]. Available: http://www.gamasutra.com/view/feature/131796/creating_character_animation_assets.php. [Accessed: 22- Feb- 2016].

[3] Warhorsestudios.cz, "Warhorse studios: HOMEPAGE", 2016. [Online]. Available: http://www.warhorsestudios.cz/index.php?page=&lang=en. [Accessed: 22- Feb- 2016].

[4] Eu.blizzard.com, "Blizzard Entertainment", 2016. [Online]. Available: http://eu.blizzard.com/en-gb/. [Accessed: 22- Feb- 2016].

[5] Ubisoft.com, "Ubisoft", 2016. [Online]. Available: https://www.ubisoft.com/en-GB/. [Accessed: 22- Feb- 2016].

[6] Freelancer.com, "Hire Freelancers & Find Freelance Jobs Online - Freelancer", 2016. [Online]. Available: https://www.freelancer.com. [Accessed: 22- Feb- 2016].

[7] Orenvid.com, "OrenVideo | Cross Platform Video Solution", 2016. [Online]. Available: http://www.orenvid.com. [Accessed: 22- Feb- 2016].

[8] Assetstore.unity3d.com, "Asset Store", 2016. [Online]. Available: https://www.assetstore.unity3d.com/en/#!/content/30226. [Accessed: 22- Feb- 2016].

[9] Assetstore.unity3d.com, "Asset Store", 2016. [Online]. Available: https://www.assetstore.unity3d.com/en/#!/content/908. [Accessed: 22- Feb- 2016].

[10] Steamcommunity.com, "Steam Community :: Steam Workshop", 2016. [Online]. Available: http://steamcommunity.com/workshop. [Accessed: 22- Feb- 2016].

[11] Creativecommons.cz, "Varianty licence | Creative Commons Česká republika", 2016. [Online]. Available: http://www.creativecommons.cz/licence-cc/varianty-licence/. [Accessed: 22- Feb- 2016].

[12] Unity3d.com, "Unity - Asset Store Terms of Service and EULA", 2016. [Online]. Available: https://unity3d.com/legal/as_terms. [Accessed: 22- Feb- 2016].

[13] Assetstore.unity3d.com, "Asset Store", 2016. [Online]. Available: https://www.assetstore.unity3d.com/en/#!/search/page=344/sortby=price/query=category:0. [Accessed: 22- Feb- 2016].

[14] Turbosquid.com, "3D Models for Professionals :: TurboSquid", 2016. [Online]. Available: http://www.turbosquid.com/. [Accessed: 22- Feb- 2016].

[15] Mdx.ac.uk, "BA 3D Animation & Games Degree | Middlesex University London", 2016. [Online]. Available: http://www.mdx.ac.uk/courses/undergraduate/3d-animation-and-games. [Accessed: 22- Feb- 2016].

[16] Herts.ac.uk, "BA (Hons) 3D Games Art & Design with Optional Sandwich Placement/ Study Abroad | University of Hertfordshire", 2016. [Online]. Available: http://www.herts.ac.uk/courses/3d-games-art-and-design. [Accessed: 22- Feb- 2016].

[17] Www1.rmit.edu.au, "Search Results - RMIT University", 2016. [Online]. Available: http://www1.rmit.edu.au/browse/About%20RMIT/Help/;CURPOS=21991. [Accessed: 22- Feb- 2016].

[18] J. Mackenzie Gavin, J. Mackenzie, G. Baily, M. Nitsche and J. Rashbass, Gaming Technologies for Anatomy Education, 2003, Online available: http://www.virtools.com/news/pdf/2004/CARET.pdf

[19] F. Jansen, P. V. Nieuwenhuizen, Computer graphics education at Delft university of technology, 1995, Online available: http://repository.tudelft.nl/assets/uuid:1de51979-1d7f-4b44-9d1d-39c4fcdab2ba/elscomp_jansen_1995.pdf

[20] J. Jirkovsky, Game Industry - Vyvoj pocitacovych her a kapitoly z herniho prumyslu. Praha: D.A.M.O., 2011.

[21] J. Jirkovsky and collective, Game Industry 2 - Vyvoj pocitacovych her a kapitoly z herniho prumyslu Praha: D.A.M.O., 2012.

[22] J. Jirkovsky and collective. Game Industry 3. Praha: D.A.M.O., 2013. 22- Feb- 2016].

Software Solution Incorporating the Steganographic Principle for Hiding Pictures within Pictures

Josef Brozek*, Jaroslav Marek**, and Vaclav Svoboda***
*Department of Information Technologies, University of Pardubice, Pardubice, Czech Republic
**Department of Mathematics and Physics, University of Pardubice,Pardubice, Czech Republic
***Department of Software Technologies, University of Pardubice, Pardubice, Czech Republic
*Jaroslav.marek@upce.cz; **mail@jobro.cz; ***svoboda@asote.cz

Abstract - Although steganography has been known since the 15th century, its bloom as a scientific branch came as late as the second half of the 20th century. Contemporary information technology allows for use of steganographic principles to place hidden content in both visual and audio media.

The article covers application of mathematical methods and software engineering methods for hiding pictures within pictures, and their subsequent recovery. The techniques described and applied for concealing pictures are: a) LSB modulation, b) Y-channel addition, c) pixel-set addition, d) frequency method and e) addition for jpg compression. Apart from introducing the methods and describing them, a software solution allowing for demonstration is also described.

The next part of the article covers the problem of concealing audio content, with the help of the following methods: a) LSB modulation, b) parity encoding, c) phase shift, d) tone addition.

I. INTRODUCTION

Steganography is a branch of cryptography. Steganography is a science discipline which studies the act of information concealing. The main purpose is to keep the communication a secret. The whole principle is based on the fact that the communication itself is never revealed. In case the message is revealed in spite of the concealing efforts, steganography may be combined with encryption methods. [1], [2].

The fate of Mary, Queen of Scots or breaking a German Enigma machine shows, how clever people can be at making secret codes and ciphers, and at breaking them.

With enough time, a man can break any secret code. For example, many people thought that famous Vigenère cipher was unbreakable without forcing one person, who shared the secret code word, to reveal the code word. During Crimean War the English scientist Charles Babbage discovered how to break it.

In stories by many writers, the main characters use secret codes or break ciphers. For example Sherlock Holmes breaked secret symbols in story Dancing man. Like Arthur Conan Doyle, Edgar Allen Poe was a writer who was interested in ciphers. His „Gold Bug story“ is not only an exciting story about the discovery of an old treasure, but also a great introduction to cryptography and code breaking. It involes cryptography with a detailed description ciphers using letter frequencies. Cf. [3]

Using radio waves, digital codes with public-key, made breaking code more difficult.

An important weakness of the secret codes is fact that everyone sees that someone is trying to keep secrets. That's why people hide secret codes. Hiding of ciphers is the main goal of steganography. By help of modern steganography photography or page of text can be hidden inside other photography and then sent by computer.

A. Concealing messages

The basics of steganography date back to the Middle Ages. The need to hide and conceal one's activities was greatest above all, in the times of war and conflict.[1]

However, it must be said that as early as 480 B.C. a Greek called Demratus sent a concealed message to Sparta, engraved into a wooden tablet. The message was covered in wax, to prevent the message from being intercepted. That can be considered to be the first use of steganography. Today's conception of steganography is different, however, and mechanical concealment is considered a category of itself, and thus it is not considered a steganographic principle.[1], [4].

The next example of hiding messages is provided by Hérodotus[3].In his book Histories he describes how Histiaeus, during a rebellion against the king of Persia, shaved the head of a slave and then tattooed it with a message. Then he waited for his hair to grow back, concealing the message. To read the message, only a shave was needed.

Shortly after the beginning of our era, a new mixture of milk and the plant called tihymalus was invented in Rome. The mixture slowly fades after being applied, creating an invisible ink. After proper heating and carbonification, the message becomes legible again. Historically, invisible ink is considered to be the most famous method of hiding messages.[5]

After 1500, Sir Francis Bacon designed a method very similar to the techniques used today. His encoding used binary expression of the letters of the alphabet. The message was written in two styles of handwriting. Every

group of five letters matched one letter of the secret message.[6]

One of the most interesting techniques from a historical point of view is the one invented by the Italian scientist Giovanni Porta. He managed to create a special ink, which soaked through the shell of a hard-boiled egg. After short period of time, the egg carried no signs of altering, but after peeling off the shell, the message was easily legible.[1]

Steganography became very popular during the Second World War. In this period, many of the previous techniques were used, but many more were invented. One of the most advanced is, without doubt, the microdot, invented by the Nazi party. The message was captured on photographic paper, since regular paper did not provide sufficient resolution. Then, it was glued in place of a regular period. The great advantage of this method was the possibility of capturing entire pages of text or pictures on the area of a single dot of ink. The biggest flaw of the method is the glossiness of the photographic paper, which is easily discovered.[7]

B. Modern steganography

With the growth of information technology, the previously unimaginable possibilities are becoming possible. Computer technologies simplified the concealing and complicated the decryption of hidden messages. Today, most of the concealed messages are transmitted in the form of pictures, and audio and video recordings. With the improved printing techniques, it is possible to print letters smaller than on hundredth of a millimetre. With binary code, it is possible to printa very big amount of data, invisible to the human eye.

The increasing necessity and popularity of steganography can be attributed to the increasing amount of music, film, art and other production. Marking by hidden signs and messages is one of the ways of at least partial prevention of breaking the copyright during illegal distribution. [8], [9], [10], [11].

Modern steganography principle check is also one of the aims of our software demonstrator.

II. STEGANOGRAPHIC PRINCIPLES

The primary element of every steganographic principle is quite simple. The object must hide information, without immediately giving away their presence, or ideally, with the information being undetectable. The principles vary in terms of insertion methods, detection difficulty, and also in their resilience to manipulation with the carrier (e.g. edits to the image carrying the message).

The basic rule when adding the message is that no data may be lost during the insertion. After extraction from the carrier, no data loss is allowed. Information we are trying to hide may be heavily altered and encoded during concealment, but there must always be a way to restore the encoded message into its original form. That is why we can never be absolutely sure that the message will not be intercepted. The risks are much greater in case the attacker means only to destroy the message, and prevent it from being delivered, instead of decoding the message.

Very often, the concealment of the message itself is done by a key, needed for the extraction process as well. In some cases, the original object is needed as well to decode the information. It is used by comparing the differences after alterations.

It is profitable to choose the steganographic principle according to the specific use (e.g. compressed vs. uncompressed carrier). Also the scrutiny of the message must be taken into account – a great difference lies between passive observation, when the message is only looked upon, and active observations, when the carrier is deformed on purpose to destroy the carried information.

Concealing secret messages into pictures is the most popular method of steganography of all. The method is based on the imperfections of the human eyesight. It relies on the fact that the human eye cannot distinguish minor degradations of the picture, caused by the encoding of secret information. It is desirable to choose a picture with high amount of detail, where this degradation is less obvious. On the contrary, pictures with many single-colour areas are not appropriate. Stated methods describe work with raster graphic. [4] [9] [11] [12] [13]

A. LSB modulation

A simple method utilising the three least significant bits (the last ones) of individual pixels for concealingpurposes. The assumption is that this alteration does not cause visible changes to the picture. Apart from the small influence the last bits have, next advantage is that only about half the bits are altered. It is possible to use more bits to conceal the information, but that leads to greater degradation of the original image.

Common images use RGB values. The last bits are altered to the bits of the message to be concealed. With three colour channels, it is possible to hide three bits of information in a single pixel. That allows for transmission of a high amount of data in a single carrier.

The message discoverability depends on the addition algorithm. When adding the bits sequentially, significant noise is created, increasing the risk of discovery. That is why it is generally the best solution to use pictures roughly matching the size of the concealed message, or to add data redundantly. However, even redundant addition (repeated addition of data to the whole image) suffers from disadvantages. A repetitive pattern can be created, increasing the risk of detection. Another problem is the increased time requirement, caused by processing of all the bits of large carriers, especially when using more complex algorithms.

B. Y-channel addition

In this method, the data carrier is transformed from the colour RGB model into the YUV model, and then the message is added into one of its components. The image is then transformed back to the original model. The main disadvantage of this method is the possibility of erroneous bits. When decoding the message, a solution similar to the encoding is used. This method is still more robust than LSB modulation, above all in the case of JPEG compression, which greatly decreases the accuracy of the U and V components.

C. Pixel-set addition

When using this method, the image carrier is split into two roughly equal disjunctive sets. The selection is made on the basis of a private key. Following, is the calculation of average intensity, which should be similar in the two sets, given the random selection process. Then, we increase the intensity of one of the sets by *k*, and decrease it by *k* in the other set. In case a watermark is used and we use a correct key, we get a difference of 2*k* during extraction.If an incorrect key is used, the values will be roughly the same again. Thus pairs of sets with similar intensity for every bit are chosen.

This principle is appropriate for hiding high amounts of information. While using only two sets, the procedure of adding a watermark is robust. It shows high resilience to changes of brightness and methods modifying the image. The image has a high probability of trimming and geometrical changes however.

D. Frequency methods

These methods use frequency expression and the encoding is executed on the created coefficients. This approach eliminates noise creation on the image carrier. However, the possibility of creating other visible changes of the images, like transitions between processed blocks, are still a liability. There are many methods of work with the coefficients, the most common one being value alteration, or change of the individual order of the coefficients.

E. Discretecosinetransform (DCT)

When using this algorithm on images, DCT is considered a 2D transformation for blocks of N x N pixels. The created matrix contains coefficients ordered in a way that the more significant (lower) frequencies are placed in the top left corner of the matrix. The less significant (higher) frequencies are placed in the bottom right corner.

It is necessary to decide which of the frequencies will be used for encoding. Encoding in lower frequencies is more robust; however, it is prone to greater degradation of the original image. That is why a compromise is usually chosen. An example of the equation itself may be as follows:

$$C_u = \begin{cases} \frac{1}{\sqrt{2}} & for\ u = 0 \\ 1 & for\ u > 0 \end{cases}$$

$$F(U,V) = \frac{2}{N} C(u)C(v) \sum_{y=0}^{N-1} \sum_{x=0}^{N-1} \left[f(x,y) \cdot \cos\left[u\pi \frac{(2x+1)}{2N} \right] \cos\left[v\pi \frac{(2y+1)}{2N} \right] \right]$$

There is a way of performing reverse DCT (IDCT), without incurring any data loss. The IDCT equation is as follows:

$$f(x,y) = \frac{2}{N} \sum_{v=0}^{N-1} \sum_{u=0}^{N-1} \left[C(u)C(v)F(u,v) \cdot \cos\left[u\pi \frac{(2x+1)}{2N} \right] \cos\left[v\pi \frac{(2y+1)}{2N} \right] \right]$$

Some of the algorithms may use the one-dimensional version of DCT. There are eight different versions of one-dimensional DCT, however only four are commonly used. The most common algorithm is the DCT-II, which is also often referred to as simply DCT:

$$F_n = \sum_{k=0}^{N-1} x_k \cos\left[\frac{\pi}{N} n(k + \frac{1}{2}) \right]$$

To this algorithm, an inverse operation DCT-III exists (after multiplication by 2/N), sometimes referred to as the IDCT:

$$F_n = \frac{1}{2} x_0 + \sum_{n=1}^{N-1} x_n \cos\left[\frac{\pi}{N} n(k + \frac{1}{2}) \right]$$

F. Discrete fourier transform (DFT)

Next of the methods using frequency modification is the Fourier transform. It is used to express a time-dependent signal. Its discrete form is used in cases where we have measured values, but not the mathematical expression of the signal. DFT has found many uses in Informatics, and is used for encoding messages as well. The formula of DFT is as follows:

$$X_k = \sum_{n=0}^{N-1} x_n \, e^{-j2\pi \frac{kn}{N}}$$

Inverse operation formula, i.e. IDFT:

$$X_k = \frac{1}{N} \sum_{k=0}^{N-1} X_k \, e^{j2\pi \frac{kn}{N}}$$

G. Discrete wavelet transform (DWT)

This is the so-called wavelet transformation. A set of filters is used. The first step is to use the lower floodgate with impulsive feedback (g), as expressed by the convolution:

$$x_n g_n = \sum_{k=-\infty}^{\infty} x_k \, g_{n-k}$$

Next the upper floodgate is used, through the upper filter (h), while creating precise coefficients from the upper floodgate and at the same time creating approximate coefficients from the lower floodgate. If half the frequencies are removed, it is possible to remove up to a half of the samples, as secured by the Nyquist rule. Then undersampling takes place:

$$y_{Down}[n] = \sum_{k=-\infty}^{\infty} x[k]h[2n-k]$$

$$y_{Top}[n] = \sum_{k=-\infty}^{\infty} x[k]g[2n-k]$$

This decomposition takes place as needed through a set of filters, creating more outputs.

H. Addition for JFIF compresion

Steganographic principles add information to compressed images, and are quite specific. Compressed image suffers significant data loss, so these methods try to utilise the specific characteristics of the compressed images. The JPEG compression is composed of the following steps:

COLOUR TRANSFORMATION – the image is transformed into the YcbCrcolor model, where the brightness element is represented by Y. Cb and Cr are the blue and red chrominance. The reason for this step is to compress the Cb and Cr components, without losses of the brightness component Y, which is utilised in the next step.

COLOUR REDUCTION –the Cb and Cr components are edited in this step. In most cases, four adjacent pixels are averaged into one value, while the brightness component Y is preserved. This lowers the data volume to roughly a half (as $2 \times \frac{1}{3} \times \frac{3}{4}$of data mass is removed). This method plays on the imperfections of the human eye, which perceives changes in brightness more accurately than changes in the colour component. Averaging the colour component of a small group of pixels thus causes minimal changes as perceived by the human eye.

LEVEL SHIFT – colour components are shifted as to correspond with the sign representation by simply subtracting 128 from values of all components.

DCT – using the 8 x 8 pixel transformation we transfer the colour components into their frequency expression. The highest coefficients can be found in the top left corner of the block, and descend gradually. These pixels have the lowest frequency and are thus the most significant. Pixels with higher frequencies may be neglected. In this phase, no data loss occurs apart from rounding errors. The coefficients still allow for high precision decoding of the original image.

QUANTIZATION – this phase has the greatest data losses of all. The individual coefficients are divided according to a rigid quantization table and rounding. There are several quantization tables, and the choice depends on the used compression. This process can be imagined as comparing the frequency curve along with its smoothing and neglecting irregularities and deviations. Quantization tables are designed in a way that causes most of the data loss in the area of higher frequencies (less effect on the resulting image).

COEFFICIENT REORDERING – coefficients are reordered with the Zig-Zag method, creating a sequence used for following encoding with the most significant changes up front..

RLE AND HUFFMAN´SENCODING – a loss-less compression of the remaining data. Not an image manipulation method in itself, it is only a decrease of the final data volume through replacement of repetitive patterns.

Steganography for JPEG compression uses DCT coefficients, into which it hides data. The problem is a serious decrease in usable coefficients, as many of them are equal to zero, which in turn means they are unfit for information encoding.

During the addition itself, it is necessary to perform inverse operation up to the point of getting the DCT coefficients in the required form. For common use, the Zig-Zag method ordering does not matter. Then information addition is done on the coefficients, which are then reversely encoded. There are many techniques for JPEG compression. Most of them are based on adjusting the frequencies created during compression. Often these frequencies are worked with in a way similar to work with audio frequencies.

I. Watermarks

Digital watermark, sometimes called a digital water-print, is a technique in Informatics which represents adding additional information into a digital document (image, sound, video). Common edits and copying of the documents will not remove the watermark. The document may carry several watermarks at the same time. Revealing them is difficult, for the algorithm used to encode them is not known.

Digital watermarks are divided in two categories – visible and hidden. A visible watermark is embedded into the document, removal of which is very difficult (due to the watermark blending in with the original image). Hidden watermarks are added into the documents with steganography. Watermarks can be also classified into robust and fragile watermarks, with the robust ones being resilient to common transformations of the document (conversion, compression), and serve to uphold the copyright laws. Fragile watermarks are used to protect the integrity of digital works, and are violated by even minor interference.

III. DEMONSTRATOR

Steganographic principle used in the application puts high stress on robustness, without causing visible changes to the carrier image. However, perfect invisibility of changes cannot be guaranteed with all carriers. That is caused by the design of the algorithm, which is most efficient during transmission of short messages.

With watermarks, invisibility and amount of added data was prioritised. Methods allow us to input whole images. They are public watermarks, meaning that only the edited carrier image is necessary for extraction, with no need to possess the original.

It must be noted that many of the steganographic principles are considered degenerative, and parts of the information of the carrier may be irreversibly damaged. An example might be Figure 1, where we can observe that the resulting image is in grey scale (although the original was coloured). These effects are covered in chapter IV.



Figure 1. Output image

To decrease time requirements, the demonstrator does not use any keys or random functions. It is not an application for real use. Time complexity is high enough without the selection of random pixels, and all used methods would suffer from overwriting errors with every next watermark applied, if random function was used. That would cause the watermark to be illegible without the necessary key, with the robustness being virtually unchanged. The implementation of the functionality itself would not be very difficult, as only appropriate bit order change would be necessary.

When using steganography to add long messages, it is advisable to check through the decoding to determine whether or not the message was added whole. When adding short messages, it is advisable to enter it redundantly, so that it is legible even when part of the data is lost. Sadly, steganographic principles do not allow for data capacity estimation, as the algorithm chooses appropriate blocks of coefficients, and only a part of the carrier's area is utilised.

The first step when launching the application, is the choice of the steganographic principle to be applied for encoding. The choice affects the supported image formats. Changing the principle unloads the loaded sample. After loading, it is possible to use a text field to input a message or load an image of the watermark. With steganographic principles, only a text message is allowed, however, with the frequency based principles, it is not possible to encode text. All other methods (LSB modulation, Y-channel and pixel-set) support both text and image input. The program scales down the loaded sample if it is too big. On the contrary, if the sample is too small, the demonstrator allows for scaling up the sample to the size of the carrier. Principles adding the watermark from white or black pixels allow for setting up the borderline brightness used in the algorithm deciding whether a pixel is black or white. With the LSB modulation principle, it is possible to set the colour channel (R, G, B, RGB) used for addition. Pixel-set principle allows for index to be set in the range of two through sixteen. The demonstrator allows saving the image in an appropriate format after editing. For extraction, only loading up the image and selecting the used encoding method is necessary.

The fundamental application options are:

1. Basic steganography
    a. Steganographic text addition
    b. LSB modulation
    c. Y channel
    d. Pixel set
2. Steganography through invisible watermark
    a. LSB modulation
    b. Y channel
    c. Pixel set
    d. Frequency method

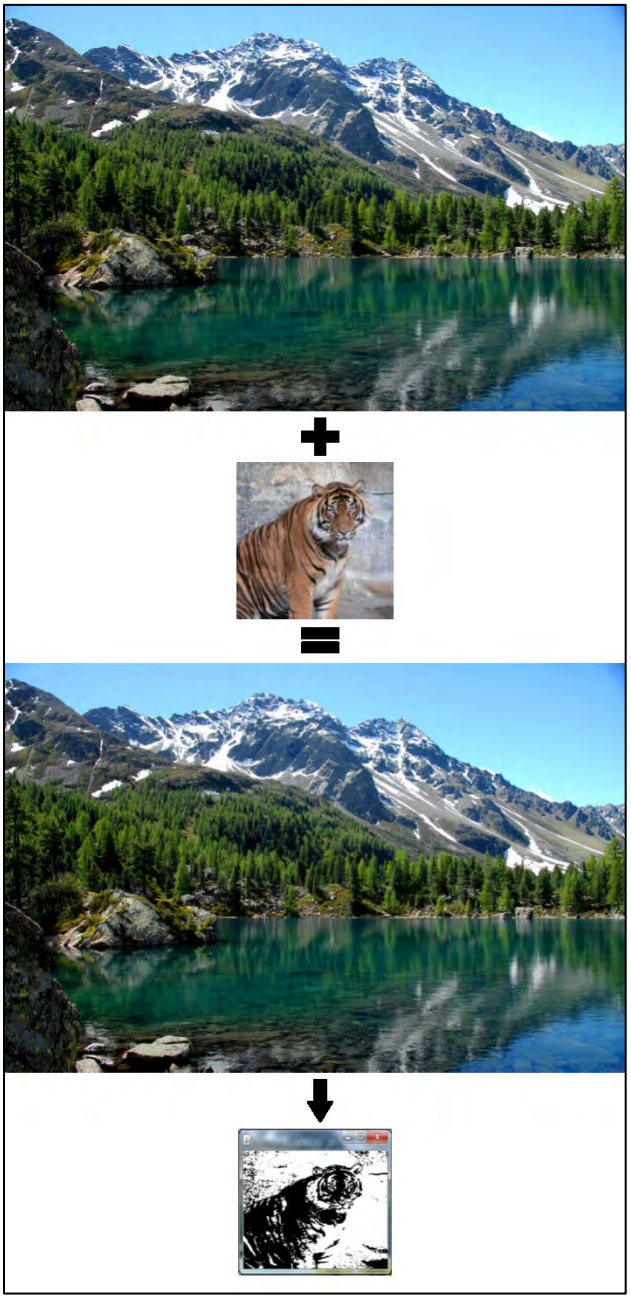


Figure 2. Image processing

IV. DISCUSSION

During the testing phase of the algorithms and their deployment, several events were discovered, that are not always mentioned in literature. In terms of result interpretation, they were emphasised in this article.

A. Image degeneration

In Figure 2, the individual steps of watermark addition and its decoding are depicted. From the top, we can see: carrier image original. The image was chosen intentionally, as a carrier image should not contain large areas of one colour, where watermark addition would be visible in plain sight. Next the image is inserted as a watermark. An image several times smaller than the carrier image was chosen on purpose. This image was then added to the carrier image's B channel through LSB modulation. The previous image is the carrier image with the watermark in place. No changes are visible to the naked eye, the only difference being an increased file size. This file was then decoded, again with the carrier image's B channel through LSB modulation, which resulted in the last image, where we can see the original watermark. Although some deformation occurred in comparison with the original watermark file, it is still recognisable. With this method, the deformation is considerable, but is balanced out by smaller file size increase, and faster execution of the encrypting algorithm.

In Figure 1, we can see the deciphered watermark when using the LSB method utilising all channels (RGB). It is immediately obvious, that the decoded image is far more superior in quality. The disadvantage is the increased file size, and slower run of the encrypting and decrypting algorithm.

B. Image file size increase

A big disadvantage of most of the steganographic principles is the change of file size. When using the LSB modulation principle with B channel, the original file sizes of 1350 kB for the carrier image and 129 kB for the watermark image (both in the png format) we obtain an encrypted image of 1590 kB (pngformat).

C. Calculation complexity

Another big issue is calculation complexity. The application was tested on a laptop with 4 GB of RAM, where the encoding process with carrier image size of roughly 10 Mpx took a disproportionately long time. That is why a carrier image with the resolution of 1024 x 672 pixels was chosen. The application was then tested on a machine with 64 GB of RAM. Since the demonstrator was created as a single-threaded application, the resulting times did not improve greatly.

D. Possible use potential

Despite the obvious disadvantages, steganography has many different uses, above all:

1) Use in education

Steganography could be used in education to discover cheating. Students' files could be signed with a watermark through scripting. It would then be very easy to discover not only who was copying, but also from whom he was copying.

2) Use in project works

If a group of people works on a single project, it is ideal if everyone has a steganographically signed copy. If the project is then leaked, it is very easy to identify the person responsible. This method is often used by big Hollywood studios to discover who is responsible or data leak.

3) Targeted author identification

If we edit an image, or we have an original author craft, we can mark it steganographically, or with a watermark. If later a law-suit arises with someone who used our image, it is identified. The advantage of visible digital watermark is the blending with the original picture, which in turn means difficult removal. Hidden watermarks are the perfect tool to support ownership rights of documents. Often even the perpetrator has no idea, that the document has a watermark. It is important to choose the objective of the watermark. If we require indestructibility, we choose robust methods. If we require a proof of unauthorised manipulation with the document, we choose fragile methods.

V. CONCLUSION

The demonstrator checked the steganographic principles and watermark addition. The application is fully functional, although a real use would require more complex methods. The inaccuracies of the frequency based methods are caused by the inaccuracies of calculation. The main reason of data loss of steganography is the JPEG format compression, but the losses in the solution are adequate. During encryption and following extraction of the watermarks, the used methods do not require possession of the original image with no alterations made. Due to redundancy, encrypting key was not used in the demonstrator, for calculation times were a greater priority. In spite of this fact, the time requirements are considerable when using high-resolution carrier images.

More complex and robust methods would be necessary for real use, however, the main idea would remain identical to that of the demonstrator.

The aim of future research is decrypting steganographic principles and watermarks, finding ways of revealing the watermark and destroying it, while keeping the damage to the carrier file minimal.

ACKNOWLEDGMENT

The software was created thanks to the master's thesis of Filip Krudenc [14], as led by J. Marek, the leader of the Department of Mathematics and Physics. The thesis was properly defended before the state exam comitee. The authors would like to hereby thank the former student. Creating the scenarios, method attestation, weighing the advantages and disadvantages, code checking and preparation of a specialised publication was entrusted to the team of Applied software technologies of the Department of Information Technology of the Faculty of Electrical Engineering and Informatics, University of Pardubice, whose members are co-authors of this article.

The article was created with the financial support of the University of Pardubice, and is financed through the project of the student grant competition SGS2016_020.

REFERENCES

[1] G. Kipper, Investigator's guide to steganography. Boca Raton, FL: Auerbach Publications, 2004.

[2] K. Bailey and K. Curran, Steganography. S.l.: Booksurge, 2005.

[3] J. Marek, K. Deligiannaki and M. Gkosdi, Statistical View on the "Gold Bug" Story of E. A. Poe. In: Proceedings 15th Conference of Applied Mathematics Aplimat 2016, STU Bratislava. pp. 799-806.

[4] J. Fridrich, Steganography in digital media: principles, algorithms, and applications. New York: Cambridge University Press, 2010.

[5] W. Mazurczyk, S. Wendzel, S. Zander, A. Houansade and K. Szczypiorski, Information Hiding in Communication Networks: Fundamentals, Mechanisms, and Applications (IEEE Press Series on Information and Communication Networks Security). 1st ed., Wiley-IEEE Press, 2016.

[6] T. Calabrese, Information security intelligence: cryptographic principles and applications. Clifton Park, NY: Delmar Learning, 2004.

[7] N. F. Johnson, Z. Duric and S. Jajodia, Information hiding: steganography and watermarking : attacks and countermeasures. Boston: Kluwer Academic, 2001.

[8] S. Lian and Y. Zhang. Handbook of research on secure multimedia distribution. Hershey, PA: Information Science Reference, 2009.

[9] S. Katzenbeisser and F. A. Petitcolas, Information hiding techniques for steganography and digital watermarking: Stefan Katzenbeisser, Fabien A.P. Petitcolas, editors. Boston: Artech House, 2000.

[10] M. Kumar, Steganography and steganalysis of JPEG images: a statistical approach to information hiding and detection. Lexington, KY: LAP Lambert Academic Pub, 2011.

[11] F. Y, Shinh, Digital watermarking and steganography: fundamentals and techniques. Boca Raton: CRC Press, 2008.

[12] I. Cox, Digital watermarking and steganography. 2nd ed. Boston: Morgan Kaufmann Publishers, Morgan Kaufmann series in multimedia information and systems, 2008.

[13] M. Dobes, Zpracovani obrazu a algoritmy v C#, 1st ed., Praha: BEN - technicka literatura, 2008.

[14] F. Kudrnec, Steganografie a vodoznaky, Pardubice: Univerzita Pardubice, 2013.

A platform independent tool for programming, visualization and simulation of simplified FPGAs

Marko Čupić, Karla Brkić, Željka Mihajlović
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10 000 Zagreb, Croatia
{marko.cupic, karla.brkic, zeljka.mihajlovic}@fer.hr

***Abstract*—During the last few decades complex programmable circuits have seen a widespread usage in various digital circuit applications. One prominent example are Field Programmable Gate Arrays (FPGAs). Teaching FPGA technology has become an integral part of introductory digital logic courses. However, implementing Boolean functions in this technology requires understanding of several steps that are not trivial, including Boolean function decomposition, mapping the design into physical programmable units and routing. We have developed a portable Java-based tool which allows students to experiment with the required steps for arbitrary Boolean functions and to simulate the obtained implementation. In this paper, we give an overview of the developed tool and discuss its usage in class.**

I. Introduction

At the University of Zagreb, Faculty of Electrical Engineering and Computing, Digital Logic is a first-year course. On this course students are taught various concepts including basics of Boolean algebra, combinatorial circuits, sequential circuits, D/A and A/D conversions, implementations of memory modules as well as transistor implementations of basic logical gates and the implementations of standard programmable circuits. We cover Programmable Array Logic - PAL, Programmable Logic Array - PLA and Field Programmable Gate Array - FPGA. PLAs and PALs are rather simple, intuitive and easy to understand, so students can see the meaning of each bit used in their programming. This is not true for FPGAs, whose conceptual structure is known, but the direct correspondence between programming stream bits (generated by commercial development tools) and each programmable element (either pass transistors, LUT SRAM cells etc.) is unknown.

The working of an FPGA can be demonstrated to students by describing some digital circuit using a hardware description language (e.g. VHDL), then applying some magic (use multi-gigabyte commercial programming tools) which will properly configure the black box (i.e. the FPGA) and then observing that the black box behaves as the described digital circuit should. However, this approach is not adequate if deeper understanding of the technology is desired. In order to demystify this process, students are often taught the conceptual structure of FPGAs (for example, LUT-based implementations) and then are left to imagine how everything else works.

Our research goal is to propose a tool that allows students to explore the complete process which starts with a Boolean function to be implemented, goes through each implementation step and ends with a fully programmed FPGA. We believe that introducing such a tool in class lectures could help students to attain a much better understanding of the topic. Additionally, it would be beneficial if for the resulting FPGA students would be able to see the complete programming (every switch, every memory cell, etc.) and if it would be possible to interactively change the values on input pins and observe how signal changes in each programmed element eventually result in a new output value. Therefore, adhering to constructive research methodology, we have developed a graphical portable Java-based tool which emulates a variety of simplified LUT-based FPGA architectures and which allows students to define an FPGA architecture and a Boolean function to be implemented, performs all implementation steps and finally offers a visual inspection of the programmed FPGA as well as an interactive simulation of its work.

II. Related work

The behavior of an FPGA can be defined either through a schematic design or using a hardware description language such as VHDL [1] or Verilog [2]. The process of mapping the desired behavior to FPGA architecture is then carried out, typically using proprietary software specific to each FPGA manufacturer [3]. This process consists of Boolean function decomposition, which determines how the Boolean function will be fitted to CLBs, placement, which determines which physical CLBs will be used, and routing, which determines how the physical CLBs will be connected [4]. We now briefly outline some important considerations in each stage and the solutions proposed in literature.

1) Boolean function decomposition: The decomposition of the user-specified Boolean functions needs to be carried out in a manner that ensures that the functions can be realized using the available type of CLBs. For instance, it could be the case that a Boolean function is defined with n inputs, and the available CLBs have k inputs where $k < n$. Furthermore, redundancy should be minimized, so that CLBs should be shared between different Boolean functions whenever possible. As there are no computationally efficient algorithms that guarantee optimal decomposition, a number of solutions have been proposed in related work, optimizing according to various criteria. For instance, Łuba and Selvaraj [5] propose representing the function by a set of r-partitions over the set of minterms and interleaving serial and parallel decomposition strategies.

Nowicka et al. [6] propose an approach that produces a number of solutions that meet different optimization criteria.

2) Placement and routing: When mapping to physical CLBs, one major consideration is how to assign physical CLBs so the latency between CLBs is minimized. Moreover, determining which routing channels to use to connect two CLBs is a complex optimization problem with many trade-offs. For instance, one might connect two CLBs with the shortest possible route to ensure minimum latency, which might have the consequence that some other CLBs cannot be connected using that channel resulting in unwanted latency in some other components. Enzler et al. [7] propose a measure of quantification of an FPGA design that enables early trade-off considerations. Placement and routing can be optimized using heuristic methods, such as e.g. genetic algorithms [8], combined genetic algorithms and simulated annealing [9], or other optimization techniques such as min-cut or quadratic optimization [10].

3) Hardware constraints: When the logical design is transferred onto an actual physical chip, physical constraints of the hardware environment should be respected. Physical FPGAs are typically connected to other hardware in a fixed way (for instance, pin 1 might be connected to a physical LED). As an additional constraint, placement and routing should take into account this hardware environment, so that logical signals are mapped to appropriate pins. Being able to constrain placement and routing is also important when optimizing power consumption [11].

In our experience teaching FPGA design in early stages of university education, students find it difficult to use proprietary vendor-specific software for FPGA programming. We attribute this to the fact that this type of software is mainly intended for hardware professionals, exposing many complex options that students do not learn in their curriculum. Additionally, many of the inner workings of the synthesized FPGAs are hidden from view, being protected as a trade secret. The tool proposed in this paper help us to alleviate the described problems.

Existing e-learning solutions for FPGA design [12], [13] mainly emphasize having the students interact with a physical board. El Medany [12] proposes an FPGA hardware remote laboratory that can be accessed by students over the internet. Garcia-Zubia et al. [13] propose a system intended for remotely teaching digital electronics that combines a step-by-step digital circuit designer, a real FPGA board and a virtual simulation of a complex watertank that is controlled by the physical board. The virtual simulation aims to expand the range of possible exercises that can be done using a remote FPGA board, given that the outputs of such board are limited. Further details on the proposed system can be found in [14].

We believe in the importance of teaching FPGAs to students, as FPGA technology is widely used and constantly developed. For instance, recently there has been an interest in combining FPGAs with embedded microprocessors, as in e.g. Xilinx Zynq-7000 programmable SoC [15], offering the possibility to easily integrate and intermix the software programmability of a processor and the hardware programmability of an FPGA.

III. Simplified FPGA architecture

In order to explain the functioning and the programming of an FPGA we have developed a simplified FPGA architecture combined with a Java-based visualization and simulation program. The main parts of this FPGA architecture are I/O blocks, configurable logic blocks (CLBs), wiring segments, switch boxes and programmable switches for signal input and output from CLBs.

I/O blocks can be used to bring signals from the FPGA's environment into the FPGA (when configured as inputs) or to bring Boolean functions calculated by the FPGA to the FPGA's environment (when configured as outputs). In our simplified architecture, once programmed, a pin can not change directionality.

Configurable logic blocks in our architecture mimic typical CLBs in todays mainstream FPGAs. Each CLB (see Figure 1) is composed from a look-up table (LUT), a D flip-flop and a multiplexer which selects what is routed to CLBs output: the output of the LUT or the output of the flip-flop output.

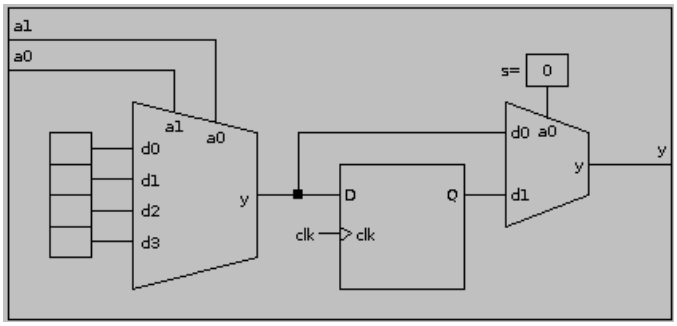


Fig. 1. The structure of a CLB

A LUT is a structure which uses 2^k user-programmable SRAM cells and one multiplexer $2^k/1$ with k address inputs which are also CLB inputs. Each SRAM cell is connected to one data input of the multiplexer. This structure allows us to directly realize any Boolean function of up to k variables simply by writing its truth table into SRAM cells. To illustrate how this works, Figure 2 shows an implementation of the Boolean function $f(A, B) = A \cdot B$ using a two-variable LUT.

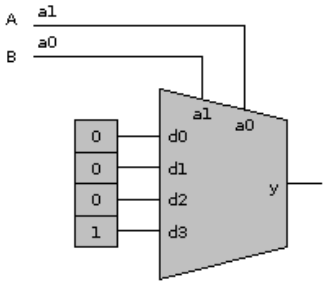


Fig. 2. A LUT programmed to calculate $f(A, B) = A \cdot B$

In our simplified architecture, each CLB receives inputs from the wiring segment located left from the CLB and

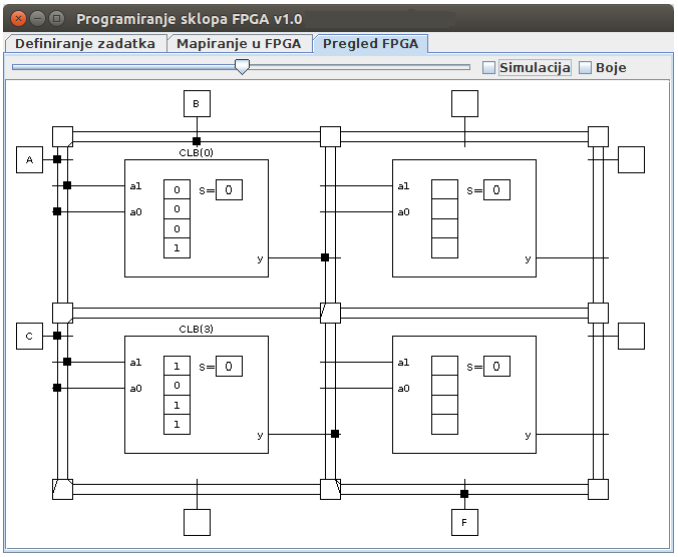


Fig. 3. FPGA programmed to generate function $f = A \cdot B + \bar{C}$

generates output which can be placed onto the wire from the wiring segment located right from the CLB. Each CLB input has its own programmable multiplexer which can be configured to route a signal from a single segment wire. Outputs are routed into a segment wire using pass transistors.

The function of switch boxes is to route signals from input pins or CLB outputs to output pins and CLB inputs. There are various implementations of switch boxes, for example the Disjoint Switch Box [16], [17], the Universal Switch Box [18] and the Wilton Switch Box [19]. The switch boxes we use in our simplified architecture allow the creation of connections between wires from each segment and wires from remaining segments.

An illustrative example is shown in Figure 3, where FPGA is used to implement the Boolean function $f = A \cdot B + \bar{C}$. The FPGA shown has four two-input CLBs in a 2×2 matrix, eight I/O pins, six horizontal and six vertical wire segments, each comprised of two wires. Wire segments are connected using nine switch-boxes. Only two CLBs are used for the implementation of the given Boolean function (the left side of FPGA). There are three input pins (denoted A, B and C) and one output pin (denoted F) on which the function is produced. The LUT programming in each CLB is also shown, as well as the routing configured in switch boxes. Connections established in each switch box are visualized in Figure 3 by joining the connected signals with a line within the switch box. For example, the upper-left switch box only establishes a connection between the second horizontal wire and the second vertical wire.

IV. From Boolean functions to programmed FPGAs: the workflow

For students to acquire a deeper understanding of FPGA technology (and to grasp why the commercial tools are so complex), it is important to demonstrate the complete work-

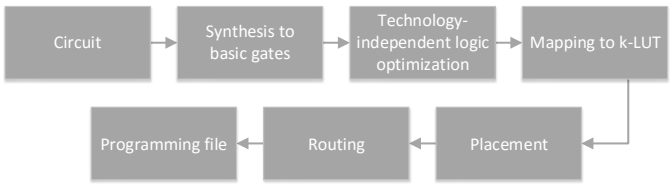


Fig. 4. The FPGA programming workflow

flow: how does one come from an abstract Boolean function to a programmed FPGA chip. A simple diagram shown in Figure 4 illustrates this workflow.

One starts with a circuit description which is usually given in some high level hardware description language such VHDL [1] or Verilog [2]. This description is in a process called *synthesis* converted into networks of basic logic gates. Using rules of Boolean algebra, this representation can then be converted into more adequate one (e.g. using minimization). Then, the resulting Boolean functions which can be functions of arbitrarily many variables should be decomposed into a network of k-LUTs, where each LUT implements a Boolean function of at most k-variables (k is typically a fixed parameter of the available architecture). Then, for each LUT generated in the decomposition procedure a physical LUT (i.e. CLB) must be assigned in the FPGA chip and configured appropriately; this process is called *placement*. After placement is done, I/O pins must be allocated as well and then the switch boxes must be configured so that each LUT gets the required input values. This process is called *routing*.

Each of the aforementioned steps is non-trivial. For example, how can we optimally create a function decomposition? If we have more than one function to implement, how can we decide if there are common subfunctions which can be implemented once and then shared among function implementations? As mentioned previously, today these problems are solved using heuristic algorithms.

When considering routing, one could start with the idea to find shortest path between connected CLBs. However, the routing problem we face here is the problem of simultaneously finding routes for many pairs of CLBs which, considering limited available wires in wire segments, can be hard or even impossible for some bad LUT placements.

Additionally, the routing procedure can have an additional constraint if the FPGA chip is in advance placed into a bigger circuit where the external wiring is already decided so the procedure does not have the freedom to select I/O pins for each variable and function, but has to obey predefined placement (e.g. function f must be routed to pin 5).

We will illustrate the difficulties one meets with the decomposition procedure on a simple example of finding the optimal decomposition for the Boolean function

$$f(A, B, C) = A \cdot \bar{B} \cdot C + B \cdot \bar{C}$$

using 2-LUTs.

One approach could be to utilize the Shannon decomposition theorem [20] (sometimes also referred to as an expansion)

which states that for each Boolean function F the following holds: $F = \bar{x} \cdot F_{\bar{x}} + x \cdot F_x$ where x is any variable, $F_{\bar{x}}$ is the function obtained from F by setting $x = 0$ and F_x is the function obtained from F by setting $x = 1$. We proceed recursively to decompose $F_{\bar{x}}$ and F_x until we end up with subfunctions of only two variables. This decomposition is shown in Figure 5.

$$f(A, B, C) = \underbrace{\overline{A}\ \underbrace{\underbrace{(\overline{B}\,C)}_{f_1(B,C)}}_{f_3(A,f_1)} + A\ \underbrace{\underbrace{(B\,\overline{C} + \overline{B}\,C)}_{f_2(B,C)}}_{f_4(A,f_2)}}_{f_5(f_3,f_4)}$$

Fig. 5. Decomposition based on the Shannon decomposition theorem

Using this approach we have used five 2-LUTs. The 2-LUTs are used to implement following subfunctions.

$$f_1(B,C) = \bar{B} \cdot C$$

$$f_2(B,C) = B \cdot \bar{C} + \bar{B} \cdot C$$

$$f_3(A,f_1) = \bar{A} \cdot f_1$$

$$f_4(A,f_2) = A \cdot f_2$$

$$f = f_5(f_3,f_4) = f_3 + f_4$$

We see that this decomposition could not be implemented into the FPGA from Figure 3 which only has four 2-LUTs. We can try a different approach hoping to end up with a smaller number of 2-LUTs. Since the 2-LUTs cannot implement a Boolean function of three variables, a naive approach would be to group variables and subfunctions into subfunctions of only two variables. This approach is shown in Figure 6.

$$f(A, B, C) = \underbrace{A\ \underbrace{\underbrace{B\,\overline{C}}_{f_1(B,C)}}_{f_2(A,f_1)} + \underbrace{\overline{B}\,C}_{f_3(B,C)}}_{f_4(f_2,f_3)}$$

Fig. 6. Naive decomposition

Using this approach we have used only four 2-LUTs. However, the described approach lacks a clear strategy on the grouping order.

Finally, it is possible to rewrite the given function in such a way that only three 2-LUTs are required for its implementation. This decomposition is shown in Figure 7.

For commercial tools non-introductory course would now continue to describe many heuristic approaches for solving this problem. However, on an introductory course, examples such as the described one are quite enough to illustrate the problem, to create interest in students into the possible ways for solving it and to demonstrate that such problems can be solved more-or-less successfully, and various solvers are an integral part of commercial packages for programming FPGAs.

$$f(A, B, C) = A\,B\,\overline{C} + \overline{B}\,C = \underbrace{\underbrace{(A + C)}_{f_1(A,C)} \cdot \underbrace{(B \oplus C)}_{f_2(B,C)}}_{f_3(f_1,f_2)}$$

Fig. 7. Better decomposition

The final step of the described workflow is preparing a configuration file for FPGA. This configuration file (called *bitstream* and often saved in JEDEC file format) is a stream of bits where each bit represents some configuration detail in the FPGA chip (for example, content of SRAM cells in LUTs, configuration of the output multiplexer in a CLB, configuration for each pass transistor, configuration for each I/O block, configuration for each input multiplexer which passes a value from single segment wire to address input of the LUT's multiplexer, etc). For commercially available FPGA chips the meaning of each bit in this bitstream is not publicly known; that way FPGA chip producers tie the users to buy and work with their own set of tools.

V. The Developed Tool

In order to demonstrate the most of the described workflow steps, we have developed a Java based tool for programming, visualization and simulation of a family of simple FPGA chips.

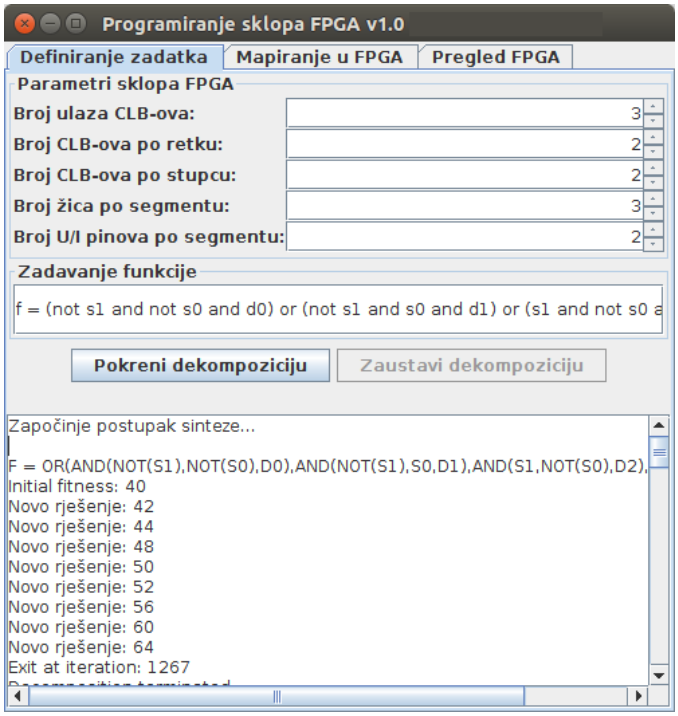


Fig. 8. The developed tool: first step

When the program is started, the user begins by specifing the FPGA architecture (see the top part of Figure 8). The user can define the following parameters.

- Number of CLB inputs (the k in the previous section). The user can decide to use 2-LUTs, 3-LUTs, etc.
- Number of CLBs in each row.
- Number of CLBs in each column.
- Number of wires per each wire segment.
- Number of I/O pins per wire segment.

Using these parameters a wide variety of FPGAs can be simulated. The architecture defined by parameters shown in Figure 8 used four 3-LUTs in 2×2 matrix having 2 I/O pins per wire segment and 3 wires per wire segment.

In the middle of Figure 8 a text box is shown, in which the user can specify a Boolean function to be implemented. At this moment, we only allow one function. The function can be entered using a simple infix notation with standard Boolean operators AND, OR, XOR, NAND, NOR, XNOR and NOT. The function shown in Figure 8 is a 4/1 multiplexer having data inputs d0, d1, d2, d3 and address inputs s1 and s0:

$$f = \bar{s}_1 \cdot \bar{s}_0 \cdot d_0 + \bar{s}_1 \cdot s_0 \cdot d_1 + s_1 \cdot \bar{s}_0 \cdot d_2 + s_1 \cdot s_0 \cdot d_3.$$

The decomposition process is started when the left central button is pressed (Croatian: *Pokreni dekompoziciju*). The progress and output is written in the text area on the bottom. In this example, the decomposition process finished with following:

```
Function is at output: 2
CLB(0): D2 S0 D3
00011101
CLB(1): D0 D1 S0
11100100
CLB(2): CLB(1) S1 CLB(0)
11010001
CLB(3): D0 D2 CLB(2)
00011000
```

For each CLB the signals for LUT inputs are written (starting with the most significant) and then the LUT configuration is written. Let us examine a bit closely the function of CLB(0). The function it realizes is

$$\begin{aligned} f(D_2, S_0, D_3) &= \sum \text{minterm}(3, 4, 5, 7) \\ &= \bar{D}_2 \cdot S_0 \cdot D_3 + D_2 \cdot \bar{S}_0 \cdot \bar{D}_3 + D_2 \cdot \bar{S}_0 \cdot D_3 + D_2 \cdot S_0 \cdot D_3 \\ &= \bar{S}_0(D_2 \cdot \bar{D}_3 + D_2 \cdot D_3) + S_0(\bar{D}_2 \cdot D_3 + D_2 \cdot D_3) \\ &= \bar{S}_0 \cdot D_2 + S_0 \cdot D_3 \end{aligned}$$

which is multiplexer 2/1 which multiplexes D_2 or D_3 based on S_0. The CLB(1) is also a multiplexer 2/1 which multiplexes D_0 or D_1 based on S_0. Finally, the CLB(2) on which the function is realized is also multiplexer 2/1 which multiplexes output of CLB(0) and CLB(1) based on S_1. It is trivial to see that the decomposition procedure has created a multiplexer tree realizing multiplexer 4/1 using a tree of multiplexers 2/1.

Once the decomposition is done, the user can proceed to the second tab which is shown in Figure 9. On this tab one can select the mapping algorithm (currently, we offer only one) and then can start the mapping procedure by pressing the central left button (Croatian: *Pokreni mapiranje*). This will start both the placement procedure and the routing procedure. The results and the final report will be written in the text area at the bottom of the window.

Once this is done, the user can proceed to the last tab which

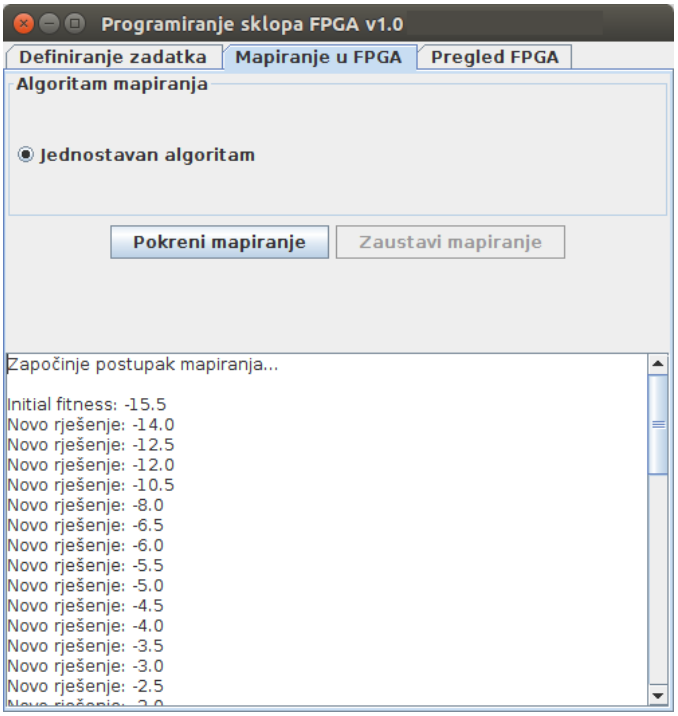


Fig. 9. The developed tool: second step

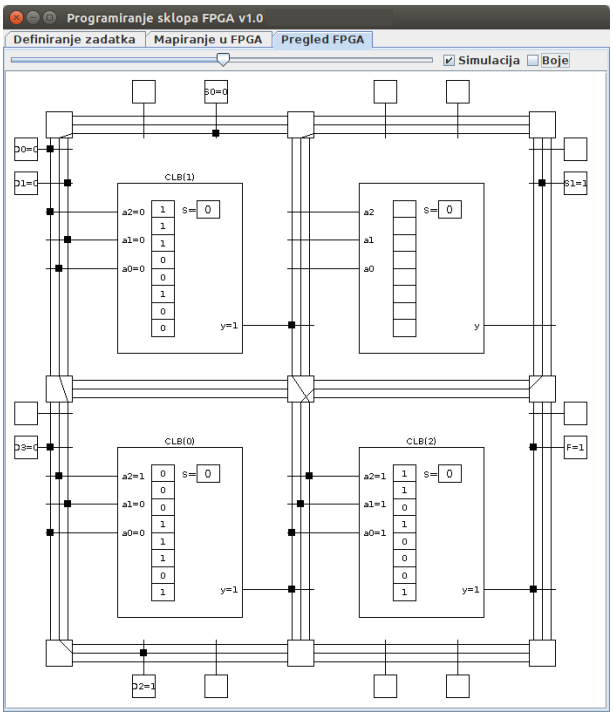


Fig. 10. The developed tool: third step

is shown in Figure 10.

Here we can see the final configured FPGA chip. We can inspect all wiring and switch-boxes and CLB configuration, and by clicking on any CLB while holding the CRTL key, we can get a detailed schematic of the CLB. In the shown case, the placement and routing algorithm placed inputs D0, D1 and D3 along the left side of chip, D1 along the bottom side, S0 along the top side and S1 and output F along the right side. If user enables Simulation mode (checkbox *Simulacija*), values on input pins can be selected by clicking while the outputs will be automatically calculated for each CLB. In the shown example we have $d_3d_2d_1d_0 = 0100$, $s_1s_0 = 10$ so the multiplexer output should be equal to d_2 and so $F = 1$.

By using the synthesised FPGA in interactive mode, students can observe each aspect of its functioning in real time. Using this tool allows us to demystify the fundamentals of FPGA technology and explain what commercial tools do.

It is important to note that the algorithms implemented in this tool for various stages are not commercial-grade algorithms: our intention was to build an educational tool. The algorithms we used are based on evolutionary computations. Specifically, we have implemented a variant of a genetic algorithm for Boolean function decomposition which is similar to Cartesian Genetic Programming [21]. Another variant of a genetic algorithm was used to solve the placement and the routing problem as well as configuration simplification.

VI. Conclusion and future work

In this paper we have described a platform independent Java based tool for programming, visualization and simulation of simplified FPGAs. We have developed this tool in an effort to enable students to better understand what an FPGA is, how it works and how it is programmed.

In academic year 2015/2016 we have introduced the tool in class and were able to obtain a much better interaction and discussion with students while presenting the FPGA topic. Students tended to ask much more focused and related questions which was not the case in previous years. This is a strong indication that the students acquired a better understanding of the topic, which was our research goal.

Currently, the developed tool allows the user to customize the FPGA architecture but to enter only a single Boolean function to be synthesized. For educational purposes this is quite enough since it allows us to illustrate most of the steps which commercial tools do. However, we are planning an extension of this work on several fronts. We plan to allow the synthesis of multiple Boolean functions as well as the synthesis of sequential circuits. Also, to make the user experience closer to the one offered by the commercial tools, we plan to implement support for circuit synthesis described by the user using a simple subset of the VHDL language (which we also use on laboratory exercises). Finally, another extension of current work will be to allow the user to manually program the FPGA circuit directly by mouse (clicking on a connection, dragging to configure switchboxes, clicking to set SRAM content in LUTs etc.).

References

[1] P. J. Ashenden, *The Designer's Guide to VHDL*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2008.

[2] D. E. Thomas and P. R. Moorby, *The Verilog Hardware Description Language*, 3rd ed. Burlington, MA, USA: Springer, 1996.

[3] A. Sangiovanni-Vincentelli, A. El Gamal, and J. Rose, "Synthesis method for field programmable gate arrays," *Proceedings of the IEEE*, vol. 81, no. 7, pp. 1057–1083, Jul 1993.

[4] K. Compton and S. Hauck, "Reconfigurable computing: A survey of systems and software," *ACM Comput. Surv.*, vol. 34, no. 2, pp. 171–210, Jun. 2002.

[5] T. Łuba and H. Selvaraj, "A general approach to boolean function decomposition and its application in FPGA based synthesis," *VLSI Design*, vol. 3, no. 3-4, pp. 289–300, 1995.

[6] M. Nowicka, T. Łuba, and M. Rawski, "FPGA-based decomposition of boolean functions. algorithms and implementation," *Proc. of Sixth International Conference on Advanced Computer Systems*, pp. 502–509, 1999.

[7] R. Enzler, T. Jeger, D. Cottet, and G. Tröster, *Field-Programmable Logic and Applications: The Roadmap to Reconfigurable Computing: 10th International Conference*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, ch. High-Level Area and Performance Estimation of Hardware Building Blocks on FPGAs, pp. 525–534.

[8] M. del Solar, J. Perez, J. Pulido, and M. Rodriguez, "Placement and routing of boolean functions in constrained FPGAs using a distributed genetic algorithm and local search," in *Parallel and Distributed Processing Symposium*, April 2006, p. 7.

[9] M. Yang, A. Almaini, L. Wang, and P. Wang, "FPGA placement using genetic algorithm with simulated annealing," in *ASICON 2005. 6th International Conference On*, vol. 2, Oct 2005, pp. 806–810.

[10] S.-J. Lee and K. Raahemifar, "FPGA placement optimization methodology survey," in *Electrical and Computer Engineering, 2008. CCECE 2008. Canadian Conference on*, May 2008, pp. 001 981–001 986.

[11] A. Gayasen, Y. Tsai, N. Vijaykrishnan, M. Kandemir, M. J. Irwin, and T. Tuan, "Reducing leakage energy in FPGAs using region-constrained placement," in *Proceedings of the 2004 ACM/SIGDA 12th International Symposium on Field Programmable Gate Arrays*, ser. FPGA '04. New York, NY, USA: ACM, 2004, pp. 51–58.

[12] W. El Medany, "FPGA remote laboratory for hardware e-learning courses," in *Computational Technologies in Electrical and Electronics Engineering, 2008. SIBIRCON 2008. IEEE Region 8 International Conference on*, July 2008, pp. 106–109.

[13] J. Garcia-Zubia, I. Angulo, L. Rodriguez-Gil, P. Orduna, O. Dziabenko, and M. Guenaga, "Boole-WebLab-FPGA: Creating an integrated digital electronics learning workflow through a hybrid laboratory and an educational electronics design tool," *International Journal of Online Engineering (iJOE)*, vol. 9, 2013.

[14] L. Rodriguez-Gil, P. Orduna, J. Garcia-Zubia, I. Angulo, and D. Lopez-de Ipina, "Graphic technologies for virtual, remote and hybrid laboratories: WebLab-FPGA hybrid lab," in *Remote Engineering and Virtual Instrumentation (REV), 2014 11th International Conference on*, Feb 2014, pp. 163–166.

[15] L. H. Crockett, R. A. Elliot, M. A. Enderwitz, and R. W. Stewart, *The Zynq Book: Embedded Processing with the Arm Cortex-A9 on the Xilinx Zynq-7000 All Programmable SoC*. UK: Strathclyde Academic Media, 2014.

[16] N. H. E. Weste and D. Money Harris, *Principles of CMOS VLSI Design: A Systems Perspective*, 4th ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2011.

[17] G. G. Lemieux and S. D. Brown, "A detailed routing algorithm for allocating wire segments in field-programmable gate arrays," in *In Proc. ACM/SIGDA Physical Design Workshop*, 1993, pp. 215–226.

[18] Y.-W. Chang, D. F. Wong, and C. K. Wong, "Universal switch modules for FPGA design," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 1, no. 1, pp. 80–101, Jan. 1996. [Online]. Available: http://doi.acm.org/10.1145/225871.225886

[19] S. J. E. Wilton, "Architectures and algorithms for field-programmable gate arrays with embedded memory," Ph.D. dissertation, University of Toronto, 1997.

[20] C. E. Shannon, "The synthesis of two terminal switching circuits," *Bell System Technical Journal*, vol. 28, pp. 59–98, 1949.

[21] J. F. Miller and P. Thomson, "Cartesian genetic programming," in *Genetic Programming*. Springer, 2000, pp. 121–132.

Digital risks and experiences of future teachers

Tomaž Bratina
University of Maribor, Faculty of Education, Department of Basic Pedagogical Studies
tomaz.bratina@um.si

Abstract - Suitable level of knowledge about digital risks is a vital part of the teacher's responsibility for safe manipulation with web contents. The teacher often presents the first barrier and safe haven for the children against the potential digital risks and the demand for this knowledge rapidly increases. The appropriate attitude toward digital risks should start developing from the early beginning of individual's web exploration. It is desirable that the future teachers should enhance their knowledge and responsibility concerning digital risks during their study; before entering the classroom. The paper presents the analysis of contemporary knowledge and experiences of future teachers about the digital risks.

I. INTRODUCTION

The different types of risks in careless web browsing are well known. Most of users are unaware about the potential online threats and/or they pay too little attention to the known risks. The reasons may be the lack of knowledge or they have false confidence that avoiding some web contents protects them and that installed protective software is bulletproof. Many persons are convinced that there is nothing wrong in the offerings of certain web pages and they regards such offerings without any critical consideration. Such persons are also prone to publish their critical personal data to receive some false advantages. Users are often unaware that personal data: name and family names, birthdates, location etc. can be abused. The majority of users are completely unaware that even published pictures can reveal sensitive personal data [1]. Very small percentage of users seek permission from other people that are on their photos for the consent to publish them. According to the Sophos Ltd. research over 90 % of Facebook users did not approve posting of their photos [2]. We consider this as compromising people's privacy and even exposing them to unwanted risks. Such behaviour may easily lead to digital violence, blackmailing or other risks. Publishing different photos or other personal data on social media is a contemporary risk, which can grove up to cyberbullying. The cyberbullying is the most common risk for teenagers. It consists of false communication, and the spreading hostile information about other person(s) in order to embarrass them. Another considerable risk is online harassment, which can cause profound psychosocial outcomes like depression, anxiety, isolation, and even lead to the tragic consequence [3].

II. DIGITAL RISKS

There are many types of digital risks beside previously described. **Sexting** means sending and receiving sexually explicit messages, but most of young people consider sexting as sending their naked photos. Such conversations and photos can be huge embarrassment for involved person if published (deliberate or by mistake) on social or other digital media. Many people do not know to remove these posts on social media therefore consider such removal action as impossible. Therefore, the victims are very exposed for long time and this have profound psychological effects of them. **Phishing** is very common risk focused on gaining personal data or even access to the financial data. Most efficient are the phishing attacks on the social networks. The **identity theft** is focused on acquiring crucial personal data for any kind of often illegal activity where stolen identity is used to hide the real perpetrator. The consequences for the victim are serious and irreversible in general [4]. **Data loss** can occur because of the hardware failure, user's careless handling with their devices, or ransomware. The careless handling may be understand as keeping the data on single media without backup. Perceive negligible importance of data backup is still very common among our students despite many accidental data losses [5]. Ransomware are the considerable risk since retrieving of the data is often impossible or it is very expensive especially because often no backup exist. The ransomware are considerable potential threats since attacks are hidden and capable to deceive even a professional. A user can often become a victim without his or her fault. The computer viruses are small programs to gain control of the victim's computer causing damage and/or data loss. With appropriate antivirus software, knowledge and behaviour the threats are easy to avoid.

The most vulnerable population are undoubtedly pupils or younger children who are not aware of possible consequences from any of mentioned digital risks. Therefore, the teacher should be the first barrier and safe haven for them and should have suitable knowledge about digital security and/or risks. According to the Slovenian Educational Network [6] one from the six basic e-competences comprehends the safe usage of the ICT means regarding to the legal and ethical principles of exploitation and publishing of information. Instead of the term e- competences, the term digital competences may be more appropriate. The term digital competences describes

much wider set of skills in using of ICT inclusive the digital security [7]. Nevertheless, we cannot expect that teachers will always be prepared to cope with every new evolved risk or situation. To help teachers in finding adequate support or advice, the certain level of support should be provided at the institutional level. The national project of Safer Internet Centre was established in Slovenia. The main objective of the project is to promote and ensure a better internet for the kids [8]. The activities cover different areas: advising teacher, children and parents; providing latest information to the organizing the seminars; and preparing the workshops.

Reports from government institutions [9], security software companies, and journals [10] shows the same internet threats. Most often risks among various others are:

- Phishing
- Identity theft
- Malware

The mentioned complications raise the necessity to educate the active and future teachers and prepared them to cope with potential digital risks. Therefore, teacher should acquire the knowledge and experiences before he/she enters the classroom. Teachers should also acquire the adequate knowledge about the hardware and software to understand the mechanisms of digital risks.

III. OBJECTIVE

The main research question was how are the students, of different study programs and ages, prepared to cope with the most often digital risks. The focus of the research was set on:

- The understanding of the conception of phishing
- The understanding of the conception of identity thefts
- Distinguishing frauds vs malware
- Security of personal files?
- Publishing pictures on social networks?

IV. DATA COLLECTION

We prepared the on-line survey. The participants in the survey were 153 students – future teachers, age from 18 to 24. Students were from the studying programs of elementary education, preschool education, pedagogy, fine arts, and music. To achieve the balanced numerus in each group, the age of the students is represented by two groups (from 18-20 years, and over 20 years). Groups also correspond with the first and the second degree of study. The students of fine arts and music are represented in single group *Art disciplines* to achieve the proportional sample.

V. RESULTS

A. The understanding of the conception of phishing

The fishing is currently one of most often applied technique to acquire sensitive personal data or financial data. According to researches, the phishing techniques were part in five most common attacks in the year 2014 [11]. Despite campaigns from government and non-profit institutions, people are still victims of phishing attacks. The awareness and recognition of potential threats is becoming higher though. Only the valid cases are included in the analysis.

TABLE I. THE UNDERSTANDING OF THE CONCEPTION OF PHISHING BY STUDY PROGRAM

Study program / What is phishing?	Elem. education		Preschool education		Pedagogy		Art disciplines	
	f	f %	f	f %	f	f %	f	f %
Searching for information on the web	2	3.8	2	9.1	5	14.3	1	2.9
Introducing with false identity	11	20.8	4	18.2	7	20.0	6	17.6
Acquiring usernames or passwords by using fake webpage	36	67.9	13	59.1	22	62.9	24	70.6
Accumulating friends in social networks	4	7.5	3	13.6	1	2.9	3	8.8

Between about 60 % and 70 % of students of all studying programs understand the conception of phishing. Student from the preschool education are more exposed in this threat because little less than 60 % understand the concept of phishing. On the first sight these results may appear satisfying but about 20 % of students still not understand the conception. However, it is necessary to emphasize, that missing percentage (about 20 %) chose the option "Introducing with false identity" which essentially is not totally wrong. Phishing attacks are often presented on the examples showing false e-mail messages or not existing persons asking for certain data. Therefore this may be the reason that the students choosing the opinion "Introducing with false identity". If this was their thinking we may consider this opinion as correct. This fact brings the positive outcomes into the range of about 80 % to 89 % of opinions representing the good understanding of the conception of phishing. The results are showing no statistically significant difference (P=0.648) among students of different study programs in understanding of the conception of phishing. We may evaluate this result, as very positive result since understanding of the conception of phishing is good and independent from the study program.

TABLE II. THE UNDERSTANDING OF THE CONCEPTION OF PHISHING BY AGE

Age / What is phishing?	18 to20 year		Over 20 year	
	f	f %	f	f %
Searching for information on the web	6	7.4	4	6.3
Introducing with false identity	16	19.8	12	19.0
Acquiring usernames or passwords by using fake webpage	52	64.2	43	68.3
Accumulating friends in social networks	7	8.6	4	6.3

The results in table II show no statistically significant difference (P=0.940) among students of different age in the understanding of the concept of phishing. The understanding is not dependent from the age of the students. Again, the same discrepancy between two similar opinions is present as described above but with the same outcome. Around 84 % of correct opinions of both groups can be considered as good understanding of the conception of phishing.

B. The understanding of the conception of identity theft

The main purpose of the identity theft is to acquire personal data for using them for different activities, commonly illegal, or trading the data to the third party. It is important for everyone to understand the conception of the risk and to take the adequate steps to avoid being victim.

TABLE III. THE UNDERSTANDING OF THE CONCEPTION OF IDENTITY THEFT BY STUDY PROGRAM

Study program / What is the identity theft?	Elementary education		Preschool education		Pedagogy		Art disciplines	
	f	f %	f	f %	f	f %	f	f %
Stealing of the personal documents	4	7.5	4	17.4	3	8.6	3	8.8
Personal data misusing	0	0.0	0	0.0	1	2.9	0	0.0
Message posting using nick name	46	86.8	18	78.3	31	88.6	30	88.2
Attending the exams in someone's name	0	0.0	0	0.0	0	0.0	1	2.9

The majority (up to 88.6 %) of students of all study programs understand the conception of identity theft. The students of preschool education show slight negative deviation of about 10 % against other results. It is not concerning result but pretty high percentage of 17.4 % shows certain lack in distinguishing manipulation with paper data against the digital personal data. In general, the students are well aware of the existence of the risk but we do not have reliable information about their protective measures against the risk. There is no statistically significant difference between groups (P=0.349). Student's understanding of the conception of identity theft is similar and independent from study program.

TABLE IV. THE UNDERSTANDING OF THE CONCEPTION IDENTITY THEFT BY AGE

Age / What is the identity theft?	18 to20 year		Over 20 year	
	f	f %	f	f %
Stealing of the personal documents	8	9.8	6	9.5
Personal data misusing	1	1.2	0	0.0
Message posting using nick name	70	85.4	55	87.3
Attending the exams in someone's name	0	0.0	1	1.6

According to the results in the table, the understanding of the conception of identity theft is not dependent of the student's age. There is no statistically significant difference (P=0.491) between students of different age. More than 85 % of students understand the risk conception. About less than 10 % of students in every group are not yet very familiar with the difference between traditional or digital personal data and possible consequences by the theft.

C. Distinguishing frauds and malware

Distinguishing the fraud(s) versus the malware may be not simple for the majority of less experienced users. These two types of threats often cause the confusion and the same risk like fraud or scam is categorized according to the author's personal preference. Unlike the malware that is more oriented to the various harmful software causing the damage to the operating system, the user software or the data, the frauds are more personally oriented. The consequences of frauds usually means the loss of money or involuntarily spreading of personal data [12].

Our strategy to discover student's ability to distinguish both risk was to present the students the two questions. In the first question, we asked the students to find the correct definition of malware. In the second (multiple-choice) question, the correct and incorrect descriptions of possible frauds were available to choose. The students have to decide if particular description defines the fraud or not. Some students skipped the questions therefore only valid cases are included in the analysis.

TABLE V. THE UNDERSTANDING THE CONCEPTION OF MALWARE

What is malware?	f	f %
The software for shops advertising	9	6.4
Unaware installed SW causing certain damage	93	66.0
Advertisements	19	13.5
Helpful add-on	20	14.2
Total	141	100.0

About the 66 % of students are familiar with the concept of the malware. The results are not satisfying since about 27.7 % of the students (13.5 % and 14.2 %) consider the malware as harmless in form of helpful add-ons or advertisements. According to the results, we suggest to organize some more aggressive campaigns to inform about the threat from malware from the various institutions including educational institutions. Additional analysis show no statistically significant differences in understanding the conception of malware by age (P=0.236) and study discipline (P=0.184).

TABLE VI. RECOGNIZING POTENTIAL FRAUDS

Descriptions of potential frauds	f	f %
Traffic information*	113	28.3
Free toll vignette without buying any article	24	6.0
Mobile network status report by users SMS request*	60	15.0
Free 30 litres of fuel after sending SMS	17	4.3
Prize winning award of 200 € by calling the commercial phone number	17	4.3
Free snacks for one year providing friends phone numbers	12	3.0
Free wellness club membership for two e-mail addresses sent	18	4.5
Payment card transactions notification by free SMS*	55	13.8
Free train traffic delays notifications*	68	17.0
Money award after sending of the three SMS's	16	4.0
Total	400	100.0

The results of multiple choices question show cumulative frequencies of particular description. The asterisk-marked descriptions are the correct descriptions. The all four correct description were selected in about 73.8 % of cases. The traffic information was selected as correct in most cases (28.3 %) following by train traffic delays notification (17.0 %). Surprisingly the payment card transactions were not selected as often as expected (13.8 % only) despite of the security campaigns performed by the banks and the government institutions. The mobile network status reports by request were selected by only 15 % of students. We assume that the service is not used by the students often enough. According to the additional analysis, the correct selections are not dependent by study discipline or age. The percentages of particular description are similar in every category.

D. Securing of personal files

Losing of personal data can be very inconvenient or disastrous for students or professional teachers. Especially in teacher's profession it is very important to keep different data secured. Even the students can also ran into problems by losing their portfolios or other important digital materials necessary for the study. In our survey, we asked the students about their habits in securing files. Some students skipped the questions therefore only valid cases are included in the analysis.

TABLE VII. SECURING FILES BY CREATING BACKUPS

Are you create backups of your files?	f	f %
On daily basis	2	1.3
Weekly	9	6.0
Twice a month	6	4.0
Monthly	23	15.4
A few times in a year	44	29.5
Not creating backups	65	43.6
Total	149	100.0

The results show that the students are very careless about securing their data. About 43.6 % of students do not create backups of their files. Among students who create backups less then third (29.5 %) of students create backups only a few times in a year. Just 15.4 % of students create backups monthly. Additional analysis shows that the behaviour is not dependent by study program. The statistically significant difference (P=0.023) was discovered between students of different ages. The 52.3 % of students of age 18-20 do not create backups. But older students of age 20 and more are bit more careful about their data and they create backups a few times in a year (42.9 %). The results are not encouraging on either side.

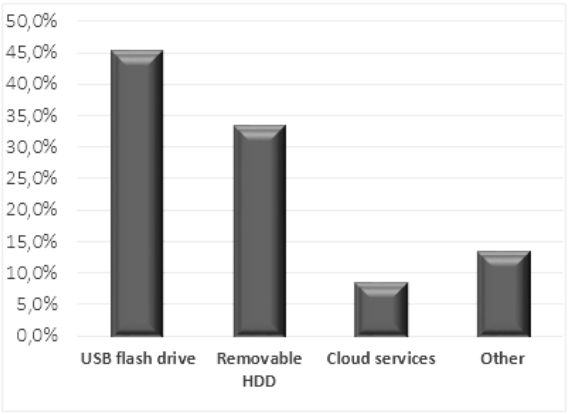


Figure 1. Media for backup of students' files

The most common media used for backups among students are USB flash drives (45.2 %) or external hard disk drives (33.3 %). Only 8.3 % of students use cloud services for the backups. The results correlate in some way with the attitude toward backups in general. Regarding to the known situations and researches [13] the USB flash drives are not appropriate for files backups. They are just for data transporting. This corresponds to our observations in our classrooms. We have witnessed many times how the USB flash drive were mechanically destroyed by pulling it out

from the socket or because of the electrical discharge. The results of additional analysis show that among users of external hard drives the students of art disciplines are prevailing (66.7 %). USB flash media is more often used by the students of elementary education (57.1 %), preschool education (69.2 %) and pedagogy (36.4 %).

E. Publishing pictures on social networks?

Most users are not aware what happens with pictures after they publish them on the social networks. The pictures becomes public domain instantly without real possibility to remove them completely from the network. Once online pictures may be used by almost anyone and for variety of purposes. Since the majority of pictures are created by smartphones, specific information are embed in metadata or EXIF (Exchangeable Image File Format) data. Metadata holds not only technical aspects like camera model, shooter speed, aperture value etc. but also some personal information. Personal information are considered the location data (GPS), date and time, owners name etc. Despite people should be aware about the potential risk they often publish their pictures on social networks without exceptions. In the survey, we asked the students about publishing their own pictures with or without other people on them on the social networks. Some students skipped the questions therefore only valid cases are included in the analysis.

TABLE VIII. PUBLISHING OWN PICTURES ON SOCIAL NETWORKS

How do you publish your own pictures on social networks?	f	f %
Real	136	91.9
Fake	1	0.7
Not publishing	11	7.4
Total	148	100.0

The majority (91.9 %) of students publish their own pictures real pictures on social networks despite of well-known risks. Such behaviour is common among younger population acknowledged by the results of additional analysis. The students of age 18-20 year are more careless about publishing and 96.5 % of them publish their own real pictures on social networks. Students of age 20 and more publish their own real pictures in 85.7 % of cases. About 12.7 % of students do not publish their own pictures. The difference between ages is statistically significant (P=0.044).

TABLE IX. PUBLISHING OWN PICTURES WITH OTHER PERSONS ACCOMPANIED ON SOCIAL NETWORKS

How do you publish your own pictures with other persons accompanied on social networks?	f	f %
Real	185	91.2
Fake	1	0.7
Not publishing	12	8.1
Total	148	100.0

About 91.2 % of students publish their own real pictures with other persons accompanied on social networks and very likely without permission of the other accompanying persons. According to the results of the additional analysis, the younger students are more careless than the older students. The students of age 18-20 year publish their own pictures with other persons accompanied as real in 96.5 % of cases. About 84.1 % of students of age 20 and more publish the real pictures and about 14.3 % of the students do not publish their own pictures with other persons accompanied on social media. The difference between ages is statistically significant (P=0.023). Neither younger students nor the older students act responsible regarding the threat of exposing their pictures on social media.

VI. CONCLUSION

The future teachers understand the concept of contemporary most common digital risk of phishing. The understanding of phishing do not dependent on study program or student's age. Despite of relatively good results such education is still required to increase the percentage of well aware students. In general, students of all observed study programs and both age groups (18-20 and over 20 years) are familiar with the concept of the identity theft. However, we still have to deal with students that do not differ between real personal documents and the digital identity. Understanding of the risk of malware is low and regretfully regarded as harmless. The students in general recognize the possible frauds. As the fraud techniques are constantly developing, the perfect protection is impossible to achieve. Therefore, the adequate vigilance and caution is required. The students' care for their data is not satisfying. Creating backups is more sporadic than regular and without adequate protection. Backups are mostly made only a few times in a year, most often on USB flash drives, which were proven to be very unreliable and with security loopholes. Publishing own real pictures with or without other persons is irresponsible. The research results show that more efforts should be made to raise the awareness of responsibility toward personal data revelation and adequate securing of digital data of any form. Since the results of the research represents the future teachers, no effort should be spared for the raising their digital security

competences. We will continue in conducting the same surveys on forthcoming generations in purpose of comparing the results. A new kind of threats will be included into the survey especially all kind of ransomware that becomes ever more present digital risk.

REFERENCES

[1] L. Koman Perenič, „Protection of personal data and media (Slo.: Varstvo osebnih podatkov in mediji),“ 26 5 2009. [Online]. Available: https://www.ip-rs.si/fileadmin/user_upload/Pdf/smernice/OP_in_mediji.pdf. [Accessed 28 1 2016].

[2] L. Vaas, „Over 90 % of Facebook users hate having photos of them posted without approval,“ 16 2 2012. [Online]. Available: https://nakedsecurity.sophos.com/2012/02/16/facebook-users-hate-photos/. [Accessed 10 2 2016].

[3] G. Schurgin O'Keeffe in K. Clarke-Pearson, „Pediatrics,“ *The Impact of Social Media on Children, Adolescents, and Families,* Izv. 4, pp. 800-804, 2011.

[4] T. Bratina, „Journal Of Elementary Education,“ *Future Teachers' Knowledge on Internet Safety,* Izv. 4, pp. 57-71, 2012.

[5] T. Bratina in M. Krašna, „Students' attitude toward digital security,“ v *5th International Technology, Education and Development Conference (INTED)*, Valencia, 2011.

[6] „SIO (Slovenian Educational Network),“ 5 11 2014. [Online]. Available: https://izobrazevanje.sio.si/navodila-in-postopek-prijave-za-katalog-zrss/pot-do-e-kompetentnosti/.

[7] M. Krašna in T. Bratina, „Digital competences in education: digital security,“ v *International Technology, Education and Development Conference (INTED)*, Valencia, 2011.

[8] „Safer Internet Centre,“ 2015. [Online]. Available: http://www.safe.si/en/center/safer-internet-centre. [Accessed 15 1 2016].

[9] „Common threats to be aware of,“ 1 8 2016. [Online]. Available: http://www.getcybersafe.gc.ca/cnt/rsks/cmmn-thrts-en.aspx. [Accessed 4 2 2016].

[10] „8 Types of Cyber Attacks Your Business Needs to Avoid,“ 14 9 2015. [Online]. Available: http://quickbooks.intuit.com/r/technology-and-security/8-types-of-cyber-attacks-your-business-needs-to-avoid/. [Accessed 12 2 2016].

[11] D. Bisson, „The 5 Most Common Attack Patterns of 2014,“ 24 6 2015. [Online]. Available: http://www.tripwire.com/state-of-security/security-data-protection/cyber-security/the-5-most-common-attack-patterns-of-2014/. [Accessed 21 1 2016].

[12] B. Christensen in D. Christensen, Protecting Yourself from Indentity Theft Internet Scams & Phone Scams, Bundaberg QLD 4670: Brett & Deborah Christensen, 2015.

[13] „Evaluation of Storage Options,“ 3 2 2016. [Online]. Available: http://researchguides.ebling.library.wisc.edu/research-data-management. [Accessed 7 2 2016].

A Study of Factors Influencing Higher Education Teachers' Intention to Use E-learning in Hybrid Environments

S. Babi¯*, M. Ki¥in-Šain ** and G. Bubaš***
* Polytechnic of Rijeka, Business Department, Rijeka, Croatia
sbabic@veleri.hr
** Society "Društvo kiberneti¥ara", Rijeka, Croatia
marina.cicinsain@gmail.com
*** University of Zagreb, Faculty of Organization and Informatics
gbubas@foi.hr

Abstract - The development of e-learning in higher education institutions depends on teachers' adoption and their levels of acceptance of e-learning technologies in the teaching process. Understanding the factors that influence teachers' intention to use e-learning technologies may help explain the barriers to their adoption of e-learning in hybrid environments. This paper investigates the relationship between the factors related to (1) online teachers' competence and (2) educational environment with the intention to apply e-learning technologies in hybrid environments by teachers who are non-users of e-learning. The results of our study showed that attitudes and educational values, computer anxiety, self-efficacy, course characteristics and social influence are related to perceived higher education teachers' intention to use e-learning technologies in hybrid environments. We also compared users and non-users of e-learning in terms of predictors of perceived intention to use e-learning. Our results may instigate further research in the field of adoption of e-learning as well as the development of e-learning strategy in higher education institutions.

I. INTRODUCTION

Virtual learning environments (VLEs), such as Moodle, have become an integral part of the complex infrastructure of higher education systems throughout the world [1]. However, both educational practice and scientific research indicate that teachers' insufficient use of VLEs in higher education teaching still represents a problem [2], [3]. Therefore, there is a need to determine factors which represent barriers to teachers' adoption of VLEs. This paper is especially focused on the problem of the insufficient acceptance of hybrid environments in which teachers combine educational activities characteristic of both traditional and virtual learning environments.

Success in implementing VLEs in the education process depends on user acceptance as well as on contextual factors such as [4]: organization, pedagogy and learning, and technology. For instance, Bingimlas [5] divides barriers to integration of ICT into education into two groups: intrinsic and extrinsic, the former corresponding to *teacher-level barriers* (lack of teachers' confidence, lack of teachers' competence, resistance to change, negative attitudes) and the latter to *school-level barriers* (lack of time, lack of effective training, lack of accessibility, lack of technical support).

The key factors of successful implementation of e-learning in both developing and developed countries are financial support from governments, students' motivation and well trained tutors [6]. Furthermore, according to Naresh and Reddy [6], user perception and readiness also play an important role in e-learning effectiveness. The barriers faced to a larger extent by developing countries than developed ones are as follows: lack of infrastructure, trained instructors, lack of financial support, government policies and less student readiness. Developed countries, on the other hand, have a strong infrastructure and face the following challenges: student engagement, student motivation, and high student drop out ratio [6]. Numerous challenges related to e-learning experienced by higher education teachers can be divided into five categories [7]: learning styles and culture, pedagogical e-learning, technology, technical training, and time management challenges.

Implementation of VLEs can be considered as an innovation in the higher education system. According to Rogers [8], the level of acceptance or refusal is determined by user perception of the following attributes of innovation: (a) the innovation, (b) communication channels, e.g. human relations, facilitated by the media and similar means, (c) time, and (d) social system, e.g. norms, the level of network connectedness and others. Rogers [8] divided adopters of a social system, according to the level of innovation acceptance, into five groups: innovators, early adopters, early majority, late majority, and laggards. Earlier research has shown that the aforementioned adopter categories demonstrate differences in communicative behaviour, socioeconomic status and personal variables, as well as the following characteristics: experience, want and need, innovativeness, and social norms. The systematic

difference between early majority and late majority adopters is especially emphasised. For instance, adopters categorized as early majority have a more positive attitude towards change, better self-efficacy, and greater aspiration to obtain formal education.

With regard to innovation diffusion, Penjor and Zander [3] argue that it is necessary to determine factors which contribute to VLE adoption in particular institutions. Namely, these authors propounds that the majority of teachers at certain institutions are early adopters, while the majority of those at other institutions are late majority and laggard adopters. For instance, significant predictors of adopting VLEs are For instance, while *Compatibility* is a significant predictor of adopting VLEs for early majority and laggards, *Complexity* is a significant predictor for innovators, early adopters, early majority, and late majority [3].

The main research aim of this paper is to determine the factors which influence the intention to adopt e-learning technology in hybrid learning environments, focusing on higher education teachers/instructors who have previously not used a learning management system (LMS) in the teaching process, based on the *Conceptual Model of Competencies of Teachers in Hybrid Learning Environment of HEI* [9]. Therefore, this study analyses the correlations between factors of teachers' online competencies and the learning environment on the one hand, and the intention to apply e-learning technology in hybrid form on behalf of educators who are not users of e-learning (*non-users of e-learning*), on the other.

The following section contains a short overview of relevant literature. First, the factor of *behavioural intention to use of e-learning* is described, followed by factors influencing higher education teachers' intention to use e-learning technology proposed in the relevant literature. A description of a case study is provided in section three, along with the results of our preliminary empirical research. A short conclusion is given in the final section.

II. BACKGROUND AND RELATED WORK

A. Behavioural intention to use e-learning

On the basis of analysing a number of theories and models of adopting technologies and innovation, we established that the Theory of Reasoned Action (TRA) [10] and its further elaborations, for instance the Technology Acceptance Model (TAM) [11] and Unified Theory of Acceptance and Use of Technology (UTAUT) [12] predict *Individual Behaviour (IB)* from *Behavioural Intention (BI).*

The construct of *Intention to use* in the UTAUT model is measured by a scale adapted according to Davis et al. [13] and represents the level of an individual's readiness to exhibit specific behaviour (Fishbein and Ajzen [10]). According to the aforementioned models, the intention of future behaviour related to technology can explain the acceptance of technology more substantially in relation to actual use of technology. That is, if a person intends to use a new technology in an environment, they will indeed do so.

Wang and Wang [14] confirm the statistically significant relationship between *Intention to use* a web-based learning system and its *Actual use*. However, Altawallbeh et al. [15] note that a discrepancy between intentions and behaviours is possible since the intention to accept e-learning can be influenced by unforeseen events. For instance, some studies show that there is no direct connection between intention to use learning management systems (such as Blackboard) and their actual use, which the authors relate to a lack of incentive or persistent downtime of the infrastructures, as well as to poor response of technical support staff [16].

B. Factors influencing higher education teachers' intention to use e-learning technology

In general, the Unified Theory of Acceptance and Use of Technology [12] proposes that *Intention to use technology* on users' behalf can predict *Performance expectancy*, *Effort expectancy*, and *Social influence*, while *Facilitating conditions* and *Intention to use* directly influence the *Actual use of technology*.

Oye et al. [17], relying on the UTAUT model [12], demonstrated that social influence directly affects higher education teachers' behavioural intention within the process of accepting e-learning technology. These authors [17] found a positive influence of computer self-efficacy, and a negative influence of computer anxiety on behavioural intention towards ICT use in the teaching process. In another study based on TAM [11] and UTAUT [12], Patil et al. [18] established that social influence did not influence behavioural intention directly, whereas performance expectancy, effort expectancy, and facilitation conditions are significantly correlated with behavioural intention.

Wang and Wang [14] demonstrated that teachers' conviction regarding self-efficacy is not sufficient, and that perceived usefulness and subjective norm, that is, personal perception of opinions about behavioural intention held by other people in their environment, have a greater influence on the intention to use e-learning systems in teaching. The authors state that the so called social pressure, which can be generated by leadership, students, teaching staff and other factors, encourages teachers to use e-learning systems.

Research by Venkatesh and Davis [19] confirms that subjective norm has a direct influence on behavioural intention in cases when using technology in an environment is not voluntary, while an indirect influence was detected in relation to voluntary use of technology.

Dlalisa and van Niekerk [16] found that perceived usefulness and attitude towards use directly influence behavioural intention to use Blackboard LMS. The same study indicates that LMS experience, computer skills and age significantly influence the adoption of a learning management system. Moreover, Altawallbeh et al. [15] showed that attitude toward adopting e-learning significantly influences behavioural intention to adopt e-learning by means of applying basic Theory of Planned Behaviour (TPB) [20] constructs.

Besides attitude and perceived ease of use, other external variables have also been established as significant predictors of e-learning intention, including the need to supply equipment and substructures of e-learning, accessibility of high speed internet, and culture. In the study by Karami Matin et al. [21], these factors can explain 46% of the variation in intention to use e-learning. Furthermore, research results by Altawallbeh et al. [22] suggest that there is no statistically significant relationship between age, gender and Internet experience on the one hand, and teachers' behavioural intention to adopt e-learning on the other. Our analysis of existing research results indicated a lack of research into teachers' acceptance of hybrid learning environments. The influence of situational factors [23] and factors concerning e-teachers' competence has also been determined.

III. PILOT STUDY

The aim of this pilot study, based on the *Conceptual Model of Competencies of Teachers in Hybrid Learning Environment of HEI* developed by Babi¯ and Bubaš [9] was to identify potential predictors of intention to use e-learning in hybrid environments on behalf of higher education teachers who do not use e-learning technology (such as Moodle, Claroline, Mahara and other institutionally used learning management systems).

A. Participans

One hundred and fourteen higher education teachers at selected higher education institutions in Croatia (based on researcher's assessment) participated in the pilot study. The sample included 49 non-users of e-learning and 65 users of e-learning. Basic differences between the two groups of participants are presented in [9].

The target population of this paper are *non-users of e-learning*. The group of non-users consisted of higher education teachers, 28 (57.1%) among whom were female and 21 (42.9%) male. In addition, most non-users of e-learning work in the humanities (49%), followed by teachers in social sciences (30.6%), while the lowest percentage of the non-users worked in natural and technical sciences, each of them comprising of 2% of the sample. The majority of teachers (59.2%) hold the highest academic degree of Doctor of Science or Arts; 22.4% had a Master's degree (either university or vocational) and only 4.1% finished postgraduate specialist studies (Specialists in given fields). Furthermore, the group of non-users consisted of teaching assistants (40.8%), assistant professors (24.5%), and the minority of 2% of full professors. None of the non-user participants were employed as research assistants or professors at polytechnics.

The analysis of results shows that the convenience sample of non-users mostly (22.4%) consisted of higher education teachers with 10 to 15 years of teaching experience, while 20.4% of participants had been teaching for 7 to 9 years, and only 4.1% of them had 25 to 30 years of teaching experience. None of the participants in the non-user group had less than one year of teaching experience.

On average, participants assessed their e-learning knowledge as rather poor (M=3.73; σ=1.10), using the 1-5 scale (1 – excellent, 2 – very good, 3 – good, 4 – poor and 5 – very poor). More than half (55.1%) of non-users of e-learning had not received education in e-learning. Moreover, most (20.4%) of the non-users who received education had been trained only in technical functionalities of the technology and e-learning tools. The minority of non-users, amounting to only 4%, had received such training during their formal education.

Results shown in Table I suggest that a great majority (93.9%) of non-users in this study use e-mail for e-learning or traditional teaching on a weekly basis, while 26.5% of them weekly spend 10 hours or more using e-mail for that purpose. In addition, only 8.2% of non-users use e-learning systems on a weekly basis. Online video records are used by 69.4% of non-users for of e-learning or teaching.

TABLE I. PERCENTAGE OF NON-USERS OF E-LEARNING IN RELATION TO INTENSITY OF USE OF SPECIFIC INTERNET TECHNOLOGIES IN E-LEARNING OR TEACHING ON A WEEKLY BASIS (N=49)

Internet technologies	Percantage of participants in relation to intensity of use (%)						
	Do not use	Use of up to 1 hour	Use for 2-3 hours	Use for 4-5 hours	Use for 6-7 hours	Use for 8-9 hours	Use for 10 and more hours
Electronic mail	6.1	18.4	20.4	8.2	14.3	6.1	26.5
Forum discussions	85.7	10.2	2.0	0	2.0	0	0
Video conferences and webinars	91.8	8.2	0	0	0	0	0
Online videos recordings (YouTube)	30.6	32.7	30.6	6.1	0	0	0
Wikis, blogs and similar tools	63.3	30.6	4.1	0	2.0	0	0
Social networks (Facebook)	63.3	12.2	14.3	4.1	0	2.0	4.1
E-learning systems (Moodle)	91.8	2.0	6.1	0	0	0	0
Multifunctional tools (Google Drive)	63.3	16.3	14.3	2.0	0	0	4.1
Online surveys (Survey Monkey)	75.5	12.2	10.2	0	2.0	0	0
Online tests (Hot Potatoes)	91.8	8.2	0	0	0	0	0

B. Instruments and Procedure

Factors selected for this study are an integral part of the *Conceptual Model of Competencies of Teachers in Hybrid Learning Environment of HEI* [9]. The research was conducted by means of an online questionnaire (http://kwiksurveys.com) and a paper version administered at Croatian higher education institutions in

October and November 2014. Participation was voluntary and anonymous. The questionnaire contained items items developed for the assessment scales defined in the aforementioned conceptual model, for which we used the 1-5 point Likert scale for responses (1 – totally untrue; 2 – mostly untrue; 3 – neither true nor untrue; 4 – mostly true; 5– totally true). The data were analysed by SPSS-19 using appropriate statistical tests including t-test, ANOVA, Pearson correlation and multiple regressions analysis (stepwise). The number of items and Cronbach alpha coefficients for the assessment scales used in the pilot study for non-users of e-learning are shown in Table II. Results show that all of the scales had good internal consistency (above 0.80).

TABLE II. NUMBER OF ITEMS AND CRONBACH ALPHA COEFFICIENTS FOR ASSESSMENT SCALES USED IN PILOT STUDY (NON-USERS OF E-LEARNING, N=49)

Assessment scale	**Number of items**	**Cronbach alpha**
ICT knowledge and skills	4	.91
Pedagogical competence to use ICT in e-learning	13	.95
Ease of use of e-learning technology in teaching	5	.86
Attitude toward the use of technology for e-learning	5	.92
Educational value of application of e-learning	9	.94
Computer anxiety	5	.83
Self-efficacy	4	.88
Characteristics of the university course	3	.81
Characteristics of enrolled students	3	.84
Technical and social support for e-learning	5	.91
ICT infrastructure, strategies and policies	4	.89
Social influence	8	.84
Intention to use e-learning in hybrid environments	3	.91

C. Findings

1) Intention to use e-learning in hybrid environments

The *Intention to use e-learning in hybrid environments* scale was measured by a 1-5 point Likert scale for responses (1 – totally untrue; 2 – mostly untrue; 3 – neither true nor untrue; 4 – mostly true; 5– totally true). The average scores of non-users of e-learning regarding their responses to the items of the assessment scale *Intention to use e-learning in hybrid environments* is 9.29 (σ=3.79), while the average score of users of e-learning is 13.32 (σ=2.14). In our study the average score was greater for the group of users of e-learning. Using the t-test we confirmed a statistically significant difference, on the $p<0.01$ level, between the two groups of participants regarding the mentioned scale. This result demonstrates that non-users of e-learning demonstrate the intention to use e-learning in hybrid environments, but to a smaller extent with regard to users of e-learning.

2) The relationship between intention to use technology for e-learning in hybrid environments and selected demographic characteristics and prior experience with ICT

T-tests for small independent samples and one way analysis of variance (ANOVA) were used to determine the relationship between non-users' intention to apply e-learning technology in hybrid environments and selected demographic characteristics, as well as their prior experience with technology use. Our analysis revealed no statistically significant connection between non-users' intention to apply e-learning technology in hybrid environments and age, as well as the scientific domain, the level of education and academic rank within the higher education system. However, results of relationship tests confirm that among the participants in the convenience sample of non-users there is a statistically significant correlation between intention to use e-learning technology in hybrid learning environments and the overall duration of computer use at home ($F=4.489$, $p=0.017$). On average, greater intention was expressed by non-users who have been using a computer at home for more than 10 years in relation to those users who have been using it for less than 10 years.

3) The relationship between intention to use e-learning in hybrid environments and other factors of the research model – differences between users and non-users of e-learning

Table III shows the values of the Pearson correlation coefficient of selected scales (mean value) in conjunction with the mean value of the scale *Intention to use e-learning in hybrid environments*, based on the data collected from the convenience samples of non-users of e-learning (N=49) and users of e-learning (N=65) in our pilot study.

TABLE III. COMPARISON OF CORRELATIONS BETWEEN SELECTED MEASUREMENT SCALES (MEAN VALUES) AND THE SCALE INTENTION TO USE E-LEARNING IN HYBRID ENVIRONMENTS FOR NON-USERS (N=49) AND USERS (N=65) OF E-LEARNING

Assessment scale	***Non-users***	***Users***
ICT knowledge and skills	.217	.239
Pedagogical competence to use ICT in e-learning	.278	.474**
Ease of use of e-learning technology in teaching	.278	.416**
Attitude toward the use of technology for e-learning	.508**	.481**
Educational value of application of e-learning	.513**	.472**
Computer anxiety	-.305*	-.501**
Self-efficacy	.402**	.375**
Characteristics of the university course	.517**	.333**
Characteristics of enrolled students	.223	.389**
Technical and social support for e-learning	.097	.122
ICT infrastructure, strategies and policies	.168	.211
Social influence	.468**	.262*

Note: * $p < 0.05$; ** $p < 0.01$.

These indicators (see Table III) show that non-users' intention of future application of e-learning technology in hybrid learning environments is most strongly correlated with the following factors: *Attitude toward the use of technology for e-learning* (r=.508, p<0.00), *Educational value of application of e-learning* (r=.513, p<0.00), *Characteristics of the university course* (r=5.17; p<0.00), *Social influence* (r=.468, p<0.00), and *Self-efficacy* (r=.402, p<0.00).

In addition, *Intention to use e-learning in hybrid environments* is correlated with all of the scales, barring the *Technical and social support for e-learning* scale (r=.097, p>0.05) and *ICT infrastructure, strategies and policies* (r=.168, p=0.00).

It is also necessary to emphasise that all of the correlations are positive, except for the mean value of *Computer anxiety* measurement scale which has a low negative correlation with e-learning technology acceptance on behalf of non-users, amounting to r=-.305 (p<0.05), which is consistent with the results of our theoretical research.

The comparison of correlation analysis results shown in Table III demonstrates that non-users' correlations are on average somewhat higher than those of e-learning users.

Greatest differences in favour of non-users were noted on measurement scales (mean values) *Social influence*, *Characteristics of the university course*, *Educational value of application of e-learning*, followed by the scale *Attitude towards the use of technology for e-learning*.

4) Predictors of intention to use of e-learning in hybrid environments for non-users of e-learning

Table III shows that correlation between certain variables with the measurement scale (mean value) *Intention to use e-learning in hybrid environments* was established. The next step included two multiple regression analyses (see Table IV) for the purpose of determining which of the variables (mean values of measurement scales) can explain the intention of future use of e-learning technology in hybrid teaching on behalf of the convenience sample of non-users.

Table IV presents the results of the multiple regression analysis. All of the factors contained in Table III were contained in *Model 1*. Results show that the only significant predictor is the variable (mean value of measurement scale) *Characteristics of the university course* (β=0.517, p=0.00). This predictor can explain 26.7% of the variance of *Intention to use e-learning in hybrid environments* (criterion variable) for non-users of e-learning. The first regression model obtained is statistically significant (F=17.153, p<.001), while the multiple correlation coefficient (R=0.517) suggests moderate correlation.

Model 2 included all of the factors listed in Table III, barring *Characteristics of the university course*. Results show that the variables (mean value of measurement scale) *Educational value of application of e-learning* (β=0.383, p=0.01) and *Social influence* (β=0.304, p=0.03) have the greatest predictive power. These predictors taken together can explain 33.9% of variance of *Intention to use e-learning in hybrid environments* (criterion variable) for non-users of e-learning.

This provides the basis for a conclusion that non-user participants who understand educational values of e-learning use and perceive that others in their environment consider it necessary to introduce e-learning in the teaching process exhibit a slightly higher level of intention of future use in their educational practices in hybrid learning environments. We can also conclude that teachers who understand the compatibility of e-learning and course subject exhibit intention to use e-learning technology in hybrid learning environments to a greater extent.

TABLE IV. RESULTS OF REGRESSION ANALYSIS ON SELECTED MEASUREMENT SCALES (MEAN VALUES) IN RELATION TO NON-USERS' (N=49) INTENTION TO USE E-LEARNING IN HYBRID ENVIRONMENTS

Model/variables	***Beta***	***t***	***Sig.***
1 Characteristics of the university course	.517**	4.14	.00
2 Educational value of application of e-learning	.383*	2.89	.01
Social influence	.304*	2.29	.03

Note: * $p < 0.05$; ** $p < 0.01$.

IV. CONCLUSION

The findings of this study how that factors selected from the *Conceptual Model of Competencies of Teachers in the Hybrid Learning Environment of HEI* [9] can be used for explaining the relationship between teachers' intention to accept e-learning and their competencies for its application at selected institutions of higher education in Croatia. In this context, items OV1 and OV2 need to be removed from the *Educational value of application of e-learning* scale. All of these scales demonstrated good internal consistency (see Table II) measured by the Cronbach alpha coefficient. In addition, the established differences between users and non-users (see Table III) suggest that our results are valid.

On average, non-users of e-learning have expressed their intention to use e-learning technology in hybrid learning environments when engaging in their teaching practice.

It was determined that selected demographic characteristics (gender, age, teaching experience, academic rank) did not have a direct influence on behavioural intention of teachers with regard to accepting e-learning in hybrid learning environments. Results of descriptive analysis show that higher education teachers on average do not possess knowledge pertaining to e-learning, and that they lack LMS experience which is a basic precondition for accepting hybrid learning environments. In addition, non-users largely do not use many of the technologies for other purposes either (see Table I). Prior experience with using different ICT did not have an effect on intention to accept e-learning, except for the overall duration of computer use at home. Those teachers who have used computers for a longer time at home have also shown a higher level of intention to apply

e-learning. These findings are similar to those of prior studies [22].

This study has found a positive correlation between intention to apply e-learning and self-efficacy, and a negative correlation with computer anxiety. Similar results were obtained by [17].

Characteristics of the university course, that is, the compatibility of the course with e-learning, is one of the most significant factors pertaining to intention to use e-learning, which is in accordance with the results reported by Rogers [8]. Attitude and educational values, as well as social influence are correlated directly with the intention to accept e-learning in hybrid learning environments, which is consistent with previous research [15], [16].

The fact that *educational values* and *social influence* can explain 33.9% of variance of *Intention to use e-learning in hybrid environments* is particularly notable. These findings can be related to the beginnings of development of an organizational culture for e-learning at higher education institutions where the study was conducted.

Our study's findings have implications for both researchers and practitioners. Researchers can use them as an initial framework for understanding the acceptance of e-learning by non-users, while practitioners can use them for the purpose of developing strategies related to organizational learning in the context of e-learning at higher education institutions. Our next paper will focus on conducting research involving a larger number of non-users of e-learning at Croatian higher education institutions.

REFERENCES

[1] E. Costello, "Opening up to open source: looking at how Moodle was adopted in higher education", Open Learning: The Journal of Open, Distance and e-Learning, vol. 28 (3). pp. 187-200, 2014.

[2] S. Babi¯, "Factors that influence academic teacher's acceptance of e-learning technology in blended learning environment", In E. Pontes, A. Silva, A. Guelfi and S. Takeo Kofuji (Eds.), E-Learning-Organizational Infrastructure and Tools for Specific Areas, InTech, 2012.

[3] S. Penjor, and P. O. M. Zander, „Predicting Virtual Learning Environment Adoption: A Case Study", The Turkish online journal of Educational Technology, vol. 15(1), pp. 69-82, 2016.

[4] T. C. Dang, and J. Foster, "Implementing e-Learning in a Vietnamese University: A Configurational Approach", iConference 2015 Proceedings, Newport Beach, CA, USA, 2015.

[5] K. A. Bingimlas, "Barriers to the successful integration of ICT in teaching and learning environments: A review of the literature", Eurasia Journal of Mathematics, Science & Technology Education, vol. 5(3), pp. 235-245, 2009.

[6] B. Naresh, and B. S. Reddy, „Challenges and Opportunity of E-Learning in Developed and Developing Countries-A Review", International Journal of Emerging Research in Management &Technology, vol. 4 (6), pp. 259 -262, 2015.

[7] N. Islam, M. Beer, and F. Slack, "E-Learning Challenges Faced by Academics in Higher Education: A Literature Review", Journal of Education and Training Studies, vol. 3(5), pp. 102-112, 2015.

[8] E. M. Rogers, "Diffusion of innovations", New York, NY: Free press, 2003.

[9] S. Babi¯, and G. Bubaš, "Assessment of competencies of online teachers: pilot study and survey evaluation." eLearning 2015-6th International Conference on e-Learning, 2015.

[10] M. Fishbein, and I. Ajzen, "Belief, attitude, intention and behavior: an introduction to theory and research", Addison-Wesley, Reading MA: Addison-Wesley, 1975.

[11] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology", MIS Quarterly, vol. 13(3), pp. 319-340, 1989.

[12] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: toward a unified view", MIS Quarterly, vol. 27(3), pp. 425-47, 2003.

[13] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. Management Science, pp. 982-1003, 1989.

[14] W. T. Wang, and C. C., Wang, "An empirical study of instructor adoption of web-based learning systems", Computers & Education, vol. 53(3), pp. 761-774, 2009.

[15] M. Altawallbeh, F. Soon, W. Thiam, and S. Alshourah, "Mediating Role of Attitude, Subjective Norm and Perceived Behavioural Control in the Relationships between Their Respective Salient Beliefs and Behavioural Intention to Adopt E-Learning among Instructors in Jordanian Universities", Journal of Education and Practice, vol. 6(11), pp. 152-159, 2015.

[16] S. F. Dlalisa, and B. van Niekerk, "Factors affecting the post-implementation adoption and usage of blackboard amongst academics at dut." 9th Annual teaching and learning in higher education conference, 2015.

[17] N. D. Oye, N. Iahad, and Z. A. R. Nor, "The impact of UTAUT model and ICT theoretical framework on university academic staff: Focus on Adamawa State University, Nigeria", International Journal of Computers & Technology, vol. 2(2), 2012a.

[18] S. Patil, P. N. Razdan, and V. Chancellor, "Creating awareness, acceptance and application of e-learning in higher education institutions: using technology acceptance model (TAM), International Journal of Multidisciplinary Consortium, vol. 2(4), 2015.

[19] V. Venkatesh, and F. D. Davis, "A theoretical extension of the technology acceptance model: four longitudinal field studies", Management Science, vol. 46(2), pp. 186-204., 2000.

[20] I. Ajzen, "The theory of planned behavior. Organizational behavior and human decision processes, vol. 50(2), pp. 179-211, 1991.

[21] M. Altawallbeh, W. Thiam, S. Alshourah, and S. F. Fong, "Do the Instructors Differ in Their Behavioral Intention to Adopt E-Learning Based on Age, Gender, and Internet Experience?", Journal of Education and Practice, vol. 6(18), pp. 41-50, 2015a

[22] B. Karami Matin, M. Mirzaei Alavijeh, M., F. Jalilian, F., and S. N. Hosseini, "Use of Adoption Technology Model to Predicting E-Learning Intention Perform among Faculty Members", Future of Medical Education Journal, vol. 5(3), pp. 34-40, 2015.

[23] S. Babi¯, "The influence of situational factors on accepting e-learning technology by university teachers", Information & Communication Technology Electronics & Microelectronics (MIPRO), 36th International Convention on. IEEE, 2013.

Development and Implementation of E-Learning System in Smart Educational Environment

Amr Elsaadany and Karim Abbas
Computer Engineering Department
Pharos University - Alexandria
amr.sadany@pua.edu.eg, karim.soliman@pua.edu.eg

***Abstract*— E-learning is an integral part of smart education. There are many e-learning systems that are widely available to educational institutions. The challenge is to easily integrate the e-learning system into a smart educational environment based on the requirements of the users. The e-learning services rely on a software system that allows access to all the materials for the educational process and makes them electronically available to all the students on the Internet whenever they need and wherever they are. The design and development of e-learning system is a critical part of the educational process as it reflects on the usage of the system. In this work, the design and implementation of e-learning systems is described where different techniques are explored and compared. The proposed e-learning system is designed using off-the-shelf and open-source software engineering model and programming tools and database models. The system is tested to prove the new design concepts and features. The method used in the back-end and front-end design and implementation allows flexible usage and integration of the e-learning systems by the educational institutions in smart cities.**

Keywords— e-learning; user interface; software models; database; programming tools; smart cities, ICT

I. INTRODUCTION

The design of e-learning systems puts a lot of challenges on the system designer. A good system should have a strong yet flexible design. This is because the system should be robust and available all the time for the intended users. As such, we have approached the design of our system using compatible models in software engineering and database design selections.

Education over the world nowadays are depending more and more on the use of technology and the integration of the new technological trends into the educational systems to introduce better learning environments for the students where ever they are around the globe. Educational technology can be defined as "The study and ethical practice of facilitating learning and improving performance by creating, using and managing appropriate technological processes and resources" [1]. However, a smart educational environment is the environment that uses the new trends of ICT to serve the educational processes and overcome the traditional problems with assistance of such technologies.

There are many problems that face traditional academic learning systems. We have categorized these problems based on our experience in the academic environment. These problems are described below in order to provide a background on the design issues that face e-learning systems development. First, there are many time and place problems that must be addressed by e-learning systems. These problems are highlighted below.

- Lectures nowadays at most of the educational institutions require the students to attend the lectures at specific location and time (location & time dependency). Students sometimes cannot attend the lectures as scheduled.
- Conflicts in time tables could lead students to drop a course and be unable to register it in the desired term/ semester. Often, there are conflicts between subjects in exams time tables (which has fixed time duration).

In a traditional learning (non-electronic) system, students must attend the classes physically and of course have to bear the time and cost of travelling to the university campus to attend the lectures. Also, lecturers and teaching staff may waste valuable time in marking assignments and quizzes manually while they can spend the time enhancing the course materials.

Second, there are many issues associated with the reviewing of the lectures materials and questions. These issues are listed below.

- Students may need to review some points from a previous lecture but they cannot find the lecturer at a suitable time.
- Some lectures take place in big lecture halls and are attended by large number of students where questions of all students cannot be answered within the lecture.
- Sometimes the lecturer delays sending the lecture materials (or upload them on university site) to make them available to the students on time.
- The students may not have the means to post a question to all the colleges in a particular course. With the new system, a student forum can address this requirement.
- Also a student cannot access previously asked questions from previous offerings (semesters) as there is no recording of this history.

The third category is related to lab experiments (the e-learning system should have a way of simulating lab work).

- Some lab experiments are time consuming and/or dangerous and need may to be done in controlled environment.

- Some experiment labs cannot fit the whole crowd of students at a specified time.
- Some experiment materials are extremely expensive or rare and need to be used wisely.

The fourth category includes miscellaneous problems related to the educational system as described below.

- Disabled students suffer from lack of educational aiding facilities and it could be difficult for them to attend their lectures on campus during the regular schedule.
- The code of conduct may be broken by few students (like the copying of assignments). This needs to be dealt with in a non-traditional way by the e-learning system.
- There is no support for a feedback mechanism for lecturers to get indications as to which parts were not fully understood by everyone or which parts need additional clarifications.
- Different instant language translations are not available in the traditional educational systems (and this can be done via e-learning system).

Based on the variety of issues given above, we came up with the requirement of an e-learning system. The requirements of the new system came from a bottom up approach where a selected group of senior students participated in the exercise through brain storming sessions. The challenge is to design and develop a system with these requirements and yet flexible enough for future enhancements.

The paper is organized as this: Section 2 describes common e-learning systems and their basic features and gives a comparison of the common e-learning systems. Section 3 provides the scope of system. Section 4 provides the main system features and section 5 details the system implantation. Section 6 summarizes the results and last Section concludes the paper.

II. E-LEARNING SYSTEMS DEFINITIONS AND COMPARISON

E-learning is basically the use of networking technologies to create, deliver, and facilitate learning, anytime and anywhere [2]. It is composed of instruction delivered on a digital device such as a computer or mobile device that is intended to support the learning process. Basically, e-learning systems must have the following characteristics:

- Stores and/or transmits lessons on selected servers on the Internet (or on university own servers).
- Includes content relevant to the learning objectives and uses media elements such as text, pictures, slides and videos to deliver the content.
- Uses instructional methods such as relevant examples, practices, and feedback to promote learning.
- Provides instructor-led (synchronous e-learning) and self-paced individual study (asynchronous e-learning).
- Helps learners to build new knowledge and skills linked to individual learning goals.

The above list helps in determining the scope of the system and the minimum set of features that makes the system capable of supporting our educational needs. The system must offer on-line education and distant learning through network connection [3]. Moreover, the requirements set by the student community must be met by the system in order to provide value services for smart education environments.

It is important to mention that the educational institutions using e-learning systems should set their required quality standards and evaluate them frequently. They should set quality standards for all the e-learning processes, for example, use standard template to present the course content. They should not only set the standard but also continuously improve the quality of materials being presented and quality of sound being recoded from the lectures. More importantly they must enhance the connection and interaction between students and the academic staff [4].

In the following comparison, we discuss and compare different e-learning systems that are available nowadays. The systems are: BlackBoard, TED ED, CodeAcademy, Udemy, Coursera, EdX, and Udacity. It is clear that e-learning has become a way of life for many businesses and it has become an essential channel for both formal and informal learning in school systems and universities [5]. We also include Moodle system in the analysis due to the increasing number of users and the availability of project developers, attach plugins and APIs [6].

Some e-learning systems (such as EdX) have developed MOOC courses (Massive Open Online Course) which are online course aimed at unlimited participation and open access via the web. The data collected from the e-learning system can be analyzed to examine the students' use of resources and their interactions with the curricular and the pedagogical components [7].

We compare the features that are presented by these systems based on the following criteria:

- Whether the system offers the courses free or paid.
- Whether the system imposes Prerequisite courses or Open system which allows taking any course without prerequisites restrictions (P or O system).
- Whether the system has assessment methods that evaluate the student and keep record of their progress.
- Availability of the student forum where the students can post questions or comments and receive feedback.
- The programming languages used by the system and its user interface.
- The type of dependency; time-based or progress-based (T or P dependency).
- The kind of media used in representing the course materials (video, audio, slides) and where these materials are hosted.
- Whether the system provides certifications to the students or not.

Table 1 shows the comparison between the various systems according to the criteria listed above.

TABLE I. SYSTEMS COMPARISON

System	Free/ Paid	Pre or Open	User Evaluation	Student Forum	Open / Closed Source
Moodle	Free	Open	NA	Yes	Open
Black board	Paid	Both	Yes	Yes	Closed
TED ED	Free	Open	Yes – Optionally	Yes	Open
Code Academy	Free	Open	Yes - Rating by points	Yes	Closed
Udemy	Paid	Open	NA	No	Closed
Coursera	Free	Open	NA	Yes	Closed
EdX	Free	Open	NA	Yes	Open
Udacity	Free	Open	NA	No	Closed

System	Prog. Languages	Media Types	Time or Progress	Video hosted	Certificate
Moodle	PHP, HTML, Js	Video, Audio	None	NA	No
Black board	ASP, Others	Video, Audio, Slides	Both	Client Servers	Yes
TED ED	HTML, JavaScript	Video – Animation	None	YouTube	No
Code Academy	HTML, JS, PHP	Flash and Editors	None	No Videos	No
Udemy	HTML, JS, PHP	Video	None	Private Servers	No
Coursera	C Language	Video and Slides	Both	Private Servers	Yes
EdX	Python, JavaScript, Ruby	Video and Slides	Both	YouTube	Yes
Udacity	Python	Video and Slides	Progress	YouTube	Yes

III. SYSTEM SCOPE DEFINITION

In this section we explore the scope of the subsystems that compose our e-learning system. We also define the possible extensions of the system. Then we describe the expected benefits of the e-learning system. Figure 1 shows the e-learning high level system description. The figure depicts a system that users can interact with via different interfaces through the network. On the other side, it interacts with a database for data exchange. The scope of the system includes the following:

- A system administration capability in order to control the system and grant the suitable permission to each user type (student, lecturer, etc.).
- A registration and login system to allow the student to enter the data that is needed to consider him/her as a member in the system.
- A course scheduling capability according to the requirements of the educational program.
- A course administration subsystem that controls the creation of new courses and attach materials to a specific course. It can also update, edit, and delete any existing course.
- A profiling subsystem to track general information about all types of users (name, birthdate, address, etc.).
- A student performance tracking subsystem that tracks the progress of the student in different courses and the grades of the student in each course he/she attended.

The system provides capabilities for assignments and quizzes preparation, correction, scoring, and recording. It also supports question banking capability and student forum to allow the student to post their questions, answers and comments. The scope of the system can expand to support lectures video conferences and provides live, online and interactive lectures. The proposed system has the following expected benefits:

- Saving time and money; reduce wasted time in traffic and the amount of money spent.
- Make course materials available all along the course; avoid awaiting the lecturer to provide.
- Eliminate conflicts between lectures schedules (which were based on physical attendance).
- Provide direct link with the professor or lecturer electronically through the system.
- Provide the disabled students with an easier way of learning.
- Provide broadcasting questions to course members and the lecturer and receive timely feedback.
- Provide instant translation on video courses.

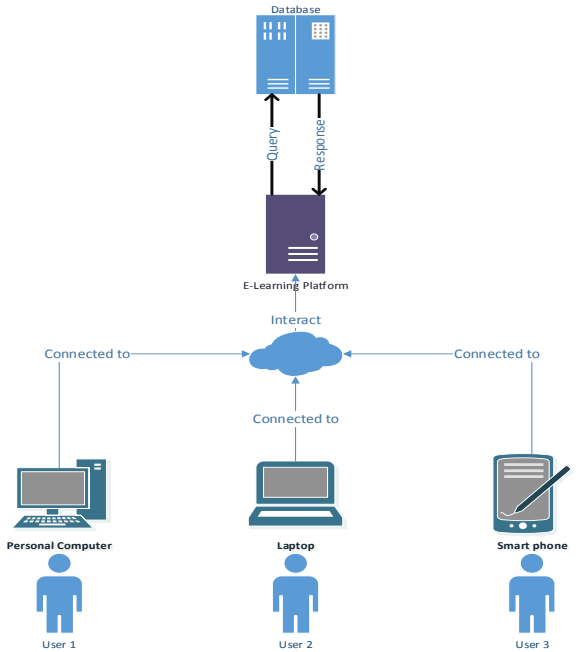


Fig. 1. High Level System Description

IV. SYSTEM FEATURES

The system includes a rich set of features that allows it to be used for wide range of educational purposes. The system features are given below:

- Management and statistics capabilities and reporting.
- Auto evaluation of questions, quizzes, and exams.
- Record keeping which keeps track of student progress.
- Quizzes and assignments generation that allow students to take quizzes and solve assignments.
- Live lectures availability feature that is managed by system administrator.
- Video hosting and slides hosting available as a choice for the system administrator (according to the organization copy rights).
- Rich user interface that has a sophisticated formatting, multimedia, and depth of interaction. The user interface platform serves the users with different roles. Some roles have special panels for observing a course or managing it.
- Progress-oriented or time-oriented features. The progress oriented systems allow the student to proceed with a course according to their individual progress. For example, a student who did not open a class/video lecture /assignment cannot open the next one unless he complete finishes the current. While the time oriented systems allow the student to proceed with a course according to the courses time schedule. For example a student finished a class he cannot take to the next class unless its pre-determined time comes.
- Security features are incorporated into the system. The database is protected from SQL (Structured Query Language) injection by methods of sanitization. Also, the passwords are stored in the database as irreversible hash format so if database is illegally accessed, stolen data cannot be used to access other services.
- Compatibility with mobile browsers that uses XHTML, CSS and other scripts are written in a way so that when rendered on a browser, the web page fills the window. However, it is a bad design to fill the same pages on a relatively small mobile screen so we use a methodology called "one solution fits all" through applying the open-source solution "Bootstrap" for resizing and reordering elements on the pages [8].

V. SYSTEM IMPLEMENTATION METHODOLOGY

The implementation of e-learning system utilized various techniques of software development. First, the database was examined using several ERD (Entity Relationship Diagram) models. These models were studied and compared in order to come up with the most suitable version that supports the current features as well as possible future expansions. The new e-learning system is implemented using the software engineering models and tools described below.

A. The Implementation Environment

After completing the design phase and settling on the most suitable entity relationship diagram for the system which supports the desired features, we selected the tools and languages that are to be used in the implementation phase. A brief description of the used tools and languages is given in Table 2. The table includes back-end and front-end tools and languages as well as tools used on both sides.

TABLE II. TOOLS AND LANGUAGES

Item	Description
XAMPP	a development tool that allows website designers and programmers to test their work on their own computers without access to a server on the Internet
Apache	a popular web server (HTTP server)
MySQL	a popular open source database. It is a relational database management system
CodeIgniter	open source development web application framework [9] that used for building dynamic web sites with PHP
PHP	a server-side scripting language designed for web development but also used as a general purpose programming language [10]
Bootstrap	a free collection of tools for creating websites and web applications
Ajax	a group of interrelated Web development techniques used on the client-side to create asynchronous web applications
HTML	Hypertext Markup Language is the main markup language for creating web pages to be displayed in a web browser
CSS	Cascading Style Sheets is a style sheet language used for describing the presentation semantics (the look and formatting) of a document written in a markup language
Sublime	a cross-platform text and source code editor
Bitbucket	a web-based hosting service for projects

B. The Implementation Crieteria

The implementation crieteria demanded the use of an architectural design that separates the backend logic from frontend presentation, handles interactivity through active and dynamic web pages, prevent user from direct access to the database and, at the same time, suited for web applications. As such we designed the system to follow the MVC (Model–View–Controller) pattern [11] as opposed to other aproaches [12,13]. Figure 2 shows the user interaction with MVC Pattern.

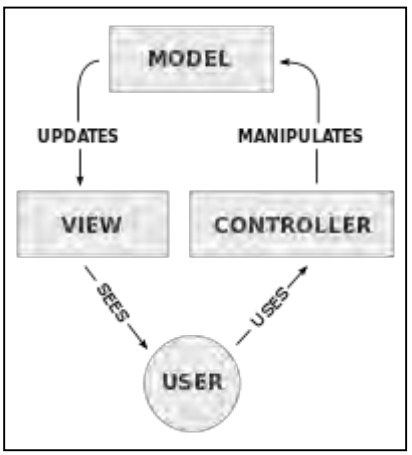


Fig. 2. User Interaction with MVC Pattern

MVC is a three-layer design architecture that decouples interface from navigation and from application behavior. It is used for interactive applications. Keeping the interface, the application, and the navigation separate simplifies implementation and enhances the software reuse.

- Model – contains all application-specific content and processing logic, including all content objects, access to data/information sources, and all processing functionality that is application specific.
- View – contains all interface specific functions and enables the presentation of content and processing logic, including all content objects, access to external data/information sources, and all processing functionality required by the end user.
- Controller – manages access to the model and the view and coordinates the flow of data between them. In a web application, the view is updated by the controller with data from the model based on user input.

There many benefits from using the MVC model as it is an architectural design pattern that separates the backend logic from frontend presentation. It handles interactions through active and dynamic web pages. It prevents user from direct access to backend database and is very well suited for developing web applications.

VI. System Results

We used the concepts described in the Implementation Methodology Section to implement our E-learing system and managed to develop a felxible system with strong set of features. We have implemented the coressponding number of views, controllers and models based on MVC pattern that we used in the implementation process. The result of our system implementation includes 36 views, 40 view helping elements, 23 controllers, 19 models, 138 functions in the controllers, and 135 functions in the models. The system was tested by a selected group of students. As an example, we include a sample of user interface for the system in Figures 3-4.



Fig. 3. User interface – Registration screen



Fig. 4. User interface – Sign-in screen

We have implemented the user views that can be run over different devices like desktops, tablets and mobiles. We also used asynchronous data exchange through AJAX and XML to exchange data between the system and the interface without the need to reload the view, making the user view more dynamic and easier to use.

The software model used supports security features to protect the system from illegal attacks (as the user interacts with the system indirectly). To implement the system with this concept, we used the CodeIgniter framework for building the dynamic web sites with PHP. We also explored other security concepts and tested them on the e-learning system. These included SQL Injection, XSS (Cross-side-scripting), session hijacking, middle man attack, and DOSA (Denial of service attack) [14, 15].

Using the concept of code versioning we were able to keep a record of the code to be viewed, edited or added during the development process. Each code portion that is edited, added or deleted is saved in the records and can be viewed as part of the whole system. In order to use the concept of code versioning, we used the GIT software [16] which is a distributed revision control and source code management system.

The new e-learning system can be integrated with other smart education environments. In fact, e-learning is now an essential component in building intelligent campus. Intelligent campus refers to an intelligent environment which encompasses the building blocks needed for smart education [17]. It includes several other components such as social networking, health care, building management, all of which are supported by smart network services and ICT platforms.

VII. Conclusion

In this work, the design and implementation of e-learning systems is described based on our experiences in the academic environment from both faculty members view and students requirements. The e-learning system is developed using commonly available (and free) tools and languages. The system is capable of satisfying the needs of educational institutions as it provides a rich set of features and flexible user interface screens (through labtops, tablets, or mobile devices). The list of features

of the system includes the ability to provide course program based on both certification and non-certification (continuing education) options. The system provides course administration and scheduling to control the creation of new courses; adding materials to any course, and update, edit, or delete any course as well as scheduling the courses according to the requirements of educational program.

The aim of our comparison was to understand how each e-learning system works, and what features and tools are available in each one, so we can start our system with strong base of knowledge and experience and add more features that will be useful to the students and the instructors.

The new system provides performance tracking features to allow the user to keep a record of his/her progress, grades and performance in each course he/she is enrolled in. It also provides question banking feature that allows the instructors to insert, edit and remove questions related to a specific course or even a specific chapter of the course. The system contains other needed features such as system administration interface (that give the ability to control and manage the system content and users privileges) and a registration system (that allow the users to enter the information that the system needs to consider as members in the system).

The implemented system is capable of holding the data for the number of students in our university. The system user interface includes 36 screens, each of which is designed to give the ease of use and the professional look and feel. As a matter of fact, the same design methodology and the tools utilized in our model can also be utilized in developing similar systems to e-learning and become part of an overall smart educational environment. The open software platform that is accessible by a variety of user devices can be used along with a location-based application in order to provide customized smart educational services. Furthermore, combining these applications with an e-health system is a topic of our current work plan.

REFERENCES

[1] Robinson, Rhonda; Molenda, Michael; Rezabek, Landra. "Facilitating Learning". Association for Educational Communications and Technology. Retrieved 18 March 2016

[2] R. C. Clark, R. E. Mayer, "E-Learning and the Science of Instruction," Wiley, 2011.

[3] J. Moore, C. Dickson-Deane, K. Galyen , "e-Learning, online learning, and distance learning environments," Internet and Higher Education, 2010 Elsevier Inc.

[4] A. Y. Alsabawy, A. Cater-Steel, J. Soar, "E-Learning Service Delivery Quality," Information Science Reference, IGI Global, Chapter 6, 2013.

[5] D. McIntosh, "Vendors of Learning Management and E-learning Products," Learning Management Vendors, Trimeritus eLearning Solutions Inc., 2015.

[6] Moodle – Developer Docs, https://docs.moodle.org/dev, April 2015

[7] L. Breslow, D. Pritchard, J. DeBoer, G. Stump, A. Ho, D. Seaton, "Studying Learning in the Worldwide Classroom Research into edX's First MOOC," Research and Practice in Assessment Journal, Volume Eight, 2013

[8] Bootstrap open source - http://getbootstrap.com, April 2015

[9] CodeIgniter – User manual – http://ellislab.com/codeigniter/user-guide, April 2015

[10] PHP Manual – http://www.php.net/manual/en/ index.php, April 2015

[11] I. Sommerville, "Software Engineering," Pearson Education, 2011.

[12] R. S. Pressman, "Software Engineering: A Practitioner's Approach," McGraw-Hill Higher Education, 2010.

[13] P. Kruchten, "Architectural Blueprints—The "4+1" View Model of Software Architecture," IEEE Software 12 (6), 1995.

[14] Google XSS reference, https://www.google.com/about/appsecurity/learning/xss/index.html, April 2015

[15] Open Web Application Security Project, https: //www.owasp.org/index.php/SQL_Injection, April 2015

[16] GIT reference, How to use GIT, http://git-scm.com, April 2015

[17] L.f. Kwok, "A Vision for the Development of i-Campus". Smart Learning Environments. Vol 2, No. 2, 2015.

Introducing Inquiry-Based Learning to Estonian Teachers: Experiences from the Creative Classroom Project

Nataša Hoić-Božić*, Mart Laanpere**, Kai Pata**, Ivona Franković*, Signe Teder***
* Department of Informatics, University of Rijeka, Rijeka, Croatia
** Tallinn University, Tallinn, Estonia
***BCS Koolitus, Tallinn, Estonia
natasah@inf.uniri.hr
martl@tlu.ee
kpata@tlu.ee
ifrankovic@inf.uniri.hr
Signe.Teder@bcs.ee

Abstract - The project "Creative Classroom" was launched in September 2014 with the main goal of assisting Estonian schools in their digital turn. The project is co-financed by the Erasmus+ program and will continue until the end of August 2016. Partners from Estonia, Finland, Slovenia and Croatia work together to design and deliver innovative training for Estonian teachers in order to enable them to use tools and methods for e-learning. Under the guidance of European educational technology experts, the group of Estonian teachers participate in four workshops on topics concerning ICT and teaching methodologies such as trialogical learning, flipped classroom, project-based learning, game-based learning, problem-based learning, and inquiry-based learning. They are expected to develop innovative lesson scenarios which are going to be published as an electronic lesson guide book and shared with other teachers in Estonia and across Europe. This paper describes experiences in delivering the workshop about inquiry-based learning (IBL) and e-tivites. The IBL is introduced as question-driven or problem-driven approach to learning based on seeking new knowledge and understanding. It involves students investigating complex, ill-structured problems. As an example of an IBL problem, teachers perform a WebQuest e-tivity using Web 2.0 tools.

I. BACKGROUND

New ICT technologies are transforming the way of accessing information and therefore the way people learn. Education is going through a major transformation. The main challenge for the 21st century education is not only to provide ICT infrastructure, but to develop teacher professional development programs that will provide them with knowledge and skills on how to use ICT to support pedagogy appropriately and meaningfully in enhancing learning experiences among students. Teachers have to understand how to facilitate and inspire student learning and creativity, support creative and innovative thinking, engage students in exploring and solving authentic problems using digital tools and resources, and promote student reflection using collaborative tools and engaging them in knowledge construction while learning with other students in face-to-face and virtual environments [1].

One of the Erasmus+ funded projects about new innovative digital curricula and educational methods development and professional development of teachers in ICT methodologies is the "Creative Classroom" project aiming to assist Estonian schools in the digital turn. The project started in September 2014 and lasts for 24 months. BCS Koolitus, one of the leading Estonian ICT training company, together with partners from Tallinn University, Metropolia University of Applied Sciences, University of Rijeka – Department of Informatics, and University of Ljubljana forms a consortium that brings together European experts in the field of innovative learning and teaching. Under their leadership and guidance the group of Estonian teachers will learn about various approaches and methodologies that can be adopted in their classrooms.

The main tasks of the project are:

- Gathering and interpreting evidence of readiness of the shift to innovative education from the aspect of teachers, students and schools in Estonia;
- Designing and delivering up-skilling workshops for Estonian teachers with the combined knowledge from Estonia, Croatia, Slovenia and Finland;
- Developing an electronic guidebook of innovative ICT learning and teaching methodologies and tools for an open use in schools Europe-wide;
- Organizing dissemination events to showcase and propagate the developed guidebook.

The main result of the project is the developed electronic lesson guide book. In addition, the project process provides teachers with a community to exchange ideas and experiences on innovative ICT in teaching practices.

II. SURVEY ON THE STATUS OF DIGITAL TURN IN ESTONIAN SCHOOLS

One of the first activities in the project was to conduct a survey about the current status of digital infrastructure,

ICT teaching practices, digital competences, and training needs of teachers in Estonian schools. The survey questionnaire was composed by BCS Koolitus and Tallinn University, and sent to all primary, basic and high schools. The questions were inspired by the recent Estonian National Lifelong Learning strategy 2014–2020, inquiring about the readiness of schools to implement the digital turn towards 1:1 computing and the new learning paradigm [2]. An open-source online survey software LimeSurvey was used for data collection. Questionnaire was not anonymous by default, but respondents were able to choose not to reveal their identity. Out of 541 Estonian schools, 165 responded to the survey. Participants were mostly school principals or vice principals, and IT specialists.

The most interesting results reveal that there are 12 pupils per one desktop or laptop computer in Estonian schools on average, nearly all schools are with Wi-Fi availability in school buildings, but only a few schools have tablet computers. Different factors may affect teachers' motivation to use ICT in teaching and learning but the key factors are school leaders, availability of equipment, teacher training, colleagues and students, and the new national curriculum. Mathematics is the subject where ICT is used most often, followed by geography, biology and foreign languages. As the biggest barriers that hinder large-scale adoption of technology in teaching and learning, the responders identify time pressure, poor technological infrastructure, insufficient ICT skills, and not enough subject-specific digital learning resources.

Respondents gave their opinions regarding the needs for teacher training in line with some topics relevant for the aims of the "Creative Classroom" project. The results illustrated in Figure 1. show that the respondents perceive every potential topic suggested by the project plan as highly relevant.

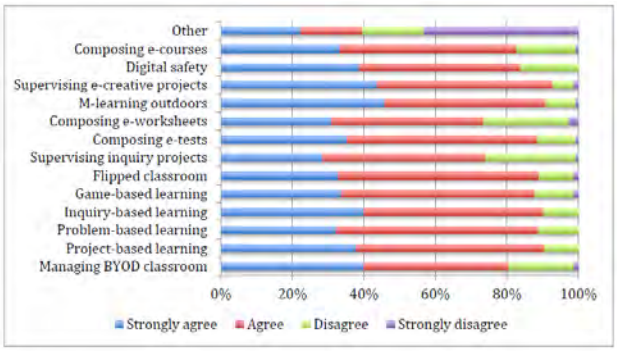


Figure 1. Perceived teacher training needs.

III. "CREATIVE CLASSROOM" WORKSHOPS

In the context of the "Creative Classroom" project, the group of Estonian teachers attend four workshop sessions handling various approaches and methodologies that teachers can adopt in their classrooms in order to advance competence and skills in innovative ICT use in education. Based on the knowledge gained in the workshops, the teachers will develop a lesson guide book with the best practise of different ICT-related innovative learning and teaching methodologies that all teachers across the Europe can use in their classes.

Under the leadership of Tallinn University, the partners develop workshop programs for teachers to introduce different ICT-related innovative teaching methodologies, modalities and tools.

The workshop program is divided into 64 academic hours and includes both theory and examples of best practices. The program includes learning objectives and framework, as well as innovative ICT-enabled assessment approaches that better capture 21st century skills, learning practices, teacher roles and infrastructure.

The focus group consists of 20 Estonian teachers chosen based on the willingness the schools have in participating in the project, (as expressed in the questionnaire). The participants were selected from schools all over Estonia. Some teachers are chosen because their schools had been equipped with a full set of tablets but the teachers miss the ideas and experience in applying those devices in teaching and learning. Another reason for participating in the project were the teachers' wishes to carry out lessons in more interesting and attractive ways for their students.

The workshops are held in Estonia over a period of 1 year. The focus group meets four times on two-day workshops to learn and collaborate on the topic of innovative ICT methods and tools in education. In every focus group meeting two methodologies are introduced with the related ICT tools and methods. For every methodology introduced an example of best practice is also presented. Based on the knowledge gained from the meeting, teachers start preparing the lesson guide during the focus group and continue working on it as homework.

Topics of the workshops (Figure 2) are: designing innovative learning scenarios and trialogical learning, flipped classroom, project-based learning, game-based learning, problem-based learning, inquiry-based learning, learning analytics, and evaluation and assessment of students' work.

In order to enable the exchange of resources and ideas between the workshops' participants a course "Digididaktika" was created with the help of Eliademy tool (Figure 3). Eliademy is a platform for instructors to create, share and teach online courses. The "Digididaktika" course is used intensively during and after the workshops [3].



Figure 2. Workshops' methodologies.

A. *The 1st workshop of the project „Creative Classroom"*

The 1st workshop was conducted by experts from the Tallinn University and the Helsinki Metropolia University of Applied Sciences, with the addition of a web lecture by an expert in learning sciences Paavola from University of Helsinki. The goal of the first workshop was to stimulate and energize teachers` creativity with regards to planning and carrying out lessons with ICT.

The workshop focused on how to use trialogical learning design [4] approach to conduct the flipped classroom lessons with the help of innovative technology. Trialogical learning refers to those forms of learning where learners are collaboratively developing, transforming, or creating shared objects of activity (such as conceptual artefacts, practices, products) in a systematic fashion. In trialogical learning design approach interaction and learning should take place around shared digital artefacts that are conveyed from one learning activity to the next and are matured with the help of appropriate digital tools and practices (e.g. self-creation, co-creation, self-reflection) during the consequent learning scenario events. When in different phases of learning scenarios these conceptual artefacts (as digital products) are reworked, personal and group usage of different digital tools enables revelation of certain aspects of knowledge and allows deeper learning.

The flipped classroom methodology proposes that students have to do pre-learning before the actual lessons and use lesson time for co-creation. Both the leadership role in the classroom and the home/classroom activities will be flipped. Combining trialogical learning design principles and the flipped classroom approach was found to be particularly promising. The first phase of a learning scenario can be done at home and should focus on individual knowledge-building to reveal personal understanding or on data-collection, and serves as input for the co-creation phase done in the classroom. The co-created knowledge and digital artefacts should then, in the next scenario phase, be used for individual reflection tasks or for addressing certain perspectives in knowledge. These reflections and new perspectives should again be brought back to creative group work in the final phase. Conducting flipped classrooms with trialogical learning design approach works particularly well with project-based and problem-based learning.

At the workshop, step-by-step flipped classroom scenarios were developed in teams of 3-4 teachers with the help of different web-based tools like Padlet, Tricider, Kahoot!, and LearningApps. The next step was to visualize the created scenarios. Adhesive papers of different colours were used to signify activities and tools as a preparation for using a lesson scenario building tool LePlanner in the next workshops [1]. After the workshop, the teachers created their own flipped classroom scenarios and tested them out in schools. They reflected on the applicability of trialogical learning design approach for flipped classroom design.

B. *The 2nd workshop of the project „Creative Classroom"*

One of the main topics of the 2nd workshop was game-based learning with ICT. Experts from the University of Ljubljana gave an overview of game-based learning in innovative learning and teaching. Games can be used in different ways and in different stages of learning. A game could be an introduction to a new topic, tool for activating group assignments or a part of homework. Participating teachers created 4 gamified learning scenarios in teams and the ICT students from the University of Ljubljana later programmed the scenarios into real digital learning games.

Second topic of the 2nd workshop was project-based learning with ICT. The group reached a mutual understanding that project-based learning is a practical form of learning that enables collaboration and integration of different subjects and that digital technologies can be used in various phases of the project or can be the actual results of the project. However, there were some disadvantages identified relating to the application of project-based learning, such as tight time-frame attributed by the curriculum, lesson format, difficulties in involving several teachers into the multidisciplinary project-work, dissimilarity with the traditional way of learning, more complicated assessment and additional, non-traditional work for the teacher.

As the next step, the teachers individually created project-based learning scenarios, tested them in their schools, and analysed the outcome through reflections they posted in their blogs.

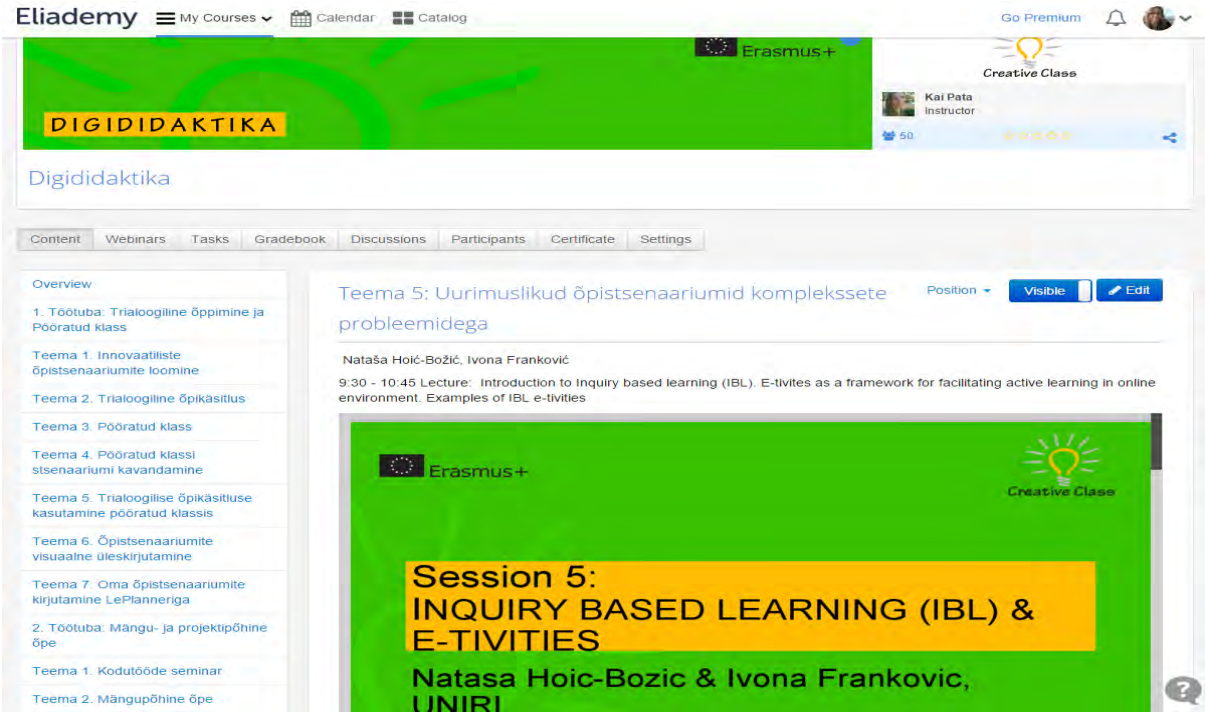


Figure 3. Eliademy Digididaktika e-course.

IV. THE 3RD WORKSHOP AND INQUIRY–BASED LEARNING

The first topic of the 3rd workshop was problem-based learning (PBL). Experts from the Helsinki Metropolia University of Applied Sciences introduced the Jonassen's (2000) problem typology that describes different well- and ill-structured problem types. Using that typology, well-structured problems (e.g., algorithmic, story problems, rule-using problems) can be addressed in the lessons with simple one-activity- or whole-lesson-based problem tasks, whereas complex problem types (e.g., diagnosis-solution problems, dilemmas, trouble-shooting problems, case problems, etc.) require complex learning scenarios such as project-based learning or inquiry learning. On the first day of the 3rd workshop, teachers tested the functionalities of ICT tools that support problem based approach. This part focused on digitally supported learning activities with well-structured problem types. These problems have known solutions, known problem solving paths, and require knowing/finding algorithms and rules [5].

During the second part of the workshop, experts from the University of Rijeka presented the inquiry-based learning approach in different subjects with examples. E-tivites as a framework for facilitating active learning in online environment, were introduced together with Web 2.0 tools for their realization. As part of group work, participants performed and tested Web 2.0 tools: Wikispace for creating wikis, Diigo for collecting and organizing bookmarks and other resources, and MindMeister for online mind mapping and brainstorming. As an example, they conducted IBL e-tivity WebQuest.

A. *Inquiry Based learning (IBL)*

IBL is question- or problem-driven approach to learning based on seeking new knowledge and understanding [6]. It involves students performing investigations of some sort in order to address questions or solve problems. IBL is an active and learner-centred approach in which the role of the teacher is to act as a facilitator. The students are taking increasing responsibility for their learning and the development of skills.

In IBL focus is on complex, ill-structured problems. In contrast to well-structured problems typical of problem based learning, ill-structured problems have either unknown problem solving paths, unknown solutions, or both, or have several alternative solutions. Problems which belong to ill-structured problems are those with unknowns in the problem space (e.g., diagnosis-solution problems), problems wide open to interpretation (e.g., dilemma solving problems), and problems with different number of solutions (e.g., decision-making problems) [5].

Inquiry–based learning is often described as a cycle which implies 5 phases: ask, investigate, create, discuss, and reflect [7]. In the first phase the questions are formulated. During the investigation phase, students or groups of students plan, read, and do research on the Internet. They collect some information which will be merged together in the "create" phase in which students may generate new thoughts, ideas and theories, and write them down in some kind of report. During the "discuss" phase, students debate about their findings with the other students, share their ideas with each other, and ask others about their own experiences and investigations. The last phase is students' reflection about their findings and about the learning process in general. Students look back and possibly take some new decisions.

TABLE I. WEBQUEST IBL E-TIVITY EXAMPLE

Name of IBL e-tivity	WebQuest – 11,000 years of Estonian History
Subject and grade	History; Could be adapted to different learners
Purpose/Learning outcomes	WebQuest activity engages learners in the use of the WWW in order to complete a task related to the course or some topics. As a collaborative learning activity, WebQuests can be used to focus groups of learners on task while providing a variety of online resources necessary for completing their assignments [11]. Goals: • Learners will utilize Web resources to complete a task • Learners will work with their peers to complete a course-related task • Learners will build skills for interacting with other learners in course • Learners will get familiar with several Web 2.0 tools. During this WebQuest, learners will investigate the history of the Estonian state, people and country. They will try to understand the uniqueness of the people who have lived here and historical events that have most influenced them.
Collaborative Learning	Groups of 3-5 students
Time required	3 to 5 days
Tools	- Access to the World Wide Web - Web 2.0 tools: MindMaster or similar tool for brainstorming Diigo for collecting and organizing bookmarks and other resources Wikispace for creating wikis as summary Blog for keeping learning diary with reflections on the learning process
Spark	Visit to Estonian history museum in Tallinn (if possible), or web page http://www.ajaloomuuseum.ee/
Structure/actions ***Ask – Investigate – Create – Discuss – Reflect***	**1. Ask** Main questions: *What do we know about Estonian history through several millennia?* Each group will pick 1-2 inquiry questions based on the "spark", for example: How old is Estonian statehood? Who are Estonians? Is Estonia known in the rest of the world? Is Estonian a simple language? Is Estonia a Nordic country? … It is expected that different groups will choose different questions/topics for investigation. **2. Investigate** Each group starts by identifying a list of concepts/key words which are related to the selected questions and searching web resources (web pages, pictures, video, etc.). They create a Mind Meister mind map while brainstorming about the concepts, and use Diigo for collecting the resources. **3. Create** Each group uses Wikispace for creating a short report summary about the information found. **4. Discuss** Based on the created wikis, learners share their ideas with each other, and ask other groups about their own experiences and investigations. They exchange their opinions about the historical events that influenced the country the most. **5. Reflect** Learners begin to write blogs reflecting on their work during the WebQuest e-tivity and about what they had learned. They express their opinion about the main question and decide by themselves which historical events were the most important.

B. E-tivity WebQuest

E-tivity is a term used to describe a framework for facilitating active learning in an online environment. An e-tivity involves learners interacting with one another and with the teacher (e-moderator) in an online environment in order to complete a particular task. At the beginning of an e-tivity teacher usually provides a 'spark' - small piece of information, stimulus or challenge, for example an attractive video or picture. Students take part in the e-tivity by responding to the 'spark'. E-tivities can be designed in line with the IBL approach. Today, Web 2.0 tools have become a common technology for performing e-tivities [8].

Before starting an e-tivity, teacher should carefully plan and design all important elements such as purpose or learning outcomes, time required, number of groups for collaborative learning, spark, tools, and structure or actions.

Table I. presents one possible e-tivity description form which was used for "Creative Classroom's" 3rd workshop. This form was designed for e-tivity WebQuest. The simplest WebQuest has a form of collaborative learning activity during which a group of students explore Web resources to find out „what is X" and write a summary with definitions and examples. "X" could be any term or subject, e.g., „fruit decomposition," „World War II", „Estonian history", „Croatia", etc. In the example presented in Table I., the structure of the WebQuest is more complex and in line with IBL approach so the students' actions follow four phases of the IBL learning cycle.

During the workshop teachers were supposed to act as students and perform a WebQuest "Learning about Croatia". They explored Web resources in order to investigate and prepare the answer to the main inquiry questions: *"What do we know about Croatia/Is Croatia a place worth visiting?"*

The evaluation in a form of an anonymous survey was made at the end of the workshop with a goal to establish in which measure participants were satisfied with the workshop content and instructors. In one part of the

survey, the workshops' participants were asked to express their opinion using the Likert scale (options from "I am not very happy with it" to "I am very happy with it"). In other open ended questions teachers were supposed to mention the topics for which they considered to be of most use in in their daily job, and give some suggestions and proposals for the next workshop.

The survey was completed by 14 participants (70% teachers who were enrolled in the project) who were mostly rather happy or very happy with relevance of the workshop topics, learning materials, e-learning environment Eliademy, competence of lecturers, and training facilities. Special value to the survey was given by teachers' comments on what workshop topics they found helpful. They mostly liked the practical work during the workshop and agreed that the different online Web 2.0 tools for PBL and IBL would be useful in their future schoolwork.

The results were not statistically significant, but helped the project team in deciding where to focus future efforts in preparing the next workshop. Based on the survey results, it can be said that teachers' experiences were very positive. All groups were engaged in performing the task and using Web 2.0 for its realization.

As the next step after the workshop, teachers had to choose a topic of their own and create two learning scenarios around well- and ill-structured problems and test these out with their students. For the time being, it is too early to see if the teaching in Estonian schools improved after the workshop. Some research could be planned as follow-up activity after the project.

V. FUTURE PLANS

The 4th "Creative classroom" workshop will be held in March 2016. The topics will be methods of assessment and evaluation, and learning analytics. The participants will be introduced to diagnostic and summative e-assessment using objective test items, formative e-assessment, assessing and evaluating group work and collaborative learning projects. They will learn how to analyse test results and how to use a selected tool for visual and numerical presentation of data for quick and accurate analyses. At the end of the second day of workshop 4, the participating teachers will start analyzing the learning scenarios they have developed and addressing them from the trialogical learning design point of view [4].

The main result of the project will be the developed electronic lesson guide book in Estonian and in English, developed in the LePlanner environment [9]. A lesson scenario building tool LePlanner was co-developed with the teachers' involvement during the Creative Classroom project. The tool allows constructing digital lesson scenarios and sharing them. LePlanner allows developing scenarios from sequential learning activities, connecting learning activities with particular learning outcomes, enriching scenarios with digital resources, tools and applications, and viewing both the teachers' and students' activities in the same timeline. The interactivity level of learning activities according to the trialogical learning principles can be added to each activity [10]. After each workshop, the teachers prepare their visual learning scenarios.

Even though the main result of the project is the development of the electronic lesson guide book, another result is important as well. The project will provide teachers with a community to exchange ideas and experiences on innovative ICT in teaching practices, which will strongly support the further development of the field. The project's consortium is hoping that the project will be a Europe-wide encouragement for the shift to the 21st century education.

ACKNOWLEDGMENT

This work has been supported by Erasmus+ "Key Action 2: Strategic Partnerships in the field of education, training and youth" project Creative Classroom (2014-1-EE01-KA201-000525).

REFERENCES

[1] ISTE Standards for Teachers, available at: http://www.iste.org/standards/ISTE-standards/standards-for-teachers Accessed 16 February 2016

[2] M. Laanpere (2015) Creative Classroom digipöörde teemaline küsitlus Eesti koolidele Available at: http://www.bcskoolitus.ee/creativeclass/ Accessed 16 February 2016

[3] Eliademy, available at: https://eliademy.com/ Accessed 16 February 2016.

[4] S. Paavola & K. Hakkarainen (Eds.), Computer Supported Collaborative Learning Practices: CSCL 2009 Conference, 2009

[5] D. H. Jonassen & W. Hung (2008) All Problems are Not Equal: Implications for Problem-Based Learning. Interdisciplinary Journal of Problem-Based Learning, 2(2). Available at: http://dx.doi.org/10.7771/1541-5015.1080 Accessed 16 February 2016

[6] A. Aditomo, P. Goodyear, A.-M. Bliuc & R. A. Ellis Inquiry-based learning in higher education: principal forms, educational objectives, and disciplinary variations, Studies in Higher Education, vol. 38, no. 9, pp. 1239-1258, 2013.

[7] M. Pedaste, Mario Mäeots, Leo A. Siiman , Ton de Jong , Siswa A.N. van Riesen , E. T. Kamp , C. C. Manoli, Z. C. Zacharia, E. Tsourlidaki, Phases of inquiry-based learning: Definitions and the inquiry cycle, Educational Research Review 14 pp. 47–61, 2015

[8] G. Salmon, E-tivities: the key to active online learning. Psychology Press, 2002.

[9] LaPlanner, available at: http://leplanner.net/ Accessed 16 February 2016

[10] T. Väljataga, H.D. S. Fiedler, M. Laanpere, Re-thinking digital textbooks: Students as co-authors. In: The 14th International Conference on Web-based Learning, 2015, pp. 143−151 Springer-Verlag Heidelberg.

[11] R. Watkins, 75 e-Learning Activities – Making Online Learning Interactive, Pfeiffer, 2005.

Mobile Robots approach for teaching programming skills in schools

W. Werth, C. Ungermanns
Carinthia University of Applied Sciences (CUAS)
Europastraße 4, 9524 Villach, Austria
Phone: +43 4242 90500
w.werth@fh-kaernten.at
c.ungermanns@fh-kaernten.at

Abstract - Teaching programming skills in schools having non-technical background is seen as a challenging task. In this paper an approach is described where mobile robots are used to get students of non-technical schools interested in programming tasks. The proposed method tries to combine pupils play instinct with the intrinsic motivation to win a challenge. Thereby physical and mathematical motivated questions lead directly to vital necessity in using software. The concept which has been performed in a class of 14 year old students is presented and the maid experiences are discussed.

I. INTRODUCTION

Following the experts of Samsung Innovation Center every child that is born from 2016 on will be educated in programming of course in addition to writing and mathematics. Furthermore the European Commission estimates around 900,000 jobs in the information and communication technology for 2020. Even today, say the experts, around 90 per cent of all jobs require certain IT skills. In order to fulfill the expectations from industry new educational approaches or even reform ideas of school's curricula are seen as a necessary step. In the UK, for example, this step has already been set. "Computing" is now part of the general curriculum. It was published recently that already pupils in primary schools in China are educated in programming in order to have an advantage in their future careers [1].

A. Educational Situation in Engineering

The current industrial situation in many European countries tells about the need of students and later on employees in the so called STEM fields (Science, Technology, Engineering and Mathematics). There are a lot if initiatives in Austria founded and supported by the ministry of science, research and economy like "young science". In this field especially Universities and Universities of Applied Sciences are motivated to participate and to share their educational experience and their knowledge in industrial needs with schools. This is often done in common projects like "Sparkling Science" program.

B. CUAS Activities

Carinthia University of Applied Sciences (CUAS) makes a huge contribution in the support of young scientists by various approaches since many years. People from secondary schools in the age of 14-16 years can get their first experience in electronics at the "TechCamp" which takes place during the summer holidays in the labs at CUAS. Classes, even of primary schools, can pass a taster day at our location in order to come into contact with technical tasks. Participation in local events like "girls days" addresses young women to increase their motivation to go into a technical education. These activities are mainly due to our infrastructure (Physical labs) and the pedagogical expert knowledge of university's staff.

On the other hand online laboratories are a modern approach suitable especially for young people. These labs provide access to experimentation equipment through the Internet. Technology for online labs together with a common framework for research based learning is the main goal of an international ongoing project at CUAS [2]. Thereby secondary school students and teachers are involved in development process.

C. A Teaching Approach

By the third approach we intend to collaborate with schools at a different level. Thereby the target group is just a specific class in secondary schools. Through team teaching students are educated in order to increase both motivation for the related subjects and quality of the corresponding learning outcomes. In this paper is referred to this approach, the main idea, the benefits/disadvantaged and results are discussed in details in the following chapter.

II. ROBOTICS APPROACH

Initially we have been looking for an appropriate way to motivate pupils, who attend a non-technical school, to build their own robots and to participate at the RoboCup Junior competition. This contest is aimed to students in secondary schools from the age of 11 to 18. Thereby according to the main directives teacher of the corresponding school should support their teams during

preparation but they must not solve the upcoming programming tasks. According to our experience motivating a teacher to become a supervisor and mentor of robot teams is a challenging task.

On the other hand CUAS represent one of the 5 regional centers of RoboCup Junior in Austria. The main responsibilities of a regional center include supporting teams, train the trainer (teacher) and organizing local competitions.

Integrating student's task in existing or even new teaching subjects was found out as suitable procedure for gaining teams. Over the past few years the scientific connection has been found in physics lessons. However it turned out that student's competition success in strictly related to their programming abilities. Unfortunately programming courses are not offered in most non-technical schools, like Bundesrealgymnasien, in Austria in a suitable way.

A. Organizational Framework

To give a contribution to the mentioned challenge in increasing programming skills for pupils, mainly to be prepared for future jobs, was the biggest (first) motivation for our approach. Therefore our support was integrated in informatics lessons. In September the teacher of the proposed class and we entered into an agreement which combines programming courses and solving robotics tasks. Schools inputs are mainly given by the following points:

- Prepare a class of 14 year old pupils for the project
- Provide a suitable classroom
- Prepare and discuss basic programming exercises
- Support students during solution process
- Provide a Moodle course as framework (e.g. for submission of results)
- Evaluate projects

The student group consists of 16 pupils (boys and girls). So we decided to form 8 robot teams. It is important to mention the class provides a bilingual education (German and English). So an English version of the software could be used without difficulty. Most pupils did not have any experience in robotics and just some knowledge in programming. They used Scratch as a first programming language in previous periods [3].

The university support can be summarized in the following list:

- Give an introduction to robotics tasks
- Present and explain the used hard- and software components
- Provide a sufficient number of robot sets for free
- Prepare challenging exercises
- Support the students in class
- Organize a visit to robotics lab at CUAS
- Evaluate projects

The support group of the university consists of 3 students and one professor. It tuned out that the integration of the bachelor students was very helpful for the project outcome. First of all they had to prepare a sample solution of the proposed exercises using the same robot and software tools. Secondly now it was possible to give an individual team support during work in classes.

B. Used Robot and Software

We decided to use LEGO Mindstorms education set as a suitable robot platform. Thus students were able to build their robot within one 4 hour session in class. EV3 Brick and EV3 Servo Motors together with a Touch Sensor, a Color Sensor and an Ultrasonic Sensor formed the main components of the robot.

To bring the robot to life, EV3 software allowed the student to do that by writing a computer program for their robots. Even if pupil's experience would count for Scratch, LEGO EV3 Software was seen as better solution for the project due to its ability to provide very simple hardware access (e.g. connect sensors and read sensor data). Furthermore no additional configuration beside the installation was needed to make the system ready for programming [4,5].

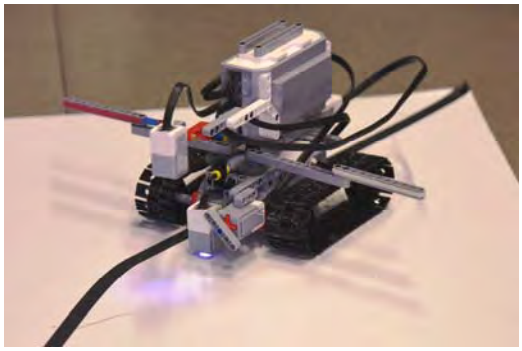


Figure 1. One of the robots used in class

C. Robot Project Steps

We started the project with an introductory session. Thereby the general motivation for the use of robots in industry and the importance of corresponding programming tasks was figured out. The main idea and mainly the need of informatics for the upcoming projects were presented. Secondly the proposed robot set was introduced. As an important outcome of the first unit every team had to build a roadworthy wheeled robot.

As a next step the unit was to familiarize the students with the operation and function of the motors and sensors of the robot. In addition the young engineer's task then was of course to become acquainted with the basic software elements and features, i.e.

- Using the programming environment
- Placing blocks and understanding sequences
- Using loops

- Controlling the flow by a Switch Block
- Applying and adjusting motor blocks
- Getting and using sensor data

The first exercise was just to make a predefined movement of the robot. We gave the students tasks like:

"The robot has to move straight forward for 50 cm."

"Move the robot from a starting position to a position which is fixed on a table."

"Make a predefined movement and force the robot to come back to the starting position."

"Redo the task but double the speed of the robot."

In this approach the needed actuation of the motors where calculated in advance. This was done for example by calculating the number of required rotations to cover the proposed distance. The speed of the motors is kept constant in that case. So just basic mathematical relationships - how can one map a distance to number of revolutions of a wheel driven by the motor - were needed to solve the problem. Overall it can be said that most of the pupils managed the given tasks quite well.

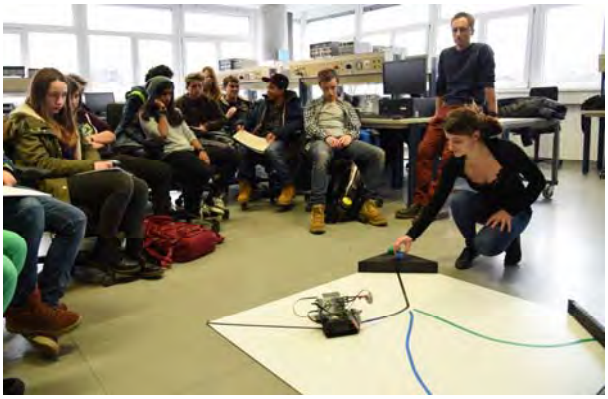


Figure 2. Final task presented to the students

In order to solve more challenging assignments students had to deal with the mentioned sensors in the second unit. Basic sensor functionality and related physical properties are covered as an important educational part. Thus the second pool of exercises included more complicated objectives the robot should solve. The students had to cover tasks like:

"The robot has to travel to a given endpoint but (predefined) obstacles are barring the way."

"Let the robot autonomously escape from a retainer enclosed on three sides."

Additional competence and abilities are now required to carry out the solutions. Basic features like

- Rotating the robot
- Measuring the distance
- Making decisions

had to be implemented in software. One of the most important elements in software design comes automatically into existence when the solution is created: reuse (elements) of your code.

Make a robot follow walls is one possibility for exploring environment. This was done by the use of Ultrasonic Sensors assisted by Touch Sensors. Another basic method to navigate the robot through a given terrain is to find and follow colored labels or lines. The robot can follow the edge of a line by using a downward-pointing Color Sensor. Thus in the third unit we increased the complexity of the tasks by solving the mentioned problem:

"Create a program for the robot to follow a line together with collision avoidance."

Elaborating on that objective was mainly done as homework. Whereas the basic tasks could be managed by the students themselves now we saw a different progress in the solution path. We noticed a different motivation of individual teams to provide a solution in time for the line following task. A feedback round regarding these difficulties showed the need of more step by step examples elaborated together in class.

After solving simple and even for beginners manageable exercises we introduced a more complex task for the remaining semester period. This should make the students feel confident and encouraged to participating RoboCupJunior competition. The main idea behind the task was to integrate solutions of the simple exercises as suitable ready to use parts of the big problem. The task was to program the robot that performs the following steps:

- Detect a specific part by a defined color
- Pick up the object
- Transport the object regarding to its color to a predefined location.
- Find the proper location by following a colored line
- Drop the object
- Travel back to the starting position and redo the task

First the already built robot could be used for that task only for the pick and place procedure the device has to be adapted. The extended functionality of the Color Sensor - distinction between different colors - is needed for both detecting the object and following the correct line.

One of the main recognitions that have been realized was the need of an initialization step before the robot starts working. This is mainly due to Color Sensor's dependency on environmental light conditions and measurement distance to the floor. Most of the line follower algorithms worked fine for a simple test path. But paths containing sharp edges or interruptions of the line required time-consuming redesign and a lot of test runs.

The big difference between giving an answer that shows just general functionality rather that developing a robust solution is seen as a central learning outcome for

the students. For solving complex tasks a clear concept is needed!

Overall is can be said that not all teams managed the final task – of course also due to time problems. From the supervisor point of view the pedagogical strategy has to be improved in two directions. First the proposed concept has to be developed together with the pupils before they start coding. In order to avoid frustrating team members a more intense mentoring is needed during the development phase.

III. FEEDBACK

A. Procedure

The presented project serves not only to spark student's interest of that class in robotics or engineering. It was our idea to define a long term collaboration framework between non-technical schools and the university. Thus the results of a questionnaire will help to improve our concept. Of course also a lot of individual discussions with students and the informatics teacher have been done.

B. Results of Feedback

The questionnaire was completed by the pupils during the last unit at CUAS. Thereby we also presented currently running robotics projects which should motivate pupils for STEM subjects in school.

A short summary of meaningful results of the questionnaire is given in the following list:

- About 90% of the students improved their programming skills by the robotics project.
- Again 50% are motivated to continue robot programming a little, 40% are highly motivated to do this.
- 40% agreed that the project helps to understand the need of math and physics, 40% were uncertain, 20% disagreed.
- 50% want to continue the collaboration, 45% maybe
- Nearly all of the students suggested to redo this project next year
- 25% decided to participate at the RoboCupJunior competition 2016

Finally some comments to the question "What should be improved?" is summarized:

"More time for programming, give a basic course in programming; explain the software in details, nice project, more explanations helpful, more units in school, detailed introduction to program needed, building bigger robots, redo it in the second semester, additional unit on Saturday afternoon only for robots"

C. General Comments

Based on the feedback we got it is our intention to continue the robotics project next study year. Students were elaborating quite well on their task unless they got lost due to missing basic knowledge in programming. We will increase the mentoring part by additional units in class and give a more detailed introduction into the used programming language at the beginning of the project.

Due to our opinion one of the main benefits of the method is seen in the intrinsic motivation to learn and use concepts of mathematics and physics for solving real world informatics projects.

IV. CONLUSION

In this paper a method for teaching programming skills in schools having non-technical background was presented. Mobile robots are used to get students interested in programming tasks. The main tasks of the project are given and discussed. The experiences lead to an improvement of the method for the next years.

ACKNOWLEDGMENT

The authors of this paper would like to thank the students and the teacher of the school BG/BRG St. Martin in Villach for the great collaboration. Without the commitment and good will of all parties, this project could not be realized.



Figure 3. Visit of students at CUAS Robotics lab

REFERENCES

[1] "Futurezone – Technology News", 2016. [Online]. Available: http://futurezone.at/digital-life/zukunftstrends-programmieren-als-hauptfach/103.264.324 [Accessed: 16-Feb-2016].

[2] "Global Online Science Labs for Inquiry Learning at School", 2016. [Online]. Available: www.golab-project.eu. [Accessed: 16-Feb-2016].

[3] "Scratch – Programming language", 2016. [Online]. Available: https://scratch.mit.edu/ [Accessed: 18-Feb-2016].

[4] "LEGO Mindstroms", 2016. [online]. Available: http://www.lego.com/en-us/mindstorms/?domainredir=mindstorms.lego.com . [Accessed: 16-Feb-2016].

[5] L. Valk, The LEGO Mindstorms EV3 Discovery Book, No Starch Press, US, 2014.

Age independent examination of algorithm creating abilities

Zoltán Attila Godó* , Dénes Kocsis** , Gábor Kiss*** and György Stóka****
* University of Debrecen, Faculty of Informatics – Department of Information Technology, Debrecen, Hungary
** University of Debrecen, Engineering Faculty - Department of Environmental Engineering. Debrecen, Hungary
*** University of Óbuda, Institute of Machine Design and Safety Engineering. Budapest, Hungary
**** Eszterházy Károly College, Faculty of Comenius - Department of Informatics, Sárospatak, Hungary
magortaltos@gmail.com

Abstract- We can clearly observe the fact that IT students' algorithmization abilities show a great difference from student to student. This affects learning programming to a facility level and in the end to the whole process of successful education. Just like in all abilities, genetic, social and learning processes all play their roles but in different amounts. The ability to create algorithms is present in humans as early as the beginning of childhood, mainly from inheritance, which, during the aging process, gets modified by all sorts of acquired other abilities. We've set a goal to create a test system – hardware and software – that can measure the individual's algorithmization abilities before learning to write. The same system can examine subjects of university levels. This universality makes statistic comparisons possible. The system is based on the so called light-programming principle, which was defined by us. We've also created specialized hardware and software to use.

Keywords: algorithm, ability, light-programming

I. Introduction

We can clearly observe the fact that IT students' algorithmization abilities show a great difference from student to student. This affects learning programming to a facility level and in the end, the whole process of successful education. In our everyday lives, we constantly create and process algorithms, we design action sequences and data-flow processes, but to understand them, we must understand their structure and core elements [2]. We have to consider developing problem-solving abilities as one of the most important tasks of education. Mathematics has been traditionally playing this role for ages. With the introduction of IT education, a new way was born, that requires less abstract thinking, but is also much more evident in terms of problems and solutions [4].

Just like in all abilities, genetic, social and learning processes all play their roles but in different amounts. The ability to create algorithms is present in humans as early as the beginning of childhood, mainly from inheritance, which, during the aging process, gets modified by all sorts of acquired other abilities. The difference is clearly present on all levels of education [6]. Some students – even in university – get the hang of things almost instantly, and soon will create astonishing things, while others can't even understand algorithms that are so basic, that even an elementary school student can solve them. Creating algorithms comes from a complex, high-level neurological activity. During the process of analysation, one has to understand both the steps that lead to the solution and also the individual subtasks. These make up the final solution together [7].

We've set a goal to create a test system – hardware and software – that can measure the individual's algorithmization abilities before learning to write. The same system can examine subjects of university levels. Therefore, we were looking for a system that relies the least on the acquired ingredients of the abilities and STM (short term memory). While planning our research, we considered followableness to be top priority [1]. This means that we'd like to conduct our experiments with the same subject at different ages, along other required dedicated tests. The universality of our system makes statistic comparisons possible at different ages [3]. The system is based on the so called light-programming principle, which was defined by us. We've also created specialized hardware and software to use.

II. Inherited and acquired components

We acquire knowledge of the methods for making accurate, symbolic calculations by learning them. However, some scientists assume that there are certain basic abilities regarding calculation and dividing into logical parts that we are born with. According to Rochel Gelman and Randy Gallistel, children know various calculation principles without learning. [5]

For example, they know that when counting objects, they only have to count every object once. They automatically understand that the numerals always follow each other in the same order, and that the last number they say is equivalent to the cardinality of the set. Algorithmization, that is, creating a series of steps that are needed to reach the desired result, shows a parallel development with the previously mentioned abilities.

Karen Wynn showed in her experiments [11], that infants of just 5 months of age are already capable of solving the simplest addition and subtraction operations. According to her results, when two objects are hidden behind a curtain, the infants do not expect to see only one item when the curtain is lifted (1 + 1 = 1). So when the

curtain is lifted and they see both objects, they don't show any sign of being surprised. Wynn found evidence to early capabilities of subtraction with the same method. This time, the infants saw them putting two objects behind the curtain, one of which was then removed in front of them. But when the curtain was lifted, and there were still two items behind it, the infants showed visible signs of being surprised (2 – 1 = 2).

It is observable at the age of 2-3, that children are capable of counting a set of objects by pointing at them. According to Karen Wynn, at this age, children are also capable of counting not only visible objects, but also sounds from different sources, meaning that they understand that counting can be applied to any visible or audible objects around them.

At the age of 3½, they understand that when determining the cardinality of objects, if they count each object only once, then the order does not matter, because they get the same result by starting with any other object. The perception of order is a key base of creating algorithms. At this age, they show sensitivity to error, which means that they notice when someone says the numerals in the wrong order or forgets to count one of the items, or counts something twice. They are also capable of isolating these errors from someone having an unusual, but correct way of counting.

Although children at the age of 3½ are capable of counting the elements in a set, that is, they know how to count, they don't understand what counting is good for, and what's the meaning of counting (e.g. they don't understand that if they count their toys, the number of toys they have is equal to the result of the counting) until the end of their 4th year. Understanding the functionality of counting, and the connection between counting and quantity becomes present just before they reach the age of 5.

By the time they reach the age of 6-7, they learn multiple counting algorithms. A portion of these is "invented" by themselves. They also copy other children's methods of the same age. Most of these early strategies involve counting with the help of their fingers.

Our creation, light-programming, is particularly spectacular and entertaining, and it requires manual operations, therefore the children do them with great affinity. This curiosity is observable all the way up to adult age test subjects.

We test our system from 4-year-old subjects up to university students successfully. We assume, that with the progress of age and learning, we can separate the inherited and acquired components of algorithmization abilities.

III. Age independent problem solving

The first system, based on light and sound, was "Simon says" in 1978 coming from the idea of Ralph H. Baer and Howard J. Morrison, and was written in Assembly language [9]. With its success, it became the pop symbol of the 1980s. "Do what Simon says: follow the pattern of lights and sounds for as long as you can… if you can!". The player has to copy the order of flashing lights for longer and longer. However, Simon only tests the memory, but it's worth mentioning that it also helps developing it. This game was the base for Cogmed's brain stimulation program package, which is successfully used to treat children with ADHD (Attention Deficit Hyperactivity Disorder) [10]. However, we had to look past memory tests and focus on examining creative problem solving.

We wanted to create a test system that can measure algorithm creating abilities in early childhood, but can also be used on adult subjects. By being able to follow the development of our subjects and analyse their inherited and acquired abilities, new perspectives are opened in terms of better didactic understanding of algorithm creating abilities.

Children at the age of kindergarten/elementary school cannot read or write at the levels that are needed for programming. We cannot possibly expect them to develop programs, but the aptitude to think algorithmically is given in many of them. Turning colourful sources of light on and off is possible without the ability to read or write, utilising buttons. If they do this by a logically built sequence, they achieve the same as if they wrote a program for it. Only the necessary hardware has to be created, that can store the various keystroke patterns and can play them back. The keystroke patterns that control the LEDs used as the light source can be considered the programs, and the playback can be considered running the programs. The task is quite spectacular, amusing and playful and is a challenge throughout the experiment's whole age interval. The professional literature considers playful digital education systems to be highly effective [8].

IV. Binary thinking

An important area of IT education is learning about digital technology. Here we work mainly with binary values, which creates an even greater diversity between students in terms of understanding. On higher levels of education, the goal is to create programs that can either produce a binary combination or process binary input signals. The binary signal system can be directly interpreted as the LEDs' on and off states. The goal is to make the subjects be able to break down the path to the solution into smaller steps, ergo creating an algorithm. Although high level programming requires much more complex thinking, the end result is the same, which is to make the LEDs light up in a given order (helping to understand binary 1 / 0). The dataflow visualised this way can indicate clearly that the examined person is on what level of ability of solving the task. On higher levels of education, the student creates the code that is responsible for the needed output, by interpreting the combinations in decimal. This way, light-programming shows a close analogy with the didactic area of problems of digital technology.

V. Hardware

The system consists of two parts. A control interface, where the user can control and check the tasks as well as read the results, and a measuring side, which belongs to

the examined person and where he/she can create, store and play back a light combination by his/her liking.

The measuring side has 8 LEDs and 8 push buttons. Each button controls the on and off state of the corresponding LED to the subject's liking, with unlimited retry count. The buttons are used to create the binary combinations that can be stored in the device's memory. The topmost button is used to run the stored combinations in the order of saving, repeatedly, i.e. to "hand in" the tasks.

The control interface's electronic circuits are not hidden on purpose. Our goal was to use non-virtual machines, to create the feeling of being used to seeing and touching the equipment (Fig. 1).

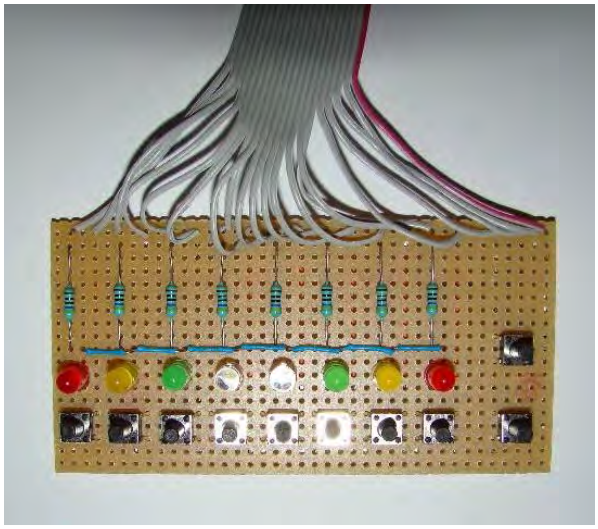


Figure 1. The tester's measuring interface.

The control interface is based on a BS2SX microcontroller, with 16 I/O ports and a RISC processor. The multiplication of peripherals is done by a 2-way, digital, I2C communication based MCP23016 demultiplexer. We used a 0.01s accuracy, I2C bus PCF8583 RTC for the evaluation data and task timing. The display is a 2x16 resolution, character LCD panel (Fig. 2).

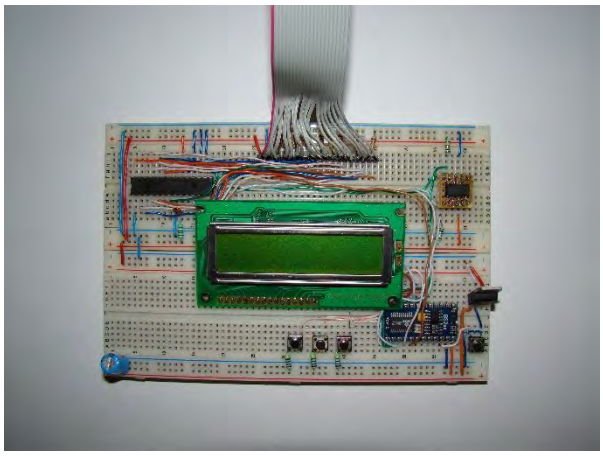


Figure 2. breadboard version, the structure of the system

Finally, it received a cover. With outer and inner power supply and data port towards PC (Fig. 3).

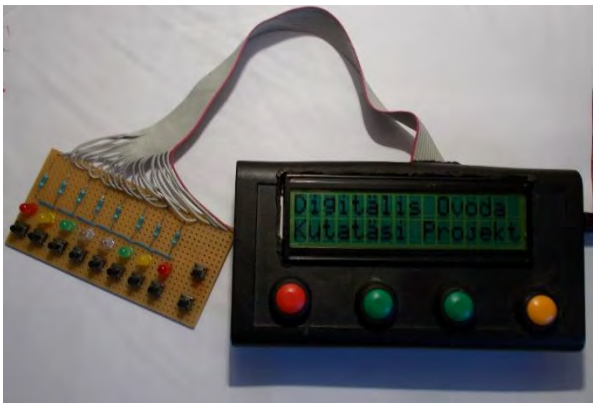


Figure 3. The system with the cover on.

VI. MEASURING

The first step is to get the subject to know his/her opportunities. He/she gets a set of instructions in a given time frame on operating the tester: "The buttons under the coloured lights can be used to turn the corresponding light on or off. If they are lit the way you desire, you have to push the rightmost button to store it in the computer. The stored light patterns can be played back with the top button, like a film." After this, the person conducting the test shows the operation with a simple example.

A. Pattern copying

The system shows a static pattern for a given time, then the presentation stops. Then the measurement starts, and the subject has to copy the pattern they saw before. As the tasks progress, the patterns get more complex. The duration of the presentation can be adjusted, e.g. made shorter, thus making the task much harder.

For example, the correct combination presented by the computer:



If the subject copies the pattern correctly, it can be considered as successful. The following attributes can be measured: the number of errors in the solution, the number of wrong attempts ("handing in" an incorrect solution), the number of all keystrokes, and the time it took to get to the solution.

An example of an incorrect solution:



The "distance" from the correct solution can also be calculated (in the above example it's 1). Another, more symmetrical example:



In this case, the following incorrect solutions are common:



In both cases, the distance from the correct solution is 2. Other attributes include:

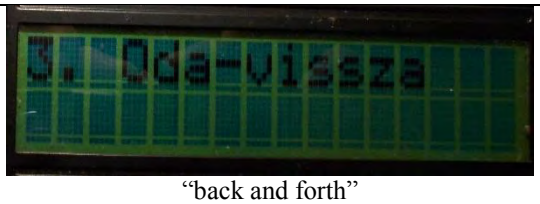
Distance from correct solution	2
Time elapsed	5.2 sec
Attempts	3
Number of keystrokes	9
Keystrokes required for solution	4

This test is meant to examine the capabilities of recreating and giving back the given pattern. Only 8 LEDs' lights have to be remembered, therefore the test is almost non-reliant on the person's visual memorizing capabilities. However, the pattern-copying ability can be well tested.

B. Pattern following

The system presents a time-varying light pattern. The task of the subjects is to break the pattern into various parts and program it. In other words, create an algorithm!

The first pattern is a simple run of lights from one side to the other. Children aged 4-6 can solve this difficulty level with a good rate. After this, the difficulty is gradually increased until the 20-30 step, complex variations. For example, two light streams that start on both sides, meet in the center and then go to the other ends, or when they meet in the center they ricochet off of each other. They can also have a "plume" consisting of a few lights, they can start from various points, cancel each other out, etc. The number of tests can reach the order of 100s. (see Fig. 4).

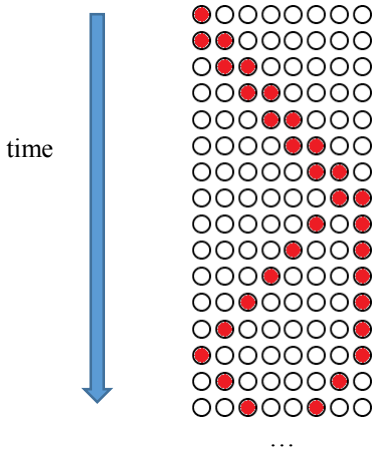


"back and forth"

00000001 = 1	11000000 = 192
00000011 = 3	10000000 = 128
00000110 = 6	11000000 = 192
00001100 = 12	01100000 = 96
00011000 = 24	00110000 = 48
00110000 = 48	00011000 = 24
01100000 = 96	…

Figure 4. Decimal coding of the binary combinations

Of course, the more complex the patterns get, the easier it gets to make an error. The subjects usually choose a wider or narrower light lane, wrong position or they don't break the task into smaller steps correctly. The almost unlimited difficulty setting separates the subjects' solving performance by age. While the simpler ones can be solved by kindergarten students, the difficult ones are a challenge even for university students.



…

The simpler tasks can be tested on the whole age spectrum, which can give quite interesting test results. While we track a person as he/she ages, we can not only examine how he/she can solve more and more difficult tasks, but also if they can solve a previous problem better than they did the last time.

Just like with the pattern copying tests, we can objectively measure the number of errors in a solution, the number of attempts, keystrokes, and the elapsed time.

VII. Evaluation

The data is stored in the microcontroller's own memory. It can be read through an SPI bus and can be processed on a PC. This provides many different ways of statistic evaluation. The light-programming tester provides many, varying and exceptionally accurate, automatically stored data. Apart from the different step and error numbers, timestamped data is also available with 0.01s accuracy.

The results are compared to the psychological test conducted on the same subjects. With young children, we examine the relevance of school maturity, motility, aggressiveness and creativity tests, while with the older generations we examine the relevance of various intelligence tests. We want to compare these tests' results' correlation to our very own test system's measurements.

The research process is currently in the stage of software development, as well as constructing a pedagogic and psychological evaluation model.

References

[1] Bakonyi Anna: Az óvodás gyermek fejlődésének nyomon követése, a kompetencia alapú óvodai programcsomag gyermekek megfigyelésén alapuló mérési és értékelési rendszere, Educatio Társadalmi Szolgáltató Nonprofit Kft (2009).

[2] Cláudio Amorim: Beyond Algorithmic Thinking: An Old New Challenge for Science Edu-cation. Eighth International History, Philosophy, Sociology & Science Teaching Confer-ence, July 15 to July 18, 2005, University of Leeds, England.

[3] Eilers, Merry L.: Older adults and computer education: "Not to have the world a closed door.", International Journal of Technology & Aging, Vol 2(1), (1989), 56-76.

[4] Geda Gábor, Hernyák Zoltán: Algoritmizálás és adatmodellek, Educatio Kht., (2011)

[5] Gelman, R., Gallistel C.R. The Child's understanding of number. Harward University Press, 1978.

[6] James A. Kulik, Chen-Lin C. Kulik, Robert L. Bangert-Drowns: Effectiveness of computer-based education in elementary schools, Computers in Human Behavior V. 1, I. 1, (1985), P. 59–74

[7] Juraj Hromkovic: Algorithmic Adventures – From Knowledge to Magic (2009), Springer.

[8] Marina Papastergiou: Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation, Computers & Education V. 52, I. 1, (2009), P. 1–12

[9] Ralph H. Baer and Howard J. Morrison: Simon Says, Marvin Glass & Associates (1980), US4207087(A) http://worldwide.espacenet.com/publicationDetails/biblio?CC=US&NR=4207087&KC=&FT=E&locale=en_EP

[10] Toni Clarke: Brain fitness seen as hot industry of the future (2008), Reuters, http://www.reuters.com/article/2008/03/12/us-brainfitness-idUSN1218668920080312?feedType=RSS&feedName=healthNews&pageNumber=1&virtualBrandChannel=0

[11] Wynn, K. (2000). Findings of Addition and Subtraction in Infants Are Robust and Consistent: Reply to Wakeley, Rivera, and Langer. Child Development, Vol. 71, No. 6 (Nov. - Dec., 2000), pp. 1535-1536

The Digitalization Push in Universities

H. Jaakkola*, H. Aramo-Immonen *, J. Henno** and J. Mäkelä***
* Tampere University of Technology, Pori Department, Pori, Finland
** Tallinn University of Technology, Tallinn, Estonia
*** University of Lapland, Rovaniemi, Finland
e-mail: hannu.jaakkola@tut.fi; heli.aramo-immonen@tut.fi; jaak.henno@ttu.ee; jukka.makela@ulapland.fi

Abstract - Digitalization is spreading across our current society and is a major source of the changes - incremental and also radical - encountered in daily life. It changes the content of jobs: some jobs are no longer needed, and new jobs are also needed. Digitalization in universities has several implications - in teaching, research, and in administration. The opportunities provided by digitalization are manifold. In our paper, we separate two digitalization related topic areas - i.e., networking and digitalization itself. Our paper focuses on academic publishing and we will analyze its role in the current academic world. The approach selected emphasizes the role of an individual researcher and, in particular, publishing the results of their work in academic forums. Digitalization provides a wide variety of new forums but also the means to recognize the impact of the forums and affect it. Networking as a phenomenon is also based on digitalization - physical networks and social networking related applications can be beneficial for researchers. Because teaching is the most visible work of universities, the role of digitalization in this sector is also addressed briefly in this paper.

I. INTRODUCTION

One of the main tasks of universities is to conduct scientific research and transfer its findings to higher education. The digital era has provided a means to measure and control research work. As a consequence, publishing has become a part of the funding instruments of universities. In our other paper [8], this aspect has already been discussed. It covers the storyline from publishing based on academic freedom to the current state, which can be described as externally controlled market driven publishing.

Digitalization provides a means for measurement – the data collected is organized and access to it is provided by a variety of information systems and network solutions. In addition, digitalization has an important role in changing academic practices. In this paper, we handle two of these – the role of digitalization in teaching and publishing. The research problem of this paper – the consequences of digitalization in universities - covers two research questions:

1. How is digitalization seen in daily university work?
2. How can an individual researcher benefit from digitalization?

Answering the research questions also dictates the structure of this paper. The approach derives from experiences from our own context (Finland, Estonia) – but may easily be generalized and used as a benchmarking reference in other contexts.

In addition to handling digitalization on a general level, we will also focus on the role of measurement in universities. The productivity of universities is measured both by internal and external indicators, which are used to guide the activities to fulfill the goals set by the university itself and by external interest groups (like government). Measurements (= getting information from the source) always have the control aspect as feedback for adapting the processes (= applying the measurement results) to fulfill pre-set goals. This topic has been handled in our paper [8] on a general level. In here, we will concentrate on the impact-related aspects, especially in the research and publication context. What impacts are measured? This covers several aspects in our study context: the impact of a single research paper, the impact of a researcher, the impact of an organization, and the impact of a publication forum.

The first research question is divided into two parts – one covering digitalization-based teaching and learning solutions (Section 2) and the other covering research-related aspects. The latter handles digital publishing (Section 3) and then impact creation through academic publishing (Section 4). Section 5 concludes the paper and summarizes the issues addressed.

II. DIGITALIZATION IN LEARNING

A. New approaches

One of the trends faced by universities is the tendency towards "part time" studies – not in an organized manner but because a growing amount of students work simultaneously with their studies. They have limited time resources and overlapping duties. In this situation, participation in traditional class studies no longer serves a growing section of students in the best possible way. The students of the "*Internet age"* are also used to using digital services in a time -independent way – instead of TV, the Internet has provided replacements, Netflix instead of movies, email and Facebook instead of face-to-face communication; surplus type services like Periscope for online mobile video streaming, Twitter for feedback etc. are a part of everyday life for them. This has created pressure for new teaching arrangements as well. Some of these services are also beneficial as a part of traditional learning – e.g. Facebook and Twitter are widely used to enrich communication and networking in universities as an embedded part of the course implementation.

From the learning / teaching system point of view, we take no stand on whether the change is good or bad. We see it as an alternative in some cases but it is not suitable for all situations. The authors have personal experience of some learning methods implemented in our curricula:

- *blended learning*: learning with mixed methods and media – covering some of the elements like class studies, video archives of lectures, shortened video-based material of lectures, combining MOOC (Massively Open Online Courses) material into studies, online video broadcasting of lectures, etc.;
- *flipped learning*: supporting students' self-learning with a variety of material (see the bullet above) and focusing on face-to-face study resources for collaborative problem solving;
- the use of (globally available) *MOOC* as a replacement or support material for local course implementation;
- the use of *groupware facilities* - chat, online group sessions, network-supported collaborative problem solving, a variety of tools for distributed development, and several additional methods.

The primary motivation factor, in our case, is to support the student's opportunity for independent (place and time independent) learning. This is done keeping in mind the conditions of those participants who do not have time to participate in classes, and also to provide all participants with an opportunity to repeat learning items. A smaller attendance makes a more efficient use of time possible, but experiences of the quality of learning vary. Our experiences have been reported by Linna et al. in their conference papers [12;13].

The adoption of new learning methods has two sides – the first (studies, students' needs) was handled above – but what about the lecturers? Following the discussion that has been ongoing for the last few years, there seems to a *respectful attitude to difference (new)* and a low quality image connected to the traditional (class-driven study method) – the less contact you have face-to-face with your students, the higher the quality of their studies. This attitude has roots in two sources: *higher productivity* and *"political" pressure*. The productivity increase comes only from the opportunity to reach larger masses of students.

The changes in the time usage profile can be estimated according to the following figures. In *traditional classes,* in the case of a new first implementation (experience of the first author), one class hour requires one working day to prepare the material and course implementation principles (1:6), and preparing for a four-hour study module needs two hours (4:2). The second implementation halves these times (1:3; 4:1). The third implementation does not need any significant time to prepare material, but preparation time for the classes remains the same. In connection with the new learning methods (flipped learning, blended learning supported by video material) the time required increases dramatically, as well as the work demanded of the teacher in evaluating material returned by the students. A rough estimation provides the following multiplying factors: in preparing a new course (material) 2-5 (new course implementation 1:12..30), in preparing a single learning activity 2 (4:4), in the time used for a single learning activity (class) 0.5-0, and in the after (class) activity 2-5. Additional activities which are difficult to count cover tutoring, which depends greatly on the number of students.

The role of lecturer is also changing - the teacher is becoming a mentor. The truth is that in new study modes, the teacher's capacity decreases easily by half (or even more) of the original (in the number of manageable courses); the throughput of students will increase, of course, but only in the case of large masses of students. One additional problem related to very technology-oriented teaching situations is that technology (related aspects) steals some of the students' observation opportunities and teachers' direct contact with the class. The second aspect mentioned above – political pressure – will be ignored here; its source is in most cases somewhere with *no deep understanding* of the primary tasks conducted by the teachers and researchers. If they had this understanding, they would leave these issues to the responsibility of the experts, i,e., the teaching organizations).

B. Massively Open Online Courses

We started our own contribution to the discussion on MOOCs two years ago in our conference paper [7] in the context of open data. In the following, we will briefly return to this issue. Additional aspects cover reported experiences on MOOC-based learning.

At the moment, MOOCs are seen as an opportunity that has great potential and some drawbacks. Our experiences are based on embedding external MOOC content in our own teaching, and participation on a small scale in some MOOC courses. The leading organizers of MOOC are among the world's leading universities, which in this format provide open access to the lectures of top world-level experts. There seems to be a wide variety of motivation factors for participating in MOOCs, which in practice host in some cases tens or even hundreds of thousands of participants. The most commonly known and globally referred to MOOC platforms are Coursera, Udacity, EdX, and Khan Academy.

MOOCs provide one alternative way to organize studies. The courses available can be used in several ways, either directly as a part of the curriculum or embedded in individual courses. In all cases, MOOC studies also need effort from the user organization – availability of mentoring and tutoring is usually expected on local level. In spite of several positive MOOC experiences, there are also some negative aspects related to them. The first – not discussed here - is globalized (outsourced) learning and its consequences in the improvement of local skills. The second is the high percentage of dropouts.

Katy Jordan is an e-learning expert who has collected data concerning MOOC dropouts (in [9]). Her findings are reported in Figure 1. The enrollment in the courses studied varies up to 230,000, with 25,000 a typical size. The chart in the original source [1] is interactive and provides

[1] http://www.katyjordan.com/MOOCproject.html

additional information for the reader. Completion rates can approach 40%, although the current average completion rate for MOOCs is approximately 15%.

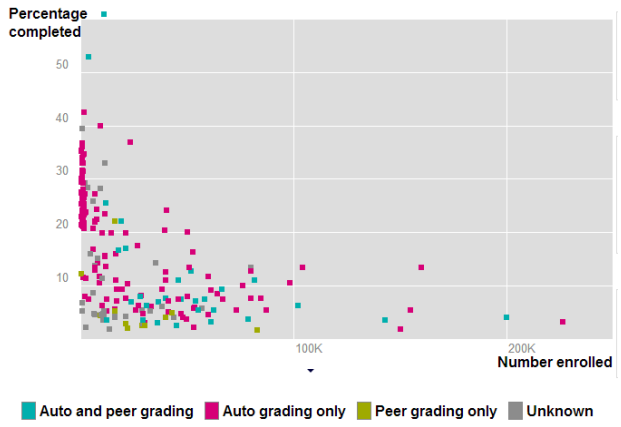


Figure 1. Completion rates (%) and Assessment type (different colors)

The study [16] reports the main reasons for dropouts: no real intention to complete; lack of time; course difficulty and lack of support; lack of digital skills or learning skills; bad experiences (a variety of bad experiences being a barrier to continuation); expectations (students may enroll with little understanding of what the course requires and may have unrealistic expectations either of the course or of their own ability to undertake it); starting late (late starters on a course may find it very difficult to catch up and outcomes are likely to be much lower); peer review (courses relying on peer grading often have much lower completion rates than others). They also noticed that the participants did not follow the original course schedule strictly: many participants were a long way behind in the course but were still working steadily to continue it at their own pace. The study of Kloft et al. ([11]) has focused in predicting dropouts in MOOC courses as a function of time. The finding of the empirical study is that the number of active users decreases in time (highest percentage in the second half). They also measured the interaction level (number of requests - page views and video clip actions), which remains reasonably stable over the course period.

One additional MOOC finding in Internet discussions refers to *MOOC Piracy*. The bloggists Amanda Connolly and Rob Conery[2] report on the case of a security expert, Troy Hunt, who has produced video-based learning material on "Ethical Hacking." Someone else ripped his video and re-uploaded it on Udemy, charging £37 ($55) to watch it. This is one more new phenomenon in the area of digitalized teaching. However, even in a traditional teaching context, this is possible. The (first) author has found his teaching material copied and included in published books as well as hearing about direct use of it by other lecturers without permission.

[2] https://medium.com/@robconery/how-udemy-is-profiting-from-piracy-5638b929ffca#.xpdq6hei3 (retrieved on February 15th, 2016).

III. DIGITALIZED PUBLISHING

Publishing has moved and is moving rapidly to digital platforms. As a consequence, the availability of publications has become easier. The vast majority of scientific publications are covered by traditional search machines and libraries of research organizations have agreements between publishers related to accessing their materials. For instance, the *Scienceport portal* of TUT covers access to over 280 scientific databases, among which are[3] IEEEXplore, the ACM Digital Library, and Springerlink (as an example of important sources for computing area researchers).

From an individual researcher's point of view, there is also an increase in publishing alternatives. Instead or in addition to the traditional printed format, in many cases there is an opportunity for digital publication to allow higher (network-based) visibility and availability for their contributions (with or without a separate publication fee; the latter usually opens the full article for free downloading). From the publisher's point of view, publication and distribution costs decrease dramatically, if no paper versions are needed. A journal having both a printed and digital version of a publication is called a *hybrid journal*.

A new category of publications is *open access journals*. They follow traditional (automated) paper submission procedures, including peer reviews and academic quality control principles that are apparently taken seriously. The problem, as in any other new publisher case, is creating trust. New publishing forums face challenges in reaching an acceptable *trust (reliability) level* as a first step and then, as a consequence, such an impact level that it can be used for academic contributions. From a researcher's point of view, open access journals are ideal: short publishing delay, journal status, low costs, no need to cover travelling costs for conferences.

DOAJ[4] maintains a list of high quality, open access, peer-reviewed journals. At the moment, it covers 3,100 publications, both pure open access and hybrid journals. The article on the history of DOAJ by Olijhoek et al. ([15]) reports growth in high quality open access journals from 1,000 to over 10,000 in the period 2004-2015 (tenfold; from 5,000 to 10,000 in the period 2010-2014).

Beall's list of predatory publishers [5] covers questionable scholarly open-access publishers. "Predatory" in this context covers different arguments ([19]; author's own analysis): Accepting articles quickly with little or no peer review or quality control, including hoax and nonsensical papers; Notifying academics of article fees only after papers are accepted, paid for and accepted papers not published; Aggressive campaigning for academics to submit articles or serve on editorial boards; Listing academics as members of editorial boards without their permission, and not allowing academics to resign from editorial boards; Appointing fake academics

[3] IEEEXplore http://ieeexplore.ieee.org/Xplore/home.jsp; http://dl.acm.org/dl.cfm2531801; http://link.springer.com/

[4] Directory of Open Access Journal – DOAJ is available on https://doaj.org/.

[5] Scholarly Open Access, available on https://scholarlyoa.com/publishers/

to editorial boards; Mimicking the name or website style of more established journals; Misleading claims about the publishing operation, such as a false location; Improper use of ISSN; Fake (e.g. use of ISI) or non-existent impact factors and use of own impact factor specifications.

Shen and Björk in [18] have studied the negative consequences of the rapid growth of scholarly predatory open access publishing funded by article processing charges. Their report includes a trend study that illustrates clearly the fast growth of this publication category (Figure 2)

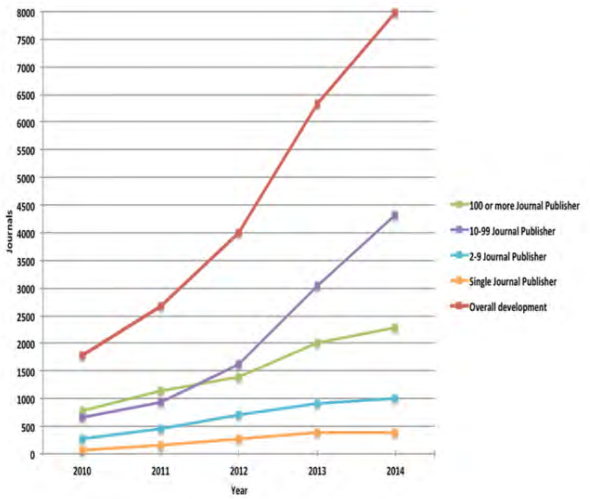


Figure 2. The development of active predatory open access journals from 2010 to 2014

Figure 2 reports the growth of open access journals from 1,700 in 2010 to 8,000 in 2015 (fivefold increase); in the same period the number of published articles increased from 53,000 to 420,000 in a reasonably linear process. Another interesting trend is seen in the number of published articles per journal: overall, it has remained quite stable (close to 50), but in the single journal publisher category it has risen from 70 to 270 in the period analyzed.

The study also reports an interesting journal case - *Experimental & Clinical Cardiology*: "The journal had for 17 years been published by a respectable Canadian subscription publisher. The journal, which had a JCR impact factor (0.7), was purchased by investors of obscure background, changed the business model to OA, funded by an APC of 1,200 USD, and rapidly increased the number of articles from 63 in 2013 to over 1,000 in 2014." This is an example of the value of trust – the impact factor remains (at least for a while) on the old level in spite of the changing publication policy.

According to Shen and Björk in [18], these journals are targeting their marketing efforts primarily to authors from developing countries (India – 34.7%, rest of Asia 25.6%, Africa – 16.7%), who are eager to be published in "international" journals. The distribution of publishing companies shows similar figures (India 27%, Africa, Asia (other than India) – 11.6%, Africa – 5.5%, impossible to determine 26.8%). The fake journal business took about $75 million in 2015 and published nearly half a million articles.

As discussed above, publishing is big business. To get a paper published, a publishing fee must be paid. This especially fits situations where the author buys open free access to the papers. If this is not paid, the access covers only extracts of the paper and downloading the full version requires money. In some cases, the fees for downloading are high and do not suit the budget of a researcher. Author-driven distribution using different platforms (ResearchGate, Academia.edu; see 4.3) or the portals of one's own research organization (if permitted by the publisher) improves the availability. The blogger Simon Oxenham has called his text "Meet the Robin Hood of Science." The same topic is handled by Fiona MacDonald under the title "Researcher illegally shares millions of science papers free online to spread knowledge - Welcome to the Pirate Bay of science."[6] The Sci-Hub site (http://sci-hub.io/) was established by Alexandra Elbakyan, a researcher from Kazakhstan, in 2011. The service notices the restrictions based on copyright laws, but points out that these are simultaneously barriers that slow down the development of science in human society. As a result, it provides a service that effectively bypasses any paywalls and restrictions. At the moment it has free access to more than 47,000,000 scientific articles. Without continuing the discussion in this context, we are content to note that this is also one way to get publicity for scientific papers.

IV. IMPACT IMPROVEMENT

A. *Referencing and Cross-referencing*

There are two important concepts related to scientific publishing – visibility and impact. One of the main purposes of academic research is to subject new information to *public criticism – visibility* serves this goal. In addition to visibility, an important aspect is the *impact* of the work. Impact (in current academic society) has two components – the impact of the publication forum measured by its *impact factor*, and the *impact of the individual work* (measured by the references to it). The *impact of an individual researcher* in turn is the resulting impact profile of his/her publications. These three factors are bound to each other: the high impact factor of a publication forum is the result of a high number of references to it; the high impact of individual work is the result of a large amount of references to it; high impact is easier to reach in highly ranked forums than in lower level ones; an individual researcher has an opportunity to increase his/her visibility by active impact improvement actions; a growing number of citations may also affect the growing impact of a publication forum. We will address this topic from different angles in the following paragraphs.

It may be worthwhile refreshing Karl Popper's original idea from 1930 about the need for falsification in academia in order to prove the truth of a scientific argument [17; 20]. Because scientific publication is to an increasing extent based on cooperation in networks, it is

[6] Simon Oxenham: http://bigthink.com/neurobonkers/a-pirate-bay-for-science Retrieved on Feb. 28th, 2016); Fiona MacDonald: http://www.sciencealert.com/this-woman-has-illegally-uploaded-millions-of-journal-articles-in-an-attempt-to-open-up-science (Retrieved in Feb. 28th, 2016).

natural that the members of the network utilize each other's studies as a part of their own work. This further leads to the growth of cross-referencing inside the network. Furthermore, this should be distinguished from the contemporary culture of "likes" (meaning references to a friend's papers) in discussion in journals for example. The tendency to generate references among contemporary communities of practice is a growing phenomenon, resulting from the evaluation measures introduced during the last decades (impact factors and the h-index for example). In the following, we discuss the most commonly known *author-level* and *journal-level* metrics.

B. Metrics in Scientific Publishing

The *h-index* (Figure 3) is an *author-level metric* that attempts to measure both the productivity and citation impact of the publications of a researcher (Hirsch, [6]). There are different versions of the h-index, like the i10-index representing the number of publications with at least 10 citations. This very simple measure is used by Google Scholar, and is another way to help gauge the productivity of a researcher [4].

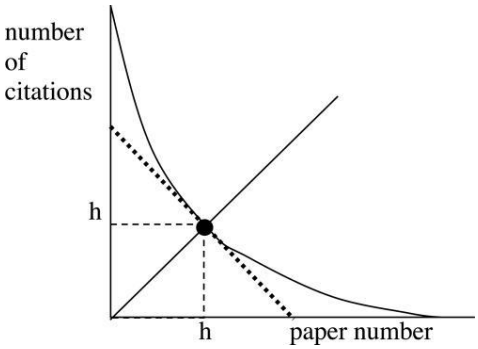


Figure 3. Schematic curve of number of citations versus paper number, with papers numbered in order of decreasing citations (Hirsch, [6])

There are some well-known disadvantages regarding the h-index and yet it has become a de facto standard in measuring researchers. Firstly, it does not give an accurate measurement for early-career researchers. Secondly, it is calculated by using only articles that are *indexed* in the Web of Science (WoS). If a researcher publishes an article in a journal that is not indexed by the WoS, the article as well as any citations to it will not be included in the h-Index calculation [4]. Among early-career researchers, new ways of affecting the h-index have emerged by utilizing social media. The more you advertise your article in social networks (Twitter, Facebook), open publishing platforms (SlideShare), or in researcher networks (ResearchGate, Academi), the more readers you gain and the more citations you receive from your community of practice (in other words, from your *followers on social media*). This is an effective way to grow your h-index [1; 10].

When following article downloads and article citations in ACM, it could be observed that there is a positive correlation between the numbers of downloads and the numbers of citations (Table 1). When studying the citations of one researcher *before social media marketing and after one-year intensive social media promotion,* 122% growth was found, rising from 20 citations to 67 citations [10].

TABLE I. CORRELATION BETWEEN ARTICLE DOWNLOADS AND ARTICLE CITATIONS IN ACM (CUMULATIVE FIGURES

	ACM article downloads		ACM article citations	
YEAR	Article A	Article B	Article A	Article B
2014	23	0	0	0
2013	135	112	2	2
2012	561	426	10	8
2011	3020	1245	33	12
2010	5298	2278	65	32

Commonly used *journal-level metrics* are: Source Normalized Impact per Paper (SNIP), Impact per Paper (IPP), Impact Factor (ISI), and 5-Year Impact Factor. SNIP measures contextual citation impact by weighting citations based on the total number of citations in a subject field [14]. IPP is the average number of citations received in a particular year by papers published in the journal during the three preceding years (Elsevier, [5]). The Journal Impact Factor (ISI) is published each year by Thomson Reuters. It is a measurement of the number of times an average paper in a particular journal is cited during the preceding two years (Elsevier, [5]). Similarly, the 5-Year Impact Factor is based on 5 years of history.

There are several ways of using these metrics to evaluate research and the researcher. The most common are ranking lists based on journal-level impact factors. There are international ranking lists for universities and journals and conferences. These ranking lists guide a researcher on which university to apply to, which journal to publish in and what conference arena to use to introduce your research. An example of a widely used international journal guide is the UK-based The Association of Business Schools (ABS) Journal Guide [3]. ABS has received permission from Thomson Reuters to use their JCR data, and permission from Elsevier to use their SNIP, as well as permission for SJR metrics powered by Scopus. Simultaneously, there is a growing number of emerging national ranking lists like the Publication Forum (Julkaisufoorumi, JUFO[7]) in Finland. The problem is that these international and national lists overlap and use very contradictory measurements/metrics.

There is also criticism among researchers about the guiding direction of such rankings. It seems to be much harder to get a qualitative study into a highly scored ranking position compared to quantitative studies: "The newly published ABS Journal Guide is a travesty. It entirely undermines a century (and more) of European scholarship that has sought to challenge the positivistic orthodoxy in social research. We have mainly succeeded in that aim with innovative theory building and ample empirical studies that offer insight into what real organizations look and feel like – not what a model tells us they should look and feel like. In an era where our research must feed into the needs of society and where, as the deliverers of business education, we need to develop responsible and reflective leaders of the future and

[7] JUFO introduced: http://www.julkaisufoorumi.fi/en; Publication forum search: https://www.tsv.fi/julkaisufoorumi/haku.php?lang=en.

encourage innovation and creativity, not a regurgitation of the same orthodoxies that please the elite."?? Referring again to the ABS (2015) ranking list, it seems that therefore the new truly innovative inventions and novel scientific thinking is absent from top ranked journals: "…this is simply a list of US journals. In my sub-disciplines, in the journals ranked, the work is conservative, narrow, and parochial to the United States." [8]

Finally, it is a common pitfall to think that these author-level and journal-level metrics are measures of scientific quality. Journal-level metrics are by definition impact factors. In other words, they do not take any responsibility for the actual quality of the research but merely indicate the popularity of subjects and methods among scholars. Recent research has shown how vulnerable the author-level h-index is. It is possible to simply boost your h-index by marketing your research in the social media. There is nothing wrong in this, but the failure is to confuse these metrics with the actual value or quality of scientific research.

C. Digitalization in Impact Creation

Academic society has found different means to increase visibility. In addition to the role of social networks (Facebook, Twitter, …) to boost visibility improvement, discussed above, there are several digital platforms supporting researcher networking. The most commonly known are ResearchGate and Academia.edu. Both have the same basic idea - researchers can submit their publications to these forums and the forum provides visibility among the platform users (and also on a more general level).

ResearchGate was founded in 2008 to serve as a social networking site for scientists and researchers. It supports sharing of papers, collaboration activities between researchers, and finding collaborators. It is the collaboration network for scientists, research professionals and affiliated people, having 8 million users [9]. ResearchGate, in addition to managing the information related to registered users, is also known for aggressive marketing to improve their network growth. It automatically emails unsolicited invitations to the co-authors of its users; these messages are written as if personally sent by the user. ResearchGate provides a platform for storing publications in pdf form as a goal to increase their availability and visibility. To increase coverage of the publications, ResearchGate uses a crawler to find PDF versions of articles (from authors' and publishers' home pages). These papers are presented as if they had been uploaded by the author; one more questionable feature of the system. The members have an opportunity to manage these automatic features through settings. A study published by the Association for Information Systems found that a ResearchGate account using default settings generated 297 invitations to 38 people over a 16-month period, and that the user profile was automatically attributed to more than 430 publications.

Academia.edu is a social networking website for academics launched in 2008. The platform can be used to share papers, monitor their impact, and follow the research in a particular field. It has 32 million users, manages 9.7 million uploaded papers and has 1.8 million research interests registered. It has over 36 million visitors a month. [10]

The Academia.edu team has published a study [11] related to the increase in citations of the papers posted on their platform. The study finds a 73% citation increase associated with articles posted on Academia.edu. The authors believe that finding objective studies in this field is difficult. However, the results above are indicative and in agreement with our own subjective findings: citations grew from 20 to 40 in a short time after registering the paper (and researcher) on ResearchGate.

V. CONCLUSION

Digitalization is a general phenomenon that affects society as a whole – even in universities and other academic institutions. The original idea of this paper was to concentrate on publishing and impact analysis in academic work, but finally we extended our work to cover other aspects of digitalization. Digitalized learning provides new approaches for organizing university studies – for example, the MOOC case in a global context and the massive amounts of participants. Local level applications are available as well, but because of the large amount of development work it has not spread so fast. Blended learning with elements of MOOC implementation principles has, in contrast, become a part of daily culture.

Electronic publishing has started a new era in publishing culture. The wide availability and fast growth in the number of open access digital publications also demands a new critical attitude from researchers. Predatory publishing ("pay and publish") does not respect the traditions of academic ethics. An individual researcher has to be more and more careful when submitting papers to the forums that are available.

Researcher networking has provided a new means for the researcher. Active researchers are networked. They use network-based platforms to increase their visibility and indirectly (via the growth of citations) their impact. The effect on impact culture is available for all researchers. In addition, to be good at science a researcher is also expected to be good at and active in marketing.

Figure 4 illustrates two loops: the *impact growth loop* (gray area) indicates the importance of an individual researcher in increasing his visibility (impact). It is

[8] The Times Higher Education Website February 26, 2015. Available on https://www.timeshighereducation.com/news/abs-ranking-reveals-world-elite-of-business-journals/2018722.article. Retrieved on Feb. 11th, 2016.

[9] https://en.wikipedia.org/wiki/ResearchGate (Retrieved on February 15th, 2016. Homepage: https://www.researchgate.net/ .

[10] https://en.wikipedia.org/wiki/Academia.edu (Retrieved on February 15th, 2016). Homepage: https://www.academia.edu/ .

[11] Reported in https://medium.com/@academia/academia-edu-citations-and-open-science-in-action-4a24a6376573#.tw8vjo50k (Retrieved on February 15th, 2016. The full study is available in https://www.academia.edu/12297791/Open_Access_Meets_Discoverability_Citations_to_Articles_Posted_to_Academia.edu

discussed in more detail in [8]. The upper part (thick arrows) indicates the *impact improvement loop*. To pour oil on the fire, we can say that first the researcher can increase his visibility – i.e. *personal impact*. In addition, indirectly the person level impact reflects the impact factor of the journal because of the growth of citations. The growth of the impact factor of the journal makes it more visible and further grows its impact factor. If we think of this as a networking phenomenon, the growth potential is exponential.

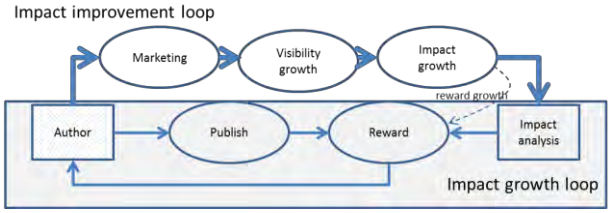


Figure 4. Impact growth loop

Impact factors (in different levels) are a measurable indicator that is also criticized. The article of Belter ([2]) separates the different targets of *impact measures*: research paper (=number of citations in it), publication forum (the number of citations in relation to all articles in a certain two-year period), researcher (different metrics, e.g. h-index; the key factor is a comparable value between researchers), organization, country, etc. On the journal level, the impact factor only covers a small fraction of its published articles. On the individual level, young and experienced researchers are in a different position, because there is a delay before publications are visible in indicators. There is also a big difference between disciplines. One more aspect worth noticing is that the rapid growth in the amount of publication forums makes longitudinal comparisons difficult.

We will close this discussion with a collection of statements from ”private thinking”: A high number of citations may indicate the positive importance of the paper, but a negative factor may also be an alternative interpretation – errors corrected by other authors are a reference as a bad example. A good study published by a low level forum does not become bad. In contrast, a bad study published by a high level forum (if possible because of a review) does not make it good. A good study that is not found and visible is useless. However, a bad paper that is useful for somebody becomes good. A well-known author may have privileges to use the fast lane. The final conclusion of the authors is that publication is no simple matter.

REFERENCES

[1] H. Aramo-Immonen, J. Jussila, J., & J. Huhtamäki, “Exploring co-learning behavior of conference participants with visual network analysis of Twitter data”, Computers in Human Behavior, 51(Part B), 1154–1162, 2015. doi: 10.1016/j.chb.2015.02.033.

[2] C.W. Belter, “Bibliometric indicators: opportunities and limits”, Journal of Medical Library Association 103(4): 219–221, 2015. doi: 10.3163/1536-5050.103.4.014. Available in http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4613388/. Retrieved on February 15th, 2016.

[3] Cabs, Academic Journal Guide 2015. Chartered Association of Business Schools. Available in http://charteredabs.org/academic-journal-guide-2015/. Retrieved in Bebruary 6th, 2016.

[4] Cornell, Measuring your Research impact, Cornell University Library, 2016 Available in http://guides.library.cornell.edu/c.php?g=32272&p=203393. Retreived on February 6th, 2016.

[5] Elsevier, Journal Metrics, 2016. Available in http://www.journalmetrics.com/ipp.php. Retrieved on February 6th, 2016. (read 6.2.2016)

[6] J.E. Hirsch, "An index to quantify an individual's scientific research output". PNAS 102 (46): 16569–16572, November 2005.

[7] H. Jaakkola, T. Mäkinen, J. Henno and J. Mäkelä, Open^n, Proceedings of the 37th International Convention MIPRO 2014, B. Biljanović (editor), Opatija, Mipro and IEEE. ISBN 978-953-233-078-6: 726-733, 2014.

[8] H. Jaakkola, J. Henno, J. Mäkelä and K. Ahonen, A Researcher Measured - Towards a Measure Driven Academy. Proceedings of the 39th International Convention MIPRO 2016, Mipro and IEEE, 2016 (submitted).

[9] K. Jordan, MOOC Completion Rates: The Data, 2016. Available in http://www.katyjordan.com/MOOCproject.html. Retrieved in February 15th, 2016.

[10] J.J. Jussila, Social Media in Science Communication, SlideShare, 2015. Available in http://www.slideshare.net/jjussila/social-media-in-science-communication. Retrieved on February 6th, 2016.

[11] M. Kloft, F. Stiehler, Z. Zheng and N. Pinkwart, Predicting MOOC Dropout over Weeks Using Machine Learning Methods, 2016. Available in https://www2.informatik.hu-berlin.de/~kloftmar/publications/emnlp_mooc.pdf. Retrieved on February 15th, 2016.

[12] P. Linna, H. Keto, and T. Mäkinen, Towards Blended Learning: A Case Study, Proceedings of the 38th International Convention MIPRO2015, Mipro and IEEE, pp. 685-689, 2015.

[13] Linna P, Keto H, and Mäkinen T, (2016). Utilizing MOOCs in the development of education and training programs, Proceedings of the 39th International Convention MIPRO 2016, Opatia, Mipro and IEEE, 2016 (submitted).

[14] H. F. Moed, H. F., “Measuring contextual citation impact of scientific journals”, Journal of Informetrics, Volume 4, Issue 3, July 2010, pp. 265–277.

[15] T. Olijhoek, D. Mitchell, L. Bjørnshauge, “Criteria for open access and publishing”, ScienceOpen Research – Section: SOR-EDU, 2015. Available in https://www.scienceopen.com/document/vid?id=85a98041-8734-4a43-b6eb-ff5903f3ae96. Retrieved on February 12th, 2016.

[16] D.F.O Onah., J. Sinclair., R.Boyatt, Dropout Rates of Massive Open Online Courses: Behavioural Patterns, 2016. Available in http://www2.warwick.ac.uk/fac/sci/dcs/people/research/csrmaj/daniel_onah_edulearn14.pdf. Retrieved on February 15th, 2016.

[17] K. Popper, The Logic of Scientific Discovery (Taylor & Francis e-Library ed.), London and New York: Routledge / Taylor & Francis e-Library, ISBN 0203994620, 2005.

[18] C. Shen, B-C. Björk, Predatory open access: a longitudinal study of article volumes and market characteristics. BMC Medicine 13:230. DOI: 10.1186/s12916-015-0469-2, 2015. Available in http://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-015-0469-2. Retrieved on February 12th, 2016.

[19] Wikipedia, Predatory open access Publishing. Available in https://en.wikipedia.org/wiki/Predatory_open_access_publishing. Retrieved on February 12th, 2016.

[20] Wikipedia, Falsifiability. Available in https://en.wikipedia.org/wiki/Falsifiability. Retrieved on February 8th, 2016.

Toby the explorer – an interactive educational game for primary school pupils

Nenad Kaevikj, Angela Kostadinovska, Biljana Stojkoska, Marija Mihova, Kire Trivodaliev
Faculty of Computer Science and Engineering (FCSE)
University "Ss. Cyril and Methodius"
Skopje, Macedonia
biljana.stojkoska@finki.ukim.mk, marija.mihova@finki.ukim.mk, kire.trivodaliev@finki.ukim.mk

***Abstract*—Technology enhanced education has been recently established as a new approach for all stages of education. Although computer games are often given little attention we believe that within the vast amount of technologies and instruments used to achieve the needed improvements it is computer games that are playing the central role in delivering the desired effects, in particular to children and teenagers. In this paper we present the "Toby the Explorer", an interactive educational game for primary school students. We show in depth the engineering behind the game, its design and structure. We also give an insight and evaluation of the importance of the game in enhancing the educational process. Our results show that the learning process is regarded as more easy and fun by the students that learned by playing, but also it increased their interest in learning other subjects not included explicitly in the game.**

Keywords—Educational games; Game-based learning; e-Learning; Desktop games; Games for kids

I. INTRODUCTION

Technology usage in the 21-st century is no longer an exception, but a rule, especially in domains close to and intertwining with computer science and engineering. People are trying to ease their lives, prominently by using a plethora of mobile applications in many, if not all, aspects of their daily life. Education and learning, as an example of the previous, has been under great influence of these technological processes, and the rise of computer-based learning applications and educational games is more evident than ever. Traditional teaching, usually based on lectures and tutorials fosters the idea of instruction-driven learning model [1] where students are passive listeners. However, the technology aided learning paradigm follow the constructivism learning theory, where learning from real-world situations is put on the first place. According to this "student centered" strategy [2][3], students collect in-deeper knowledge when they directly experience the situation. One such approach is the project based learning [4] that improves professional skills related to a particular project and specialized knowledge, provides an interdisciplinary insight to a problem, and improves the overall socio-technical skills of the students. Project based learning can be a very useful tool in learning for higher education, however it is very difficult to keep secondary, and even more so primary, school students engaged in such a process. The "fun-based" approach using games as a tool for learning is better suited for such students. Numerous fields of learning, some being: languages, mathematics, programming, science, history, geography, art and many other, are enhanced by wide range of educational games [5]. Such applications are especially useful in improving, and even enabling, the learning process in youngsters since they tend to keep children mentally active, can be used as an addition to the classroom work and as after school fun activities that extend the learning process.

Adults may make up the majority of game players, but video games are more popular than ever among kids, new research shows. According to NPD [6], 91 percent of U.S. children ages 2-17 play video games (64 million). More interesting, these numbers are up nearly 13 percent from a 2009 study. The number of kids in the U.S. has increased by 1.54 percent in that time, but not nearly enough to make up for the massive increase in game playing. Gaming among kids ages 2-5 has increased the most.

"Year-to-date through August 2011, kids comprised 44 percent of new physical software dollar sales, representing a vitally important consumer segment for the games industry," said Anita Frazier, industry analyst, The NPD Group. "Knowing how kids are spending their gaming time and dollars in both traditional and non-traditional outlets is key to staying relevant to this highly engaged audience."

Educational games are application software that helps students to learn the lesson subjects and to develop their problem solving skills by using their desire and enthusiasm to play [7]. Such games can be observed as free fun activities that provide students with the opportunity to reinforce their knowledge by repeating it in a more comfortable environment.

Prensky in [8] distinguishes between six structural factors that define an application as a game: *Rules* to impose limits which force us to take specific paths to reach goals and ensure that all the players take the same paths; *Goals or objectives* which create a sense of duty, so we play the game voluntarily and spend time and make an effort for it. *Outcomes and Feedback* that give a measure of progress towards the goals. Feedback comes when something in the game changes in response to an action a player takes; *Conflict, Competition, Challenge and Opposition* are the problems in a game that the player tries to solve. *Interaction* can be viewed both as a player – computer interaction, and an interaction amongst multiple players playing the game; and finally *Representation* which

refers to the essence of the game play, i.e. what the game stands for.

Most researchers believe that educational games have a positive impact on children and those children using educational games in the learning process have a greater chance of success as compared to children that use the traditional way of learning. Ritzhaupt at al. [9] noted several advantages of educational games over the traditional learning: children use action instead of explanation, create personal motivation and satisfaction, they are able to accommodate various learning styles and skills, they reinforce mastery, impose making decisions and collaboration with other learners. Divjak and Tomic [10] analysis of different research about the influence of educational games suggests that playing a digital computer game improves the spatial skills of the students and reduces the difference in spatial skills between boys and girls. This approach helps students in learning and improving skills for social problems solving and new/unfamiliar situations. However, it must be noted that in order to have a positive impact on students a good computer game must meet pedagogical criteria, be educational, but also interesting [11].

All of the before mentioned advantages of game based learning motivated us to create an interactive game that combines the processes of gaming and learning and produces a gain greater than the sum of the two.

Taking into account that the use of tablets and mobile phones is common for the children of today [12] and that they tend to use such mobile devices instead of desktop computers, we developed our game for both platforms. The game is designed as a three dimensional (3D) adventure game, since we believe that such a game will have greater appeal and will keep our targeted users, i.e. children, interested. The game is intended for primary school students aiming at extending their knowledge in different disciplines, like biology, geography and science.

The main advantages of "Toby the Explorer" are (but not limited to): the easy learning process, the development of the visual skills of the users, and the imperceptible transformation of fun into productive brain activity. These goals are enabled through the usage of highly visible performances and photographs that augment the learning by adding visual stimuli that accompany the written information. Additionally, the environment has a very realistic representation and is complemented with catchy music and special effects, and finally information within the game is represented in a very simple and reasonable way so it can be understood by everyone, and is amended with interesting facts.

The rest of this paper is organized as follows. The second section describes the game plot and scenario. The third section provides implementation details. Section four evaluates the game. This paper is concluded in section five.

II. GAME PLOT AND SCENARIO

The main character of this game is a little boy named Toby who is very ambitious and curious. He wants to know more about his surroundings and environment (Fig. 1). Toby lives in a small house near the lake. Near his house there is a small wood and a mountain abound with lots of animals and vegetation. There is a small village nearby and a farm where the player can learn more about the animals and nature. There are also a lot of hidden geographical spots which are harder to be found, but they discover more about some Earth phenomena. Very interesting thing is that the game has a lot of environmental sounds who make the game much more realistic. For example, if the player comes near water, splashes and stuff can be heard.



Fig. 1. Toby and his home

Since Toby is a very curious boy, he starts sniffing around the area. His mission is to find out as much as he can about the environment by collecting hidden scrolls and photos. On each of the scrolls he finds, there is a content describing the area, together with some facts about it. Photos show up very different places as well, containing graphical perspective of the area. But one thing Toby can't do is swimming, so the player must be careful to avoid deep waters. The player starts with three lives. Once they are lost, the game starts over again and the player needs to find the scrolls and photos again.

The game has a lot of detailed graphics and is made to perfection to the smallest object. We tried to compress as much as we can for the game to be easier to open on more devices. Each time the game is started, an options window pops up and prompts the player to choose graphics, controls etc. to match pc performances (Fig. 2).

The user interface of the game is very friendly and easy to navigate. It is separated on two parts:

- Main Menu User Interface
- In-game User Interface

The Main Menu User Interface consists of 4 buttons: Play, About, Scores, Quit (Fig. 3). The "Play" button starts the game and takes us to the main scene. The "About" button opens up a new popup which has a brief description of the game. The "Scores" button loads up all the scores the user has on his record. These scores are loaded from the database so an internet connection is be required for this option. The "Quit" button shuts down the game.

The In-Game User Interface is separated into four parts, and each fills a corner on the screen (Fig. 4). On the bottom left corner there is a compass which allows the player to orientate during the play. The compass has a mini-map arrow which displays the player's direction of movement. In the top left

corner there is an icon with a number on its right side which counts the lives left. Each time it hits zero, the game starts all over again. On the top right side of the screen there is a pause button used to stop the game and returns the control to the main menu. On the bottom right side of the screen there are two counters, dedicated to count the scrolls and the photos collected.

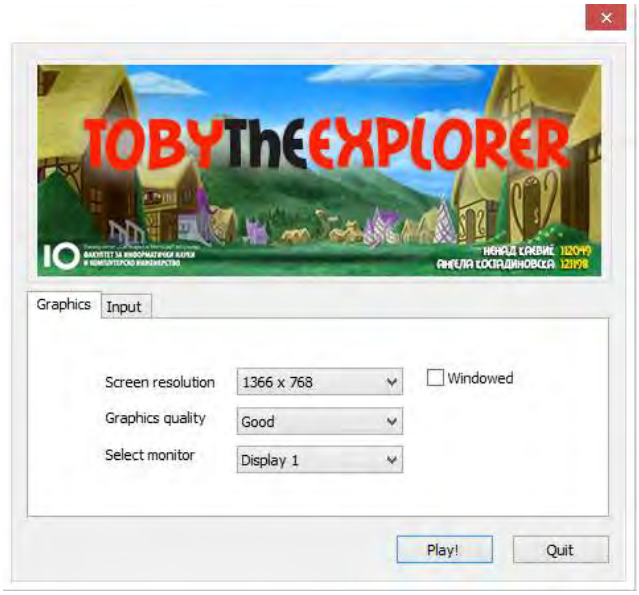


Fig. 2. Configuration Window



Fig. 3. Main Menu UI

The player's mission is to collect all the scrolls and photos in the area. They will serve as a way to learn more about the environment. When the player collects a scroll, a paper-like pop-up shows up with description about the current location. The same happens with the photos. When the player collects a photo it swirls up and pops at the screen with an imagery of the current area. Fig. 5 shows it visually.

There is quite a lot to be discovered by the player. The game abounds with actions and lots of excitement and is made for the player to go deeper in the story and to explore more and more.



Fig. 4. In-Game UI

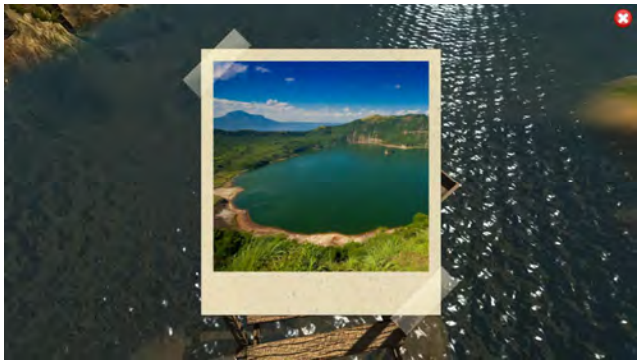


Fig. 5. Photo Popup

III. Implementation of "Toby the Explorer"

A. Used technology

"Toby the explorer" is an interactive learning game targeting kids from age 7 to age 12. Since there are a lot of technologies allowing us to implement 3D game, we decided to use the latest game engine "Unity" [13]. Unity is a cross-platform game engine developed by Unity Technologies and used to develop video games for PC, consoles, mobile devices and websites. First announced only for OS X, at Apple's Worldwide Developers Conference in 2005, it has since been extended to target more than fifteen platforms.

The reason we decided to use this engine is its clear and simply written documentation which helped us to understand the whole framework of the engine. Unity understands two programming languages: C# and JavaScript. Mainly we used C#, but there are a few scripts that are written in JavaScript. The engine gives us a chance to easily scale the game for later expansions. For example, the market is already thriving with various platforms who occupy a lot of devices. Each of those devices requires different optimizations and specifications, which can be easily achieved with Unity.

Since this game has a lot of models used in it, we had to use modeling software which helped us to import modeled objects in the game. For that purpose we used Cinema 4D [14], developed by MAXON Computer GmbH. It is a modeling program based on xyz modeling axis and modeling rigs. Our main character is designed exactly in this editor using rigs and bones to make it more realistic and acting as a real person. It

totally combined all the parts of the body, connecting them with joints to act as ankles. On the top of it, we added animations on the character which brought it to life.

Another technology we used but is less adapted and is used in very little ranges is PHP and MySQL. We used these technologies for data storage. The idea came because this game needed competitiveness, so we decided to use databases to store the scores of the player.

B. Developing the game

Unity engine was used for both front and back end of the development. We used a lot of built in tools who helped us develop the environment of the game, but we also used a lot of external software.

The terrain of the game was made with the built in Terrain Editor in Unity. It consists of brushes and terrain shifters. The helpful tool in this terrain editor is the Height Map which allows importing of black and white bitmap and converts them into terrain; depending on the blacks and whites on the bitmap, i.e. white parts make higher terrain, while black parts lower terrain. As we mentioned above, we used Cinema 4D for the modeling part. All the models were easy to import and all the textures attached to them were easy to compress and attach. The engine provides shaders, which allow us to make the surface of the models much more realistic. All the lightings and specular blooms are part of the engine. We used a lot of lightning objects to make the ambient look real. In addition to the lightning, we used a lot of ambient sounds to bring the world alive and give a full sense of the surroundings.

Apart from the frontend development and all the things visible, we implemented a lot of backend scripts. Since the game consists of many movements and physics, we had to develop all objects, starting from the floating objects on the water like the boat, to flickering leaves in the woods and waves in the water.

Most of the scripts are attached to the boy, as it is the center of everything that is triggered. In the desktop version, the motion script for the boy uses the typical system input (Keyboard) and is controlled by the “arrow” buttons or the “w”,“a”,“s”,“d” buttons. The “space” button makes the boy jump. Since the physics had big impact on this game, we had to make sure that everything will move and act properly. For that purpose, we used Unity’s Colliders and Rigid Bodies who attach Physical parameters to the objects, allowing them to bump on each other depending on their surface.

In order to store the player scores, we had to use PHP and local server to test all of the storing process. For that purpose we installed WAMP server locally and wrote two PHP scripts (display.php, addscore.php). Using the myPHPAdmin panel on the WAMP server we created an empty database and filled it with three columns using a simple query. The second stage in this process is writing a C# script in Unity where we pass the variables between the PHP and the C# script. Fig. 6 describes the communication between the database and the game.

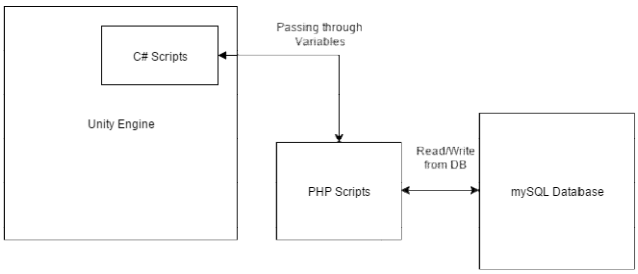


Fig. 6. Communication between the Database and the Game

This is why we found PHP very useful in this project as a mediator who handles the variables from the C# script to the MySQL Database and vice versa. Thus, it has a simple implementation and is easy to edit. The UML Class Diagram on Fig. 7 shows all the classes used to make all the communication works.

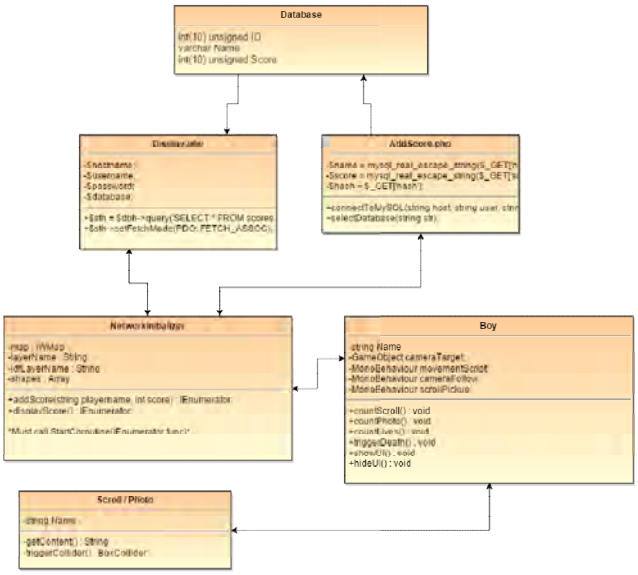


Fig. 7. UML Class Diagram

Creating the database requires installing WAMP which can be downloaded from internet for free. It will install a virtual server on the local computer with a phpMyAdmin panel. Since our implementation is based on a local server, we need to do this procedure on each device we run the game in order to save palyer’s scores.

IV. Evaluation

74% of the teachers in primary schools use digital games in the classroom. The instructors report that video games increase motivation and engagement in their students. 4 out of 5 of video game-implementing teachers use games created exclusively for educational purposes.

It’s not just teachers who are using games to assist in educating the youth of today. 56% of parents say video games positively affect their children [15].

The statistics of these studies show that interactive education might be very popular and effective in the future. This is why we think this type of games will bring the new age of education. We conducted our own small researches on how kids would react on the game. We tested 34 children between the age of 5 and 14. The research showed the following:

- 28 out of 34 children liked the game and were very attracted by it.
- 7 out of 34 (age 12 and above) think that the game could be a distraction to the kids during the classes.
- 9 out of 34 were unable to understand the plot of the game.
- 23 out of 34 think that the game can be expanded a lot more than it is now.
- 21 out of 34 think that the game should include other topics from different subjects covered at school.
- 15 out of 34 think that the game can be modified by means of covering not only science subjects, but also a society subjects.

Following the results of our research, we believe that this game could fit into the education system easily and could bring more effort into studying. Certainly this game could be used only for classes like Geography, Biology, Environments and Nature, but it could be expanded to be used for classes like Math, Physics, as well as for Society subjects. Hereafter, the game needs a couple of improvements to make it more solid for operating but mainly it is very scalable and easy to expand for further purposes.

V. CONCLUSIONS

In this paper, we describe the process of designing and implementing a 3D educational game that is very useful, and it for sure will develop the cognitive skills of the children. Because we incorporated learning material from primary school there is big chance of implementing this type of educational games in the normal learning process of the schools.

Every country should make bigger investments into this kind of projects because the technology and its own benefits are the center of the human development, in every view of our living. Game designers should work more in the domain of E-learning through the process of games, because in our opinion that is the easier way in which the kids remember things. The cognitive science states that the learning process in the children brain is mainly occurred by animations and pictures, that was the point of this game. We hope that "Toby the explorer" will have positive impact to the children and in best way will increase their environmental knowledge.

REFERENCES

[1] F. Kurzel, and M. Rath, "Project Based Learning and Learning Environment", Journal of Issues in Informing Science and Information Technology (IISIT), Vol 4, pp. 503-510.

[2] Javier Macías-Guarasa, Juan Manuel Montero, Rubén San-Segundo, Álvaro Araujo, Octavio Nieto-Taladriz, "A Project-Based Learning Approach to Design Electronic Systems Curricula", IEEE Transactions on Education, Vol. 49, issue 3, August 2006, pp. 389-397.

[3] D.O. Jorgensen, Howard, P. 2000, "Project Based Learning - A Professional Engineering Practitioner Learning Paradigm", 2nd AsiaPacific Conference on Problem-Based Learning, Singapore. December, 4-7 2000.

[4] D. Davcev, B. Risteska Stojkoska, S, Kalajdziski, and K. Trivodaliev, Project based learning of embedded systems. In *Proc. 2nd WSEAS Int. Conf. Circuits, Syst., Signal Telecommun*, January, 2008, pp. 120-125.

[5] The 70 Best Apps For Teachers And Students, http://www.edudemic.com/70-best-apps-teachers-students/

[6] D. Riley, "The video game industry is adding 2–17 year-old gamers at a rate higher than that age group's population growth," October 2011, NPD Group Press.

[7] S. Cankaya, and A. Karamete, (2009) The effects of educational computer games on studentsí attitudes towards mathematics course and educational computer games. Procedia - Social and Behavioral Sciences. 1/1: 145-149

[8] M. Prensky, Digital Natives Digital Immigrants. NCB University Press, 2001

[9] A. Ritzhaupt, H. Higgins, and B. Allred, Effects of modern educational game play on attitudes towards mathematics, mathematics self-efficacy, and mathematics achievement, Journal of Interactive Learning Research (2011) 22(2), 277-297.

[10] B. Divjak, D. Tomic, The Impact of Game-Based Learning on the Achievement of Learning Goals and Motivation for Learning Mathematics - Literature Review, JIOS, VOL. 35, NO. 1 (2011)

[11] J. Shalamanoski, B. Risteska Stojkoska, Improving learning in mathematics through 3D digital game play, CIIT 2015.

[12] M. Jovanov, E. Stankov, M. Mihova, S. Ristov, M. Gusev, Computing as a New Compulsory Subject in the Macedonian Primary Schools Curriculum, EDUCON 2016

[13] Unity Game Engine Wikipedia Page https://unity3d.com/

[14] Cinema 4D Studio http://www.maxon.net/products/cinema-4d-studio/who-should-use-it.html

[15] Children and Video Games Section- http://www.bigfishgames.com/blog/2015-global-video-game-stats-whos-playing-what-and-why/

The Use of Contemporary e-services and e-contents at Mother Tongue Classes

Vanja Jesenek
Primary School Sladki Vrh, Sladki Vrh, Slovenia
Vanja.Jesenek@guest.arnes.si

Abstract – In a modern world more and more activities take place on the web and social networks and therefore our education methods should follow and adjust. It is a high importance that the children are taught in their early age how to use e-contents and e-services in a proper way as well as having a critical eye.

Primary school Sladki Vrh was participating in a project Implementation and Usage of e-contents and e-services in a project e-Schoolbag, where we developed digital literacy and a proper use of ICT with the help of tablets at mother tongue classes. The pupils were using tablets with different e-contents and e-services, such as Socrative, Nearpod, Skitch, Padlet and were reasonably integrating these tools and contents into the classes and homework assignments and therefore giving them a significant value in education. The development and progress at learning in the class which was using the tablets was compared to the class which was not and was proven that the usage of ICT additionally motivated pupils to learn and as such influenced on work and their progress. By the same time they were upgrading their knowledge in digital literacy, which is something they will definitely need in the future.

I. INTRODUCTION

Education is constantly adjusting, changing and fulfilling due to the influence of the society, political standpoint and lifestyle. Contemporary society demands a different individual says the author of dissertation [1]. She explains that the swift technological development creates new conditions and demands of education, but on the other hand enables the solutions, accessories and tools for solving the situations as well as challenges.

Modern ICT significantly influences on contemporary didactical approaches in teaching explains [2]. The author explains that three didactical and technological fields influenced on development of ICT in education (traditional: activities connected directly on educational process; IT as didactical tool in all forms of teaching process; educational networks).

Primary school Sladki Vrh is aware of the important role of the employees in the future. Therefore it applied in year 2013 to take part in a pilot project Implementation and Usage of e-contents and e-services in a project e-Schoolbag, coordinated by National Education Institute of Slovenia. At school we decided to implement the project in the class, which was the last year at school. The students received the tablets for a whole school year. The main purpose of the project was to establish the digital literacy in several schools and to upgrade teachers' knowledge regarding the usage of IT. The emphasis was, however, on smart use of ICT and critical evaluation of it.

During the project the teachers were upgrading their knowledge in many web applications meant only for their use. We had numerous ICT lessons, online educations about contemporary e-contents and e-services at VOX conferences. Newly gained knowledge and skills were implemented into classes.

II. CONTEMPORARY e-CONTENTS AND e-SERVISES AT MOTTHER TONGUE CLASSES

Teachers of Slovene language decided to participate in the project.

The first year we used the tablets only in one class, while the next year we implemented the good practice in many other classes as well.

In school year 2013–2014 we enabled the use of the tablets at mother tongue classes only to one class of 9 graders. To be able to compare the results, effect and progress in knowledge and motivation students for work, we took as a comparison also the other 9th class, which did not use the tablets and e-contents at classes or at home for assignments.

At first we used the tablets at Slovene classes just for searching certain information, checking unfamiliar words in Slovene Dictionary and Spelling and to motivate students for work. Soon we realized that students are not able to search for information based on key words.



Figure 1. Work in Padlet

Therefore, it took a while to teach them how to search for them with the help of e-services, i.e.: Skitch, how to mark them and write them down (Padlet, Socrative, Nearpod) ...

With such activities the students expanded their vocabulary which enabled them to write better written works (abstracts, prose work, haikus, eleventh, i.e.: fakebook, storyboom …).

The students in other class, who still worked with the students' book and workbook, could not develop these skills in an appropriate way.

e-Tools and e-services, which were mentioned above, were not used by the students only to search for new information online or to translate unknown words but were used to develop contemporary reading strategies (i.e.: brain storming, Venn diagram, target, timeline, etc.).

Similar exercises were performed by the students at home and uploaded on online classrooms or were simply sent over e-mail. As teachers we were able to monitor their work regularly and offer them a prompt feedback about their progress.

Immediate feedback was quite difficult to achieve in the class without the tablets as well as individualization and differentiation during the lessons.

Also the other online applications (i.e.: viber, skype, whatsapp) had an important role in homework assignments, because we tried to encourage the students to work in pairs or teams where each and every one had his assignment and role. At dynamic forms of learning the effective use was shown at e-services where the students had to record a video, sound, make a short film, which was used afterwards for describing the way, description of a person, etc.

The students realized quite quickly that using the ICT in the class has changed the methods of work during the lessons. The lessons became more active and dynamic as well as more motivating and encouraging for the students and for us. What is more, the use of the tablets was swiftly implemented into more complex assignments where we were able to test several learning aims.

During the time the lessons in the class were performed in regular methods, which are usually frontal and non-motivational. The students were using the students' book where the key words in the texts are already stressed, so they do not have to find them.

At end of a school year we made a comparison between the classes. We gave to the class the same text and instructions to find the key words, write them down, mark them and make a summary out of them. We came to conclusion that the students who were not using e-services had some difficulties finding and writing down important key words and making a summary with the help of them. Different situation appeared with the students familiar with the work on tablets using e-services and several applications.

With the help of e-service Skitch they independently marked all important words and made a summary. It was also proven that the students familiar with e-tools were more cooperating in teamwork.

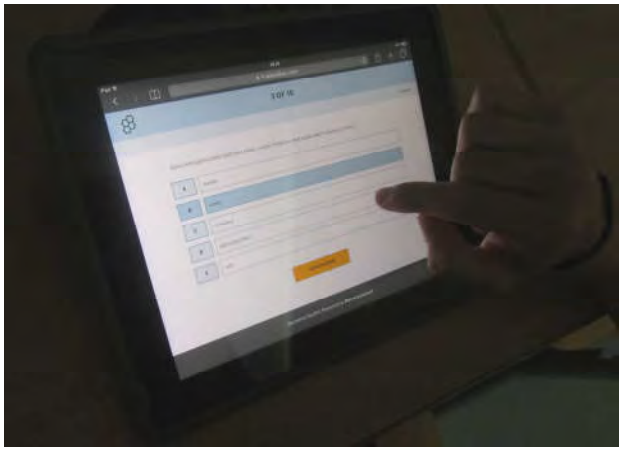


Figure 2. Working in Nearpod

To check students' work during the lesson, Nearpod, Socrative and Padlet, were shown as most efficient ones. Solving exercises in printed workbook enables to check the knowledge only for some students, because we are time limited.

We also made use of already solved exercises, which are available online, but the purpose of using such exercises was just additional encouragement and practising. The exercises were too simple and sometimes with mistakes or even incompatible with the learning process. We had to admit that there are no complex and diverse exercises on higher level when it comes to Slovene language. Therefore, we tried to reach common conclusions, plans and implementation of exercises with the help of different e-tools, e-contents and e-services, which could motivate students and would not mean additional work for the teachers. By the same time it would follow contemporary methods and teaching plan for mother tongue classes.

We began with small steps: e-tools, e-contents, e-services, and tried to upload the applications on tablets with the help of QR-code, NFC (Near Field Communication), which was often proven to be time-consuming due to technical problems.

New contents, services and tools were absorbed by students in no time therefore we needed to find new ones. During the time, when we were acquainting ourselves with all these applications, we gave the emphasis on using the online classroom at Slovene classes with the help of Moodle and tried to use all possible options offered by Moodle (i.e.: quiz, surveys, forum, debate, etc.).

We were able to monitor students' work regularly at workplace and also from home. With such activity we could see that their regular work became more intense which was shown with the statistic of visiting the online classroom.

Beside all e-tools, e-services and e-contents mentioned above, the students also discovered the meaning and advantages of using ICT. For exchanging mail, students used Arnes, which ensures network services to organizations in the field of researching, education and culture and enables their networking and mutual

cooperation in Slovenia as well with relating organizations abroad.

We used also their services such as Oblak 365 for exchanging the documents; Arnes Planner for planning the exams and learning process; online conference VOX as an example of lesson, when the kids were away; 1-ka for making different questionnaires (questionnaire for parents about the meaning of tablets at classes and questionnaire for students), etc., and we also used google and their services (Google Drive, Google calendar, google docs for writing essays or common assignments). The students also learned how to write and publish a blog.

The students had to hand over their work in a programme with no auto correction (like in Word), which enabled us to check their skill in writing and spelling on regular basis.

Month after month we have been learning and adding novelties at Slovene classes. We implemented the usage of tablets in all phases of learning and homework assignments and even though the e-services were not daily used in lessons, the work was always meaningful and followed the objective of every lesson. Mother tongue is a school subject where the fine motoric of the students and handwriting is really significant to develop the writing skill and calligraphy.

It was shown that the students in the class have forgotten many times to bring with them learning accessories and therefore could not follow the work and by the same time made it difficult for the teachers to perform their work. Such problems did not occur in the class where the modern methods with the tablets were used, because the students could work on any computer. The biggest difference between the comparing classes was shown in motivation of the students, where the motivation was much higher in class, who were using ICT and modern methods of work. Also the progress in knowledge was greater. These findings were supported by results shown in the questionnaire filled by the students who participated in the project.

III. CONCLUSION

There are numerous e-contents, e-services and e-tools online and it needs a thorough and critical assessment of what is useful or not, which demands from a teacher to be updated with the expectation of a modern society as well as students, who with the use of novelties individually dictates their own needs and interests. Teachers are obliged to educate themselves on their own initiative to keep in step with the young.

Unfortunately, while using the e-services and internet it could come occasionally to some technical problems which prevent us to perform our work efficiently.

In years to come, we will continue with such work at Slovene classes, because it was proven to be very efficient. However, before using e-tools and e-services we should check their reasonable use and values in the learning process.

Teachers must be aware that we can plan and perform modern approaches at classes only with constant monitoring, updating and developing our way of teaching as well as implementing new and different e-services. We are also positive about using e-tools, e-services and e-contents as an addition to a learning process and it is important for modern approach in education, but only with thorough consideration about the meaning and effect on motivation, work and students' knowledge.

LITERATURE

[1] P. Jesenek Bračko, *Idejni model uporabe informacijsko-komunikacijske tehnologije pri pouku geografije: magistrsko delo*, Maribor, [P. Jesenek Bračko], 2011.

[2] I. Gerlič, *Sodobna informacijska tehnologija v izobraževanju*, Ljubljana: Državna založba Slovenije, 2000.

[3] A. Flogie in A. Čuk. Kaj nam prinaša e-Šolska torba?. V *Zbornik zaključne konference projekta e-Šolska torba* (Kranjska Gora, 27.–29. 5. 2015) [Online]. A. Sambolić Beganović, ur. A. Čuk, ur. Ljubljana: Zavod RS za šolstvo, pp.. 21–30. Available: http://www.zrss.si/pdf/kaj-nam-prinasa-esolska-torba.pdf. [Date accessed: 28. 2. 2016].

[4] *Zavod Arnes* [Online]. Available: http://www.arnes.si. [Date accessed: 27. 2. 2016].

Migration from In-house LMS to Google Classroom: Case of SEEU

L. Abazi-Bexheti*, A.Kadriu*, M. Apostolova-Trpkovska*
*SEE University, Faculty of CST, Macedonia

l.abazi@seeu.edu.mk, a.kadriu@seeu.edu.mk, m.apostolova@seeu.edu.mk

Abstract – The role of e-learning systems has increased as a result of the rapid development of the information and communication technologies and their increasing relevance in all living aspects. Learning Management Systems (LMS) represent a platform for the realization of new study forms (especially the distance mode). Although primarily determined for supporting the pure set of e-learning activities, LMS supports not only study activities, but also managing activities and governing study courses. However, the appearance of Google Apps for Education has shifted teaching and learning into a completely new era of opportunities. Even though the initial belief was that Google Classroom will be used only by individual teachers and will not affect the LMS market, we are facing a complete migration towards Google classroom as a Learning Management System Solution. One can even say that we are going towards Google Drive Management. The paper presents the experience and challenges of this LMS migration at South East European University (SEEU).

I. INTRODUCTION

LMSs may not be the latest innovative educational technology, however they still remain one of the most invasive technologies in higher education. Therefore, many educational technologists are exploring LMSs from many aspects. It is important to note that despite the benefits that these systems offer in the process of learning, many researchers have considerable remarks even for the most popular LMSs and above all the commercial ones. The main remarks are that: (i) they do not have strong pedagogy; (ii) they are content focused; (iii) they present a teacher based model; (iv) they do not provide an easy way to exchange content between them (although they claim interoperability); etc. One can say that these remarks are partly true, however they do not reduce the significance of LMSs in applying successful e-learning. Google and its applications for education are shifting the complete concept of e-learning and learning management towards a new era. Starting from the latest movement in the field of teaching and learning where the shift is 'not teach but reach', Google is winning over the traditional LMS solutions. Furthermore, the current experiences in Google classroom define Google Classroom as "Google Drive Management". This description is accurate. Google Drive Management is more valuable to students and teachers than what an LMS provides because Google Apps for Education (GAFE) are essential for collaboration and feedback. Students hone future-ready skills when they collaborate and give and receive feedback in the Google ecosystem. Google Classroom automates the distribution of Google Drive files. Conventional LMSs force teachers and students into time-wasting workarounds to access and share Google Drive files.

The paper presents the experience at SEE University in migration from in house LMS to Google Classroom, including the difficulties, the challenges and the limitations of this process.

II. TECHNOLOGY INNOVATION' S DIFFUSION

The diffusion of the technology's adoption has been a constant issue in various research works for a longer period of time. One of the most influential works in this area, which although dates from many years ago, is Rogers' Diffusion of Innovations.

His study about the diffusion of innovation became broadly accepted in communications and technology adoption studies, and also entered into many other social science studies. According to him innovations have an S-shaped curve of adoption [6], as shown in Fig. 1.

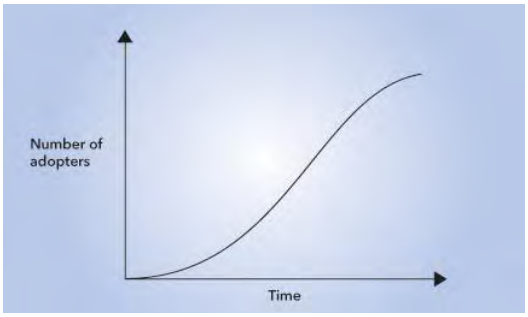


Figure 1. S-shaped curve of innovation diffusion according to Roger.

It is important to note that the innovation must not be from the technological area, although this example is often used to describe the uptake of a technology. The gradient of the curve can differ, depending on many factors such as: the nature of the actual innovation, how much benefit it provides, the audience in which it is being adopted, the easiness of adoption, etc. Anyway the most important moment in Rogers' approach is that there is usually a period of time (for any innovation) when it is used by a relatively small number of users and afterward as it enters the steep part of the curve uptake increases significantly. This is the point when innovations are accepted by a critical mass of users. Next we can notice a regression of the curve, which implies that beyond a certain point a lot

of time as well as recourses are required to enlarge the number of users.

Rogers proposes five categories of approaches towards innovation, which are now part of widespread phraseology: innovators (2.5%), early adopters (13.5%), early majority (34%), late majority (34%) and laggards (16%), based on the mathematically-based Bell curve. This model has been adapted and modified by many others, but still it remains a convenient model for both estimating and exploring the adoption of any innovation, especially technological ones.

A related work in this direction comes from Riggs and von Hippel who looked at innovations from the perspective of the developers. Actually they compared the innovations developed by users and those by manufacturers. They found "that innovations with high scientific importance tend to be developed by instrument users, while innovations having high commercial importance tend to be developed by instrument manufacturers." [9].

Similarly, according to Greller the uptake of technology within an institution follows a normal distribution curve.

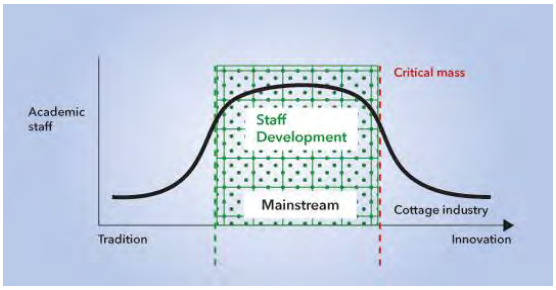


Figure 2. Pedagogic Innovation (Greller)

He presents the pedagogic innovations by taking the standard normal distribution curve and plot innovation on the x-axis and number of academic staff on the y-axis (Fig. 2). He suggests that in the process of acceptance of many technologies there comes a point when they shift beyond the early adopters and they move to the mainstream. At this moment the institutional reactions start, usually in the form of staff development, official regulations and centralization of support and resources [8]. After this period falls down the usage of technology as part of innovative solution and it moves to the part of standard usage toward the traditional end of the continuum.

III. Learning Management Systems Trends

The great interest for original pedagogical-educational practices is a reaction to the community needs for educational change. Such needs come out from the huge request and admission to higher education, the need to raise competitiveness mainly through the raise of the human potential and the necessity to take into account new approaches to learning.

These original innovative practices in education presently are mainly available through e-learning and the most pervasive e-learning tools in higher education currently are the Learning Management Systems. LMS have been present for a pretty long period in higher education, but yet they are still changing and evolving due to the technological advancements as well as the change in approaches towards teaching and learning in higher education institutions.

The latest reports show that LMSs have their common role, were a few standard features for supporting teaching and learning are yet the most used features [2]. In parallel with their common usage there is a line of new trendy features that are present in current LMSs. These trendy features usually are blogs, wikis, and similar social and connective Websites that facilitate elements to uncover and share new information.

Among the important challenges to LMSs mentioned by researchers is a necessity to include constructivist instructional design methods and pedagogy, in particular emphasizing active and shared learning and personalized attention to all students [12].

The ability of modern LMSs to integrate multimedia and Web 2.0 tools made a significant impact on instructional design and teaching practice. In particular, Web 2.0 applications may be used to enhance student interaction and collaboration, simulations and serious games are also finding their way into e-learning environment. Finally, a LMS must be adapted by using technologies in order to merge teaching, research, IT, a changing environment as well as financial accountability and managerial models.

IV. SEEU Experience in LMS Deployment

SEE University's LMS experience dates from the period 2006-2008 when SEEU initially started the usage of a commercial LMS (ANGEL) with various learning and managing tools. Although there is usually an adjustment period for most students and professors, as they learn the rhythm and patterns of online communication, the interest for using ANGEL has grown from year to year. This system was used for more than 3 years on SEEU. Although very popular and widely used among staff and students, it was very difficult if not impossible to extend, maintain or modify its business model. Also because of its price tag, it was more feasible for the university to start developing its own system. One of the requirements before building this system was to be integrated with other SEEU systems. A full integration with other electronic services on campus (enrollment services, grading schedule etc.), would be the main advantage of this LMS, something that was impossible or very hard to be achieved with other third party systems.

Therefore, in the next period a new in-house LMS-Libri (Fig.3) was developed and integrated with other e-systems at SEEU [1].

The system is in use from 2009 to present. A lot of investigation based on user voice' has been done before and after the system was designed. The main aim was to detect the most important tools in a LMS from the users' perspective.



Figure 3. LMS - Libri (http://libri.seeu.edu.mk/)

Additionally, after the system was designed, developed and implemented at SEEU, the users were again the key drivers in the process of enhancing and further system development [7].

Actually the experience that we had at SEEU, has shown that initially the users were looking for technological solution. As poorly designed as some of the institutional LMS solutions are, they typically allow automatic course and roster creation with updates, sharing of multiple document types, integrated standard gradebooks, and many other possibilities. Libri has all these elements.

Based on the approaches towards innovation deployment presented in the above part, the LMS solution at SEEU, was an in-house solution which at its most advanced level of usage included users as co-developers [5]. This was foreseen as a strategy to develop the system by its users and keep it attractive and convenient for usage. And this strategy functioned to keep the system alive and the users active for almost a decade.

Our experience has shown that the initial LMS deployment at SEEU, based on the statistical data [4], followed the traditional curve of pedagogical innovation (Fig. 4). But yet, after several years of Libri usage the motivation for managing lessons in it has dropped [2].



Figure 4. The curve of the Libri deployment at SEEU 2009 – 2016.

The reason for this decrease on the number of courses created in the LMS Libri can be the fact that the 'lead' users of the technology have already identified new ways of managing their lesson [3]. They are also identified as the drivers of the technology adoption in an institution. The other part of the curve usually just follows them.

In other words, our practice showed that after a certain period of the LMS usage the users (both teachers and students) are asking for a technological shift.

In the case of SEEU, the in-house advancement and development of the system could not reach the pace of tools and developments that are coming from manufacturers and huge companies offered for free.

V. GOOGLE CLASROOM PRACTICES ON COURSE MANAGEMENT

Google Classroom allows creation of a class with just a few clicks. This blended learning platform is progressing and forming its own generation by inviting millions of students to create a Gmail account and get familiar with the Google ecosystem. Of course, it comes with its pros and cons (Table 1), but however its main strength is the easiness of use. Using Google Classroom is as easy as it gets; setting it up, communicating with others, saving time, and organizing things. You can easily set up Google Classroom, communicate with others, save time, organize things, and many other opportunities immediately after you get the suite of tools called Google Apps for Education. Google Classroom's services and its natural collaborative benefits have already substituted a number of teaching tools.

TABLE I. THE PROS AND CONS OF GOOGLE CLASSROOM USAGE FOR TEACHING AND LEARNING.

	Advantages	***Disadvantages***
Google Classroom	Easy to use and accessible from all devices.	Difficult account management.
	Effective communication and sharing.	Limited integration options.
	Speeds up the assignment process.	Too "googlish".
	Effective feedback.	No automated updates.
	Clean and user-friendly interface.	Editing problems.
	Great commenting system.	No automated quizzes & tests.
	Is for everyone.	Impersonal.

There are many successful practices of Google classroom and its features. But among them, there are several practices that enable the users to take full advantage of this powerful online learning tool and understand why the use of traditional LMS is decreasing. These practices include the instant access to the work of the learners, the possibility to edit it, and return it as soon as the teacher finishes the editing. Furthermore, ongoing interaction of Google Classroom allows building on previous answers after feedback and resubmitting work.

Another significant practice is the possibility to create a personalized learning environment. In Google Classroom you can create different classes for different subjects, materials, learning levels, and learning needs. This promotes personalized learning by offering a variety of choices, online assignments and resources to the learners in order to address different learning preferences.

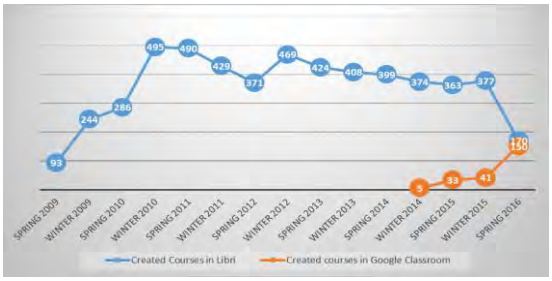


Figure 5. The number of courses in Google classroom for the last four semesters.

One can access Google Classroom from a web browser on any device, that's why it is an ideal environment for any learning setting, as well as for mobile learning. At the same time, it provides a variety of support materials by being able to attach any kind of multimedia to Google Drive, such as videos, audio files, PDF documents, etc. These are the experiences that where identified in the use of Google Classroom at our institution that were recognized as the main factor for the trend of transferring courses from the in-house LMS to Google classroom. Figure 5, shows the number of courses created by SEEU 'lead' users in Google classroom for the last four semesters.

VI. CONCLUSION

One of the main difficulties in Learning Management Systems (LMS) is the maintenance of the balance between learning and management. The primary aim of LMSs was to assist educators to manage learning. As a result of this way of thinking, the most promoted features were: (i) managing content, (ii) roster students, and others. There are numerous research studies and reports from the early period of LMS use that explore and raise the importance of the management functions of an LMS [10]. As time goes by the accent of research falls in student experiences, assessment and efficiency of the tools [12]. The use of Google Classroom as LMS alternative holds the promise of a more student-centric approach, and encourages students to reach across the limitations of academic terms.

Consequently, Google applications for education are among the most significant e-learning tools that are considered by many institutions and teachers [11]. In addition, Google Classroom is one of the main players in GAFE, who among all of its perks, such as organizing and managing online content and assignments, provides a safe environment for both online facilitators and learners to communicate with each other at no cost for the educational institution.

Of course, other powerful open source platforms, such as Moodle, Sakai and others, with a vast range of features, remain among the main players in the LMS market.

Regardless the features and the possibilities offered by an LMS, our experience shows that apart from the distribution of resources they are also used for the collection and grading of assignments. In other words, in majority of cases the LMS were used mainly for mirroring of existing paper-based procedures. Google Classroom, used in combination with Google Sites, does all of this much more easily, through a much friendlier interface. And it is all neatly integrated with students' and teachers' Google Drive, accessible on their various devices – on all platforms. There is no need for a local web server.

What are the future developments expected to be? Will the number of courses using Google increase whereas the ones that use LMS Libri decrease? This is something that we cannot claim for sure. However, as indicated from the data, this trend is very evident and the expectations are that there is going to be a very strong competition.

[1] L. Abazi-Bexheti, Z. Dika, A. Besimi, "In house Development of Learning Managment Systems Using Modular Approach". Proceedings of Ed-Media World Conference on Educational Multimedia, Hypermedia & Telecommunications (pp. 5915-5920). Vienna: Association for the Advancment of Computing in Education, Vienna, Austria, 2008.

[2] L. Abazi-Bexheti,A. Kadriu,L. Ahmedi, "Measurement and Assessment of Learning Management System Usage", Journal Proceedings of the 6th WSEAS/IASME International Conference on Educational Technologies (EDUTE'10), Sousse, Tunisia, 2010.

[3] L. Abazi-Bexheti, M. Apostolova Trpkovska, A. Kadriu, "Learning Management Systems: Trends and Alternatives", MIPRO 2014 - 37th International Convention, CE - Computers in Education, Opatija, Croatia, 2014.

[4] L. Abazi-Bexheti, B. Cico, M. Apostolova-Trpkovska, B. Ismaili, "Dimension of E-Learning Usage Based on Level of Studies: The Case of SEE University", Proceedings of the International Conference on e-Learning'15, Berlin, Germany, 2015.

[5] L. Abazi-Bexheti, E. Jajaga, M. Apostolova-Trpkovska, B. Ismaili, "An Experience in Integrating Learning Management System with user's Google Drive, 38th International ICT Convention MIPRO, Opatija, Adriatic Coast, Croatia, 2015.

[6] E.M. Rogers, "Diffusion of innovations" (1st edition). New York: Free Press, 1962.

[7] V. Shehu, A. Besimi, L. Abazi-Bexheti, M. Shaqiri, "Usability Issues While Building A New LMS", ITI Proceedings of the 31st International Conference on Information Technology Interfaces, pp. 317-322, Cavtat: SRCE, Croatia, 2009.

[8] W. Greller, "E-Learning and Pedagogy". (2005, March 22). Retrieved December 15, 2008.Available from: http://www.slideshare.net

[9] W. Riggs, E. Von Hippel, "The Impact of Scientific and Commercial Values on the Sources of Scientific Instrument Innovation". Research Policy 23, 459-469, 1994.

[10] W. Brandon, J.A. Larusson. "Strategic Directives for Learning Management System Planning". Brandeis University. ECAR Research bulletin, 2010. Available from: https://net.educause.edu/ir/library/pdf/ERB1019.pdf

[11] The Economist Intelligence Unit. "The future of higher education: How technology will shape learning". Sponsored by the New Media Consortium. 2008. Available from: http://www.nmc.org/pdf/Future-of-Higher-Ed-(NMC).pdf

[12] Y. Kats. Learning management systems and instructional design: best practices in online education, Hershey, Pa. : IGI Global, USA 2013.

Survey Analyses of Impacting Factors in ICT usage in School Management: Case Study

Bekim Fetaji*, Majlinda Fetaji*, Refik Azemi* and Mirlinda Ebibi*
* South East European University, Contemporary Sciences and Technologies, Ilindenska bb, 1200 Tetovo, Macedonia
** International Balkan University, Information Technology, Skopje, Macedonia
b.fetaji@seeu.edu.mk, m.fetaji@seeu.edu.mk, or13562@seeu.edu.mk, m.ebibi@ibu.edu.mk

***Abstract*—Today, teaching and ICT are two strongly interrelated concepts, at all levels of education and in all subjects. Dependence created by ICT teaching can hardly be compensated in another form. Data from numerous global research proved that using ICT can help students to learn and teachers to manage information more effectively. However, this does not mean that ICT will make a difference, simply being used. Different findings have suggested that although ICT can improve learning, there are a number of issues that need to be considered if such technology will eventually make the difference. Therefore, care should be taken in this broad level, in where and how and what the impact level will be. There are several issues and impacting factors that need to be taken under consideration. First is the effect of "interrelated integration" of ICT compared with other interventions, the second is the effect of the provision and use of ICT at a general level. In order to analyze this, Case study analysis of high schools in city of Mitrovica in Kosovo is realized. Insights and recommendations are provided.**

Keywords—ICT impact on education, Management of education, evaluation of technology enhanced education

I. INTRODUCTION

There has been extensive research on computer assisted instruction (CAI) and computer-based learning (CBL). In their study [2], [3] suggests two things: first, the impact of computers must and will grow, secondly, ICT helps improve the learning to a level [5]. Many studies have shown that ICT does not increase achievement, but rather "the teacher can increase the achievement of students when they use ICT to support their learning." Substantial benefits to students' achievements are more accessible when using ICT is planned, structured and effectively integrated.

ICT in teaching has an important role and its impact [4],[7],[8] on the advancement of educational processes related to effective teaching and learning, and modern research in this field is almost irreplaceable. Another important element is the use of different software platforms that facilitate learning visible and make it more concrete, more practical and applicable to everyday life [6]. Of course, the use of ICT in teaching practices cannot be imagined without the continuous professional development of teachers, to increase competence in the use of ICT and global tracking technological innovation in the field of education [4], [8].

"The integration of ICT in all major institutions and social organizations means that the need to equip young people with the capacity to understand and exploit the potential of such environments is not an alternative, but it is now a necessity " [1]. However there are a number of issues that need to be considered and solved if such technology could and will eventually make the difference.

According to [8] for years now the schools are making efforts for the integration of ICT with a substantial investment in hardware and software. In recent years the Ministry of Education seems to have gone to another level with investments in electronic platform to facilitate the integration of ICT in schools. Integration of ICT in integral system operation activities in all areas and subjects in the first requires the use of existing equipment and secondly the development of ideas on what traditional technology can be used new technologies to learning.

However there are several issues and impacting factors that need to be taken under consideration. First is the effect of "interrelated integration" of ICT compared with other interventions, the second is the effect of the provision and use of ICT at a general level.

II. RESEARCH METHODOLOGY

The research methodology used is quantitative research. Primary quantitative data has been mainly collected from questionnaires particularly designed for this research which has been distributed to the management and administration of school institution.

The target group were teachers of different subjects and school levels, schools, urban and rural, high school, vocational school, art school and special school (school for children with special needs).

The research methodology is based on the questionnaire instrument which has been distributed to educational institutions in the municipality of Mitrovica. The questionnaire consisted of 32 questions, where: 9 questions – consisting basic information and other questions have been distributed for assesing the specific impacting factors.

35 Questionnaires have been distributed to school Principles and recived 34 of responses from them. 550 electronic

questionnaires have been sent through e-mail to teachers and recived only 115 responses and waited approximately 3 months for the responses.

III. CASE STUDY ANALYSIS

In order to analyze the impacting factors in usage of ICT in school management a Case study analysis of high schools in city of Mitrovica in Kosovo is realized.

The revealing of the improper use of technology in teaching practices is reflected in the questionnaire where the use of technology in teaching school for 21 responded never, 43 times, 36 on average, and only 15 always.

This level of knowledge is not justified by the fact that the sample used in the age group 57% are new and 91% have a personal computer at home and have attended several training programs for computer work.

The rate of use of the applications of Office that represents the foundation of computer, according to survey results proves to be low, which is reflected in the quality of teaching and learning concrete and practical research, although 73% of the sample of teachers are faculty and 12% have professional master and 25% of them are with HPS. As well as their teaching experience is another indicator that should better reflect the level of use of ICT in teaching. Where 69% have 1-19 years' experience of teaching experience, 18% of 20-29 years' experience and over 30 are 13% of the target group. Also these results are questionable due to the fact that 51% have completed the training ECDL while 49% have attended other trainings.

Regarding the use of the Internet is a fundamental resource for the use of technology in teaching practices are very worrying results because 41% of respondents said they do not use at all.

The result of the test preparation for students is more gratifying because 60.86% stated that the tests produce themselves. Contrary to what a few of them use technology to prepare for the presentation of teaching units where only 21.73% said they always use.

IV. RESULTS

Out of the 35 questionnaires distributed to high school principals received 34 of them. The rate of use of the applications of Office that represents the foundation of computer, according to survey results proves low - beginner level, which is reflected in the overall management of the school, although 22% of directors are professional qualifications master and 41% of them have over 8 years experience in leadership.

As for this other software, such as software for the preparation and management of school schedules and project management software for schools (schools that can improve performance) it appears that the level of their use is very disturbing, where only two schools (6 %) use the software for preparation of schedules electronically and not in full measure. Even more alarming is the result of Internet usage where 50% have basic knowledge.

The result of processing the reports came out very satisfactory 91% in electronic form, as a result that directors engage subordinates, in most cases (45%) teachers for the preparation of these reports. These reports are often delayed by the results for two reasons: not the best approach to technology (72%) and neglect of dependents (28%).

Electronic communication internally and externally according to the results of the questionnaire carried out to the extent that 44% referring assistance again subordinates.

Gender - in terms of gender structure of school principals 75% of them are male and 25% female, which means that significantly male-dominated.

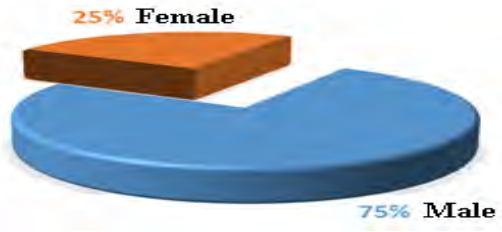


Figure 1. Gender Analyses

Age - From graphic coverage in the following figure we see that the age over 51 years dominates with 81% in the leading structures of educational institutions in the municipality of Mitrovica, and this comes out that is not a good indicator of the level of use of technology.

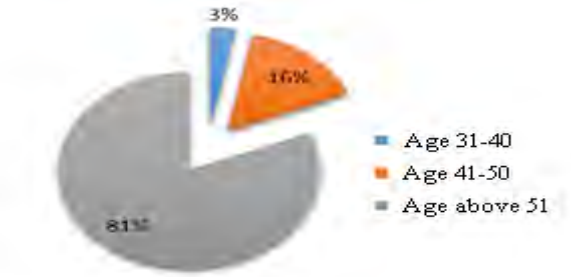


Figure 3. Analyses of Age Distribution

Experience - From the result appear in the statement of experience 41% of school principals have extensive experience in leadership, then over 8 years since one hand is a positive indicator, as this experience helps in leadership, from the other side of the mirror age did handicapped in the use of technology.

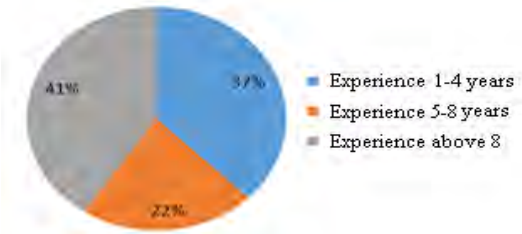


Figure 2. Analyses of Work Experience

Professional qualification - Qualification structure of school directors in Mitrovica is high, where 22 of them are in college 4 years, 7 to master and only 3 with college. This indicator should affect the efficient use of technology, but considering that all these preparations are social profiles, contributing to minimal use of technology.

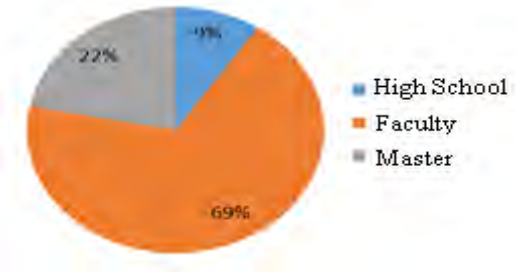


Figure 4. Analyses of Professional Qualification

Access to computers - Based on the results presented from the survey results that all principals of schools are equipped with computers, which is an indication of very important on the one hand and a disturbing indicator on the other hand, since the level of use their very low.

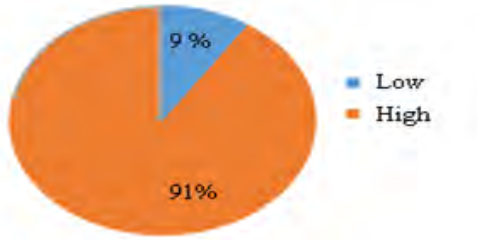


Figure 5. Analyses of Access to Computers

Long experience as Director, Computer possession, following some training in technology is inconsistent with the level of full use of technology for the management and creation of climate use in schools.

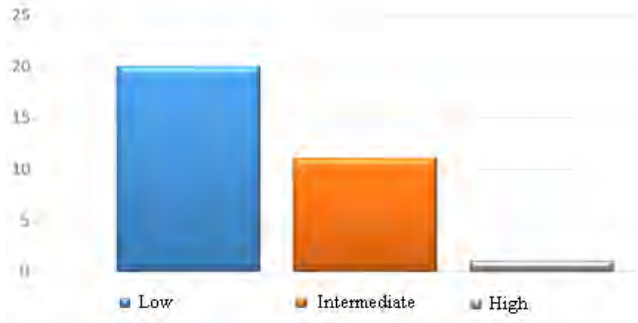


Figure 6. Analyses of Computer Skills level

Computer skills - from 63% of directors have little knowledge in the use of technology, despite the fact that all possessed a computer, 34% and 3% average knowledge of enjoyable knowledge. The following image is reflected in the rate of computer ownership by directors.

MS Word - 34% of beginner users and 66% advanced of initial users, it is known that the use of this application is more than necessary.

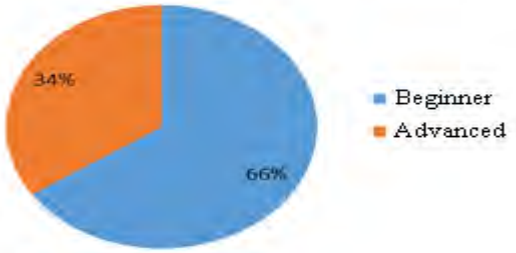


Figure 7. Analyses of usage of Word Text processor

MS Excel - worrisome is the fact that 53% of school principals at all do not use Excel application, given that the majority of statistical reports issued by this.

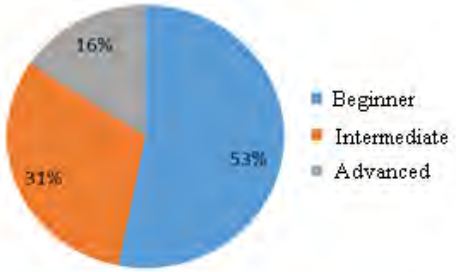


Figure 8. Analyses of Excel usage

MS PowerPoint - Even this mass application unsatisfactory use directors, which will help a lot for the preparation and presentation of data.

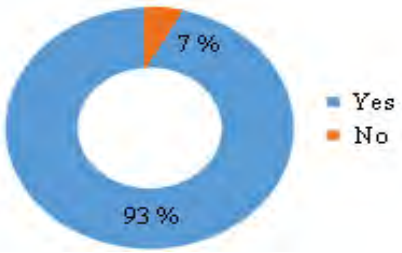


Figure 9. Analyses of PowerPoint usage

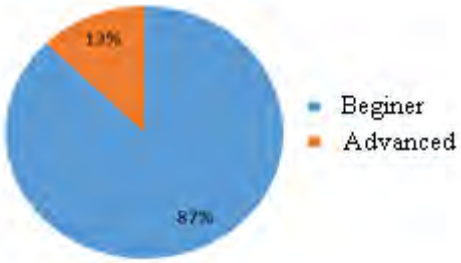


Figure 10. Analyses of Access usage

MS Access - figures 87% of directors who are basic users and do not know to use more advanced functions, which is very high percentage, given that all institutions should have a database.

MS Project - All schools in different time period the developed projects. The best way of managing these projects is to use applications which will help greatly as MS Project. None of the directors do not use this application that would facilitate the work on the projects forwarded.

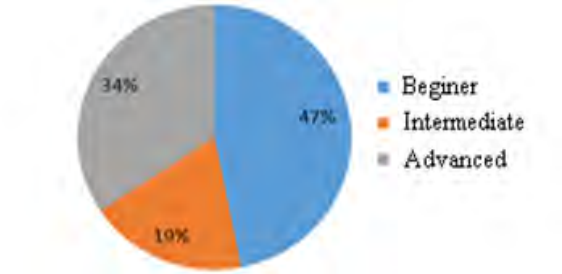


Figure 11. Analyses of MS Project usage

ASC timetable - everything happens on time to school. The tedious work of manual drafting a school schedule, where the figure 12, we see that 94% of schools have chosen this way. This will bring a lot of trouble and a lot of schools is difficult for the creation and management of overtime.

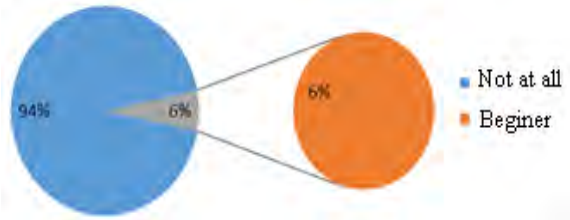


Figure 11. Analyses of usage of aSc Timetable

Schools spend great time in preparation, testing, distribution (to teachers, students and parents). Also schools also face difficulties in adjustment to schedule and its eventual modification. It often happens that two weeks before the start of the school year to spend only testing schedule. This is a handicap as a result of not using different software that enable easy creation and management of school hours in the shortest time possible. Also this time can also publish on the website of the school.

Website - figure 50% of those who are beginners in the use of the Internet is alarming, because all communication nowadays is oriented towards and through Internet technologies. Everything in the world of technology today is preceded by communication via the Internet. Development and advancement of all processes that occur in school is dependent primarily by communication through internet technologies.

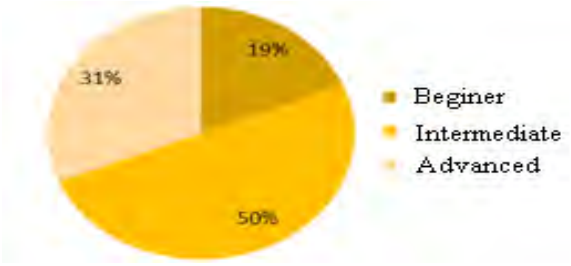


Figure 12. Analyses of Internet usage

Reports - Measure the percentage of 91% electronic reporting is inconsistent with the computer skills of directors. To this percentage is due to staff-subordinate directors engage others to prepare reports on electronically.

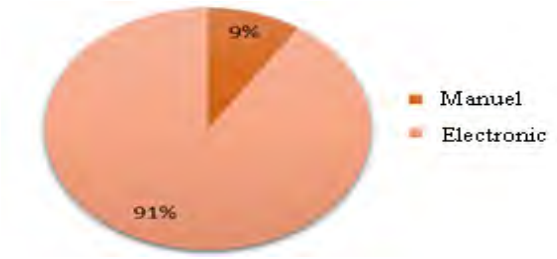


Figure 13. Analyses of Reporting

Compiler reports - greatly principals hire teachers for the preparation of reports. This result gives us realize even one of the reasons why the level of technology use is low, so they are upheld to their subordinates and this did not make them to engage and commit themselves more in the use of technology, given the fact that they have the legal obligations which sets out the duties of school principals, as p.sh management database to school, reporting through EMIS 1 -s, statistical reports, descriptive, analytical etc. The figure shown on the next page, seen rates makers electronic medical reports.

Delays reports - Clearly not the best approach to technology affects the delayed reports. These delays cause reporting problems because they know that these reports have their hierarchy of school - education director - Minister.

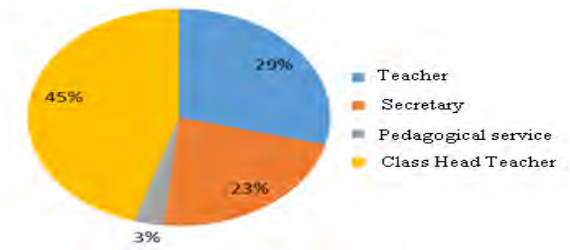


Figure 14. Analyses of Daily Reports

Communication - The tabular overview of this communication appears that the directors of subsidiaries engage in most cases teachers stand out because the result that 44% of the media use that is inconsistent with the level of use of technology by the directors.

TABLE I. COVERAGE OF EXPLOATATION TECHNOLOGIES OF COMMUNICATION EXPRESED IN NUMBERS AND PERCENTAGES

Coverage of exploitation technologies of communication directors expressed in numbers and percentages		
Communication technologies	Nr.	%
phone	2	6
Phone & Hard Copy	3	10
Phone, mail Hard Copy &	11	34
Phone, mail & Social Networks	2	6
Phone, Hard Copy, mail & Social Networks	14	44

The outlook for technology - indicators that give positive signs for the use of technology and its integration in school is the fact that all the directors in question "Will the use of technology to help management" responded positively. This gives with the understanding that they are aware of the role that the use of technology to facilitate the performance and management duties in their schools. In relation to the degree of knowledge and experience in their results that only an opinion declared not a necessity in which you should invest continuously in this direction. Directors have the opportunity to invest in knowledge for technology to do without can be large, since all possess computer, are all the time in school, have teaching technology from whose experience can benefit, provided various trainings can cooperate with each - other.

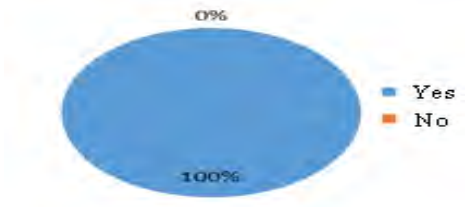


Figure 15. Impact of Technology in Management

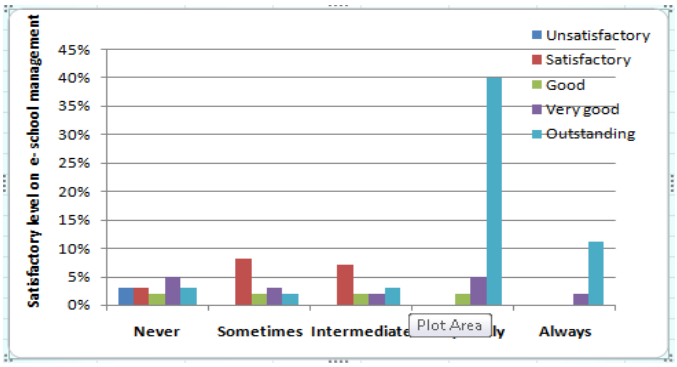


Figure 17. Satisfaction level towards use of ICT vs. previous

Technology Comments: A large number of principals surveyed (50%), marked comments about the importance of using technology in school.

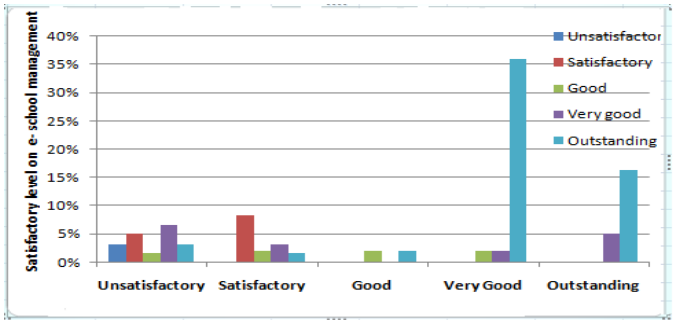


Figure 18. Satisfaction level ICT vs. confidence in using

We have used the regression equations to measure the statistical significance of the data obtained and their relationship. Now we have to test how good this regression equation predicts values of y, for given values of Xi , i=1,.., 5. The results of regression are shown in the following Tables. Results are presented in the following regression equations:

$$Y=0.99 + 0.82^* X1,\ Y=1.11 + 0.76^* X2,\ Y=1.09 + 0.76^* X3,\ Y=1.34 + 0.68^* X4,\ Y=1.27 + 0.68^* X5 \quad (1)$$

TABLE II. REGRESION LEVEL OF POSITIVE IMPACT OF ICT ON SCHOOLS MANAGEMENT X_1=GET INFORMATION IN TIME/ANY TIME

Source	SS	df	MS	Number of obs =	60
				F(1, 58) =	72.96
Model	52.7769604	1	52.7769604	Prob > F =	0.0000
Residual	41.956373	58	.723385741	R-squared =	0.5571
				Adj R-squared =	0.5495
Total	94.7333333	59	1.60564972	Root MSE =	.85052

reliable	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
getinfo	.8229256	.0963437	8.54	0.000	.6300727	1.015778
_cons	.9961506	.3424383	2.91	0.005	.3106858	1.681615

TABLE III. REGRESION ANALYSES OF SATISFACTION LEVEL OF USE OF ICT X_2=CHECK STATUS IN TIME/ANY TIME

Source	SS	df	MS	Number of obs =	60
				F(1, 58) =	97.64
Model	59.4313357	1	59.4313357	Prob > F =	0.0000
Residual	35.3019976	58	.608655132	R-squared =	0.6274
				Adj R-squared =	0.6209
Total	94.7333333	59	1.60564972	Root MSE =	.78016

reliable	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
checkstatus	.6823345	.0690518	9.88	0.000	.5441122	.8205568
_cons	1.276146	.2714185	4.70	0.000	.7328425	1.819449

The probability of the F statistic, the test for statistical significance of the regression equation (an F-value > 4.0 is usually statistically significant), for which we got the values of Fi=72, 93, 70, 93, 97, i=1,…5, means that the regression equations help us to understand the relationship between each e-service Xi and y.

V. CONCLUSION

Different findings have suggested that although ICT can improve learning, there are a number of issues that need to be considered if such technology will eventually make the difference. Therefore, care should be taken in this broad level, in where and how and what the impact level will be. There are several issues and impacting factors that need to be taken under consideration. First is the effect of "interrelated integration" of ICT compared with other interventions, the second is the effect of the provision and use of ICT at a general level. In order to analyze this, Case study analysis of high schools in city of Mitrovica in Kosovo was realized.

The rate of use of the applications of Office that represents the foundation of computer, according to survey results proves low - beginner level, which is reflected in the overall management of the school, although 22% of directors are professional qualifications master and 41% of them have over 8 years experience in leadership.

As for this other software, such as software for the preparation and management of school schedules and project management software for schools (schools that can improve performance) it appears that the level of their use is very disturbing, where only two schools (6 %) use the software for preparation of schedules electronically and not in full measure. Even more alarming is the result of Internet usage where 50% have basic knowledge.

The result of processing the reports came out very satisfactory 91% in electronic form, as a result that directors engage subordinates, in most cases (45%) teachers for the preparation of these reports. These reports are often delayed by the results for two reasons: do not use or have frequent approach to technology (72%) and neglect of dependents (28%). Electronic communication internally and externally according to the results of the survey carried out to the extent that 44% referring assistance again subordinates.

School leaders - principals need to improve very little use of technology and the opportunities that technology provides in order to achieve more efficient management of the school. Issues that schools have in most cases break down as a result of access to school directors in relation to technology and their integration into their daily work.

Directors have a basic knowledge of working on computers in a move affecting technological developments in school, difficulties in presenting reports, designs and project management schedules and school, multidimensional communications difficulties, etc. Often their work obligations carry subordinates whose cause obstacles in fulfillment of the obligations in their work.

Using technology from principals in higher level has direct impact in creating a positive school climate for technology integration and professional and efficient management.

REFERENCES

[1] Brooker, L. (2003). Integrating IT technologies in UK classrooms: Lessons for teachers from early years practitioners. Childhood Education,

[2] Condie, R., Munro, B., Seagraves L., & Kenesson, S. (2007). The impact of ICT in schools – a landscape revieë. Becta Research, January 2007. Retrieved December 13, 2007

[3] Cuttance, P (2001) "Information and Communication Technologies" School Innovation: Pathway to the Knowledge Society, Department of Education, Training and Youth Affairs pp.73-100.

[4] Duhaney, D.C. (2001). Teacher Education: Preparing Teachers to Integrate Technology. International Journal of Instructional Media.

[5] Ganesh, T. G. & Berliner, D. C. (2005). Practices of computer use in elementary education: Perceived and missed opportunities. In D. Wray (Chair), ICT in education. Paper presented at the European Association for Research on Learning and Instruction 2005 Biennial Conference, Nicosia, Cyprus

[6] Mazrek, A, (2012), International Journal of Education and Development using Information and Communication Technology (IJEDICT), 2007, Vol. 3, Issue 2, pp. 57-67

[7] Kozma, R.J. (Ed). (2003) Technology, innovation, and educational change: a global perspective. A report of the Second Information Technology in Education Study (SITES)

[8] Hirsh, Stephanie, Higgins, Steve (2010) Does ICT improve learning and teaching in schools, Published by Newcastle University-A Professional User Review of UK research undertaken for the British Educational Research Association.

Case Study Analyses of Semantic Security using SQL Injection in Web enabled ORACLE database

Majlinda Fetaji*, Bekim Fetaji*, and Mirlinda Ebibi*
* South East European University, Contemporary Sciences and Technologies, Ilindenska bb, 1200 Tetovo, Macedonia
** International Balkan University, Information Technology, Skopje, Macedonia
m.fetaji@seeu.edu.mk, b.fetaji@seeu.edu.mk, ebibi.mirlinda@gmail.com

Abstract— **The focus of the research study was to investigate the semantics security and possibilities of accessing and facilitating information within a database. Because the information in a data base is semantically related, it is possible to determine the value of an attribute without accessing it directly so this could be a start for a concern about the databases security especially in the near future. The aim of the research was to analyse security issues concerning databases, especially inference problems that are most serious in statistical databases where users can trace back information on individual entities from the statistical aggregated data. In order to analyze this, a case study experiment was realized and insights as well as recommendations are presented.**

Keywords— Security in database, SQL injection, semantics security, survey of database security, accessing database

I. INTRODUCTION

As the use of the web grows on both Intranets and the public Internet, information security is becoming crucial to organizations [11]. Especially, with many systems implementing dynamic creation of Web pages from a database, corporate information security is even more vital. As we know from previous works through our studies, one of advantages of using databases, is Data integrity and security, which in other words would mean 'If data is always accessed through the DBMS, the DBMS can enforce integrity constraints on the data [8], [12] and make our data more reliable.

We know that the web provides a convenient, cheap, and instantaneous way of publishing data. The data that we publish, we store it in a database for many reasons. We also know that it is extremely easy to disseminate information, it is equally important to ensure that the information is only accessible to those who have the rights to use it. But with all this, it is also known that in today's ways of doing business, more and more people are concerned in how to break into other peoples security models [10], [11] and do harm in many ways including compromising, making the valuable data un- accessible or other means of modern criminal activities.

These and many other issues are topics directly connected to database security that play a crucial part in the domain of the research study.

'One of the first security issues that a DBA must confront is setting up access to the database from the Web server. Whether using a CGI script, server-based middleware, or a query tool, the server itself must have access to the database'[9]

Database engines can be targeted on several levels. All modern database systems may be accessed through specific ports and anyone can attempt direct connections to the databases effectively bypassing the security mechanisms used by the operating system. These ports remain open to allow communication with legitimate traffic and therefore constitute a major vulnerability [7].

Other weaknesses relate to the actual database application itself and the use of weak or default passwords by administrators [6].

II. SURVEY OF SECURITY IN DATABASES

2.1. Survey of Table Access Control

Table access control is probably one of the most overlooked forms of database security because of the inherent difficulties in applying it.

An example as described in [1] would be allowing read access to user imputed information to the public. If a user just imputed the information why would they have to look at it within the same session?

Or, if a table is just used for system reference why it should have any other permission beside read available as discussed in [14]. Database security provides the third layer of security in a model-driven report. If you restrict access to tables or columns, the database will return an access denied error for all unauthorized access. If you include in your model any tables or columns that are subject to database security, a database error message will be returned if a user accesses a model item that maps to a table or column that he or she is not authorized to view. While database security at the table or column level is necessary in some scenarios, it is important to consider how it affects ad hoc report navigation. A user who gets a database error message while navigating a report must retrace his or her steps to get back to the portion of the model to which he or she has access as described in [13].

It is time and resource consuming to determine what we cause when we allow and deny access, even though not always we know the security concern is over fulfilled.

2.2. Survey of Vendor-Specific Security

There are three main objectives to consider while designing a secure database application:

1. Secrecy: Information should not be disclosed to unauthorized users. For example, a student should not be allowed to examine other students' grades.

2. Integrity: Only authorized users should be allowed to modify data. For example, students may be allowed to see their grades, yet not allowed (obviously!) to modify them. Or the other ways, professors that don't have to do with a given course shouldn't have access for modification of the grades of the students of that course!

3. Availability: Authorized users should not be denied access. For example, an instructor who wishes to change a grade should be allowed to do so [12].

In this context it is important to consider the user groups and roles played by certain users we implement, also other means of ensuring security, like controlling transactions, views, also using security statements provided by the database programming language and so on.

With most of the dynamic connections to databases, first we need to keep in mind that there must be a connection with full access granted to the Web server because various queries will need to access different tables or views to construct the final HTML from queries.

As we can say, the risk here is obvious: A single data source on the server must be given broad access capabilities.

So server security is crucial. For example, an ODBC data source given full administrator access could potentially be accessed by any other program on the server. A program could be designed to retrieve private information from a data source regardless of whether the program's author is permitted access. This security problem is most dangerous on a system where users are allowed to upload CGI scripts or programs to run on the server.

Most common database vulnerabilities include: the use of default configurations with default user names and passwords, buffer overflows in processes that listen on well- known TCP/UDP ports, SQL Injection via the database's own tools or web front-ends, the use of weak passwords for privileged accounts and so on. So what can we do against these and more vulnerabilities?

We can start by updating with patch maintenance. Also, in a lot of forums there is said that we should change default ports: 1521 in oracle, and 1433 in SQL Server.

To prevent unauthorized access to our data, we will have to make sure that the server that owns the database connector is physically secure and does not permit unrestricted program execution.

Different vendors offer somewhat different approaches towards a secure database?

In the next headings are analyzed some of them.

2.3. Survey of Microsoft Security Implementation

For database access, Microsoft Corp [2],[3],[4] and [10] has integrated IIS security with Microsoft SQL Server through the Internet Database Connector. User logins must occur through an HTML login form, but the information may be verified using a SQL Server stored procedure.

Better integrated security came with the release of SQL Server 7.0. Microsoft is also integrated the Kerberos security architecture into Windows NT Server 5.0. By releasing the server, Microsoft hoped to integrate the Kerberos native to NT Server with public key security. Microsoft has released a Certificate Server API in an attempt to create a Certificate Server standard.

2.4. Survey of Informix Security Implementation

Informix Corp. (Menlo Park, Calif.), like Sybase, currently relies on the logon security present in the Web server.

Therefore, any access is specified through traditional ODBC-type login channels, passing the user and password information through the connectivity middleware.

Specific drivers called Universal Web Connect are available to integrate Informix database security with both Netscape and Microsoft Web servers [3].

We can achieve both levels of database security, keeping the data intact from loss and secure, by using GRANT and REVOKE statements to set privileges at the database, table, and column levels.

Setting different types of users also separates users responsible for managing the database, DBAs, and normal users.

Concerning that the data is safe and sound after such an event, Informix-Server Engine (SE) uses transaction logging, auditing, and backups. Every activity performed by the SE server is considered a transaction, whether it's a query, an insertion, or a deletion. As SE performs the activity, it also writes what it did in a log file. When a failure occurs, SE uses an auditing process to compare what is in the log file to what is actually in the database.

If they don't match, the audit process applies the activities from the log against the database.

For data to maintain its integrity and consistency, SE uses integrity constraints, data locking, and process isolation levels (unleashed).

2.5. Survey of Sybase Security Implementation

Sybase Inc. (Emeryville, Calif.) provides a rather elegant way of protecting data access through the Web. Rather than construct its own custom Web framework, the logon security present in the Web server simply passes through to the database server for authentication, taking advantage of the native security present in the database.

Sybase provides a piece of middleware called Web.sql that is used to interface with the Netscape Web servers. Communication between the Netscape Suite- spot servers and the database passes through the API for Web.sql.

According to Sybase, 'web.sql provides easy access to relational databases from the World Wide Web and allows you

to dynamically generate customized Hypertext Markup Language (HTML) documents.

When a browser requests a URL that translates to an .hts (Hypertext Sybase) file or a .pl (Perl) file (that is, a file with only Perl statements and no HTML), the HTTP server passes the request to the web.sql program.

The web.sql program reads the specified HTS file, processes the database and Perl requests contained in that file, and then composes HTML output for the HTTP server to pass to the browser [3],[4], [5].

2.5. Survey of Oracle Security Implementation

Oracle Corp. (Redwood Shores, Calif.), while providing SSL and S-HTTP security first started using Java as a basic component of its security model. The company created its Oracle Web Server [3], [4], [5] to work most effectively with Oracle clients such as Oracle Power Browser and solutions created with the Developer/2000 development tools.

According to [9] Oracle also modified the HTTP protocol to allow a state connection to be established between the client and the server. This connection actually defines a session in which the user is identified by a generated ID to identify the user.

These enhancements are present in the Secure Network Server (SNS) included in the Oracle Universal Database. The SNS incorporates support for the Kerberos security standard. Through a Kerberos system, a single login permits access to any Oracle database in an enterprise system. The Java security classes are used by the Oracle development tools to provide complete security integration with the client.

On the newer editions of the Oracle distributions, very interesting are the new features included in 10g.

Security, when database are in question is more than just logging in.

Let's say we have a sales-type data model in which the data to be protected is stored in a CUSTOMER database, in particular in the CARD_NO column. So the SELECT * from CUSTOMER; should retrieve no records to un- authorized users. A table that seems to have no records at all will likely be less attractive to an intruder (maybe the table is not even used?) than a table that has records but that hides or masks "interesting" columns.

First we need to change settings in the DBMS_RLS.ADD_POLICY call.

Next it is a good idea to create a separate schema, such as Sec_Manager, without any privileges—not even CONNECT—as a placeholder for all definitions used in a security implementation.

All objects should be created by a DBA account in the Sec_Manager schema. Having no privileges, this username cannot be used even to log in to the database, so the owned security definitions will be well protected. Of course we need to enable SQL*Plus-level access for a few maintenance and support staff members.

Next, we need to create the security implementation schema, predicate function, and security policy. We need to mention here the cautious use of dbms_session.set_identifier function, to not open the secret door for accessing the data through a SQL*Plus connection.

Next we need to encrypt data packages, because they are going 'on the open'.

In our case, encrypting the CARD_NO data ensures another layer of data protection for sensitive data. You can do the encryption with a static key defined in the external process or stored in a column in the database.

It is very important that we use another table column as an encryption key. The value stored in the key column has to be static, because if the value changes, the CARD_NO data cannot be decrypted anymore. In this case, we can choose CREATED_BY for the key, because it should never be updated after record creation. Next we can use the ENCRYPT and DECRYPT functions from the DBMS_CRYPTO package, which offers many cryptographic choices.

When all the development has been done, the promoted code can be encrypted as well, so even a DBA who promotes the scripts cannot see exactly how the security is implemented. You implement the encryption with an Oracle-supplied utility, using an operating system command such as:
wrapiname=Secure_Package.sqloname=Secure_Package.sec

When wrapped, Secure_Package.sec can be executed at the SQL*Plus prompt just like any clear text script and the Oracle engine will interpret it just as well. This same concept can be applied similarly to any other security-related PL*SQL script.

III. ANALYSES OF NEW FEATYRES IN VIRTUAL PRIVATE DATABASES

There are two column-masking approaches: by using a view and by using Virtual Private Database (VPD). We use views to hide columns we don't want to make visible to some given users, but views are methods applicable to any version of Oracle but usually the only choice when your database release is Oracle9i or earlier.

Oracle made a real move when introducing Virtual Private Database-VPD. There is no need to create a view. Rather, the VPD policy can suppress the databases display. With the help of VPD's we allow multiple users to access a single schema whilst preventing them from accessing data that is not relevant to us.

The steps that involve the creation of a Virtual Private Databases are:

1) Setting up the environment, where we first must create a user to act as the schema owner.

2) We then create an application context, while give Grant CREATE ANY CONTEXT to the schema owner then create the context and context package.

3) Create a trigger to fire after the user logs onto the database.

4) Creating security policies. In order for the context package to have any effect on the user's interaction with the database,

we need to define a Security package for use with the security policy. This package will tell the database how to treat any interactions with the specified table.

5) Apply security policies to tables.

6) Test the Virtual Private Database-VPD.

The benefit of VPD is that it provides for row-level security in your Oracle database. VPD's row-level security allows you to restrict access to records based on a security policy implemented in PL/SQL. Security policy, in case of VPD simply describes the rules governing access to the data rows. This process is done by creating a PL/SQL function that returns a string. The function is then registered against the tables, views, or synonyms you want to protect by using the DBMS_RLS PL/SQL package.

When a query is issued against the protected object, Oracle effectively appends the string returned from the function to the original SQL statement, thereby filtering the data records. The DBMS_RLS package is not granted to everyone; administrators will require direct execute privileges on the package. The ADD_POLICY procedure requires, at minimum, the name of the object to which the policy will be applied, a name for the policy, and the name of a PL/SQL function that will implement the security policy (Knox, 2004).

The most effective way to prevent records from being accessed within an RLS policy function is to return a string that consists of something that can never happen.

The Virtual Private Database is enabled by associating one or more security policies with tables or views. Direct or indirect access to a table with an attached security policy causes the database to consult a function implementing the policy. A secure application context enables access conditions to be based on virtually any attributes an application deems significant, such as organization, cost center, account number, or position. For example, a Web order entry system can enforce access based on customer number, and whether the user is a customer or a sales representative [7]. In this way, customers can view their order status online (but only for their own orders), while sales representatives can view multiple orders, but only for their own customers.

IV. CASE STUDY ANALYSIS SQL INJECTION IN WEB ENABLED ORACLE DATABASE

We know that with SQL, we can query a database and have a result set returned. SQL includes syntax to update, insert, and delete records: SELECT - extracts existing data to read, UPDATE - updates or changes existing data, INSERT INTO - inserts or creates new records holding data and DELETE - deletes data.

Most SQL databases are based on the relational database model.

According to the 4th CODD's rule (Bangalore) that Metadata (data about the database) must be stored in the database just as regular data is, we can conclude that a database structure can be accessed, read and altered with SQL queries.

We know that it is not possible to completely block all attacks. SQL injection is probably the most common Website vulnerability today!

SQL injection is the ability to inject SQL commands into the database engine through an existing application. It is a type of security exploit in which the attacker adds SQL statements through a web application's input fields or hidden parameters to gain access to resources or make changes to data. SQL injection attacks can be used to steal information from a database from which the data would normally not be available and to gain access to host computers through the database engine. There are four main categories of SQL Injection attacks against Oracle databases [5]:

1. SQL Manipulation
2. Code Injection
3. Function Call Injection
4. Buffer Overflows

SQL manipulation typically involves modifying the SQL statement through set operations (e.g., UNION) or altering the WHERE clause to return a different result. It is the most common type of SQL injection. The most well known attack is to modify the WHERE clause of the user authentication statement so the WHERE clause always results in TRUE. Code injection is when an attacker inserts new SQL statements or database commands into the SQL statement. The classic code injection attack is to append a SQL Server EXECUTE command to the vulnerable SQL statement. Code injection only works when multiple SQL statements per database request are supported. SQL Server and PostgreSQL have this capability and it is sometimes possible to inject multiple SQL statements with Oracle. Oracle code injection vulnerabilities involve the dynamic execution of SQL in PL/SQL. Function call injection is the insertion of Oracle database functions or custom functions into a vulnerable SQL statement. These function calls can be used to make operating system calls or manipulate data in the database. The issue with function call injection is that any dynamically generated SQL statement is vulnerable – even the simplest SQL statements can be effectively exploited. So here is the catch. Application developers will sometimes use database functions instead of native code (e.g., Java) to perform common tasks. This SQL statement is not vulnerable to other types of injection attacks, but is easily manipulated through a function injection attack.

The attacker attempts to manipulate the SQL statement to execute. The changed SQL statement will request a page from a web server. The attacker could manipulate the string and URL to include other functions in order to retrieve useful information from the database server and send it to the web server in the URL. Since the Oracle database server is most likely behind a firewall, it could also be used to attack other servers on the internal network[1] SQL injection of buffer overflows is a subset of function call injection. In several commercial and open-source databases, vulnerabilities exist in a few database functions that may result in a buffer overflow. Patches are available for most of these vulnerabilities, but many production databases remain un-patched.

Almost all SQL databases and programming languages are potentially vulnerable, including MS SQL Server, Oracle, MySQL, Postgres, DB2, MS Access, Sybase, Informix, etc. The threat could be from a source done and accessed through applications developed using: Perl and CGI scripts that access databases, ASP, JSP, PHP, XML, XSL, XSQL, JavaScript, VB, MFC, and other ODBC-based tools and APIs etc. One of the very first mistakes is building queries by concatenating strings, which is the main source for SQL Injection vulnerabilities. By inserting a single quote the username string is closed and the final concatenated string would end up interpreting *or 1=1* as part of the command. The -- (double dash) is used to comment everything after the *or 1=1* and avoid a wrong syntax error. This could also have been achieved by inserting the following command:

' or '1'='1

By injecting any of the two commands discussed, an attacker would get logged in as the first user in the table.

This happens because the WHERE clause ends up validating that the

username = *' '* (nothing) *OR 1=1* (*OR '1'='1'* in the second statement).

The first conditional is False but the second one is True. By using OR the whole condition is True and therefore all rows from table users are returned. All rows is not null therefore the log in condition is met.

The single quote character (‘) closes the string field and therefore allows all of the following text to be interpreted as SQL commands. It means many things: It closes the string parameter, everything after is considered part of the SQL command. What if the login fields are numeric?

Then the SQL injection is: *$formacct* = **1 or 1=1 #** *$formpin* = 1111

When going through injecting SQL statements, we first need to find out what database engine is we connected to.

Most times this will be easy if we have error messages coming back.

ODBC will normally display the database type as part of the driver information when reporting an error.

In those cases where the error message is not an ODBC message that can also be useful.

First, you know you are most probably not on a Windows box. By knowing what operating system and web server we are connecting to it is easier sometimes to deduce the possible database.

Using specific characters, commands, stored procedures and syntax we can know with much more certainty what SQL database we have injected into.

V. CONCLUSION, INSIGHTS AND RECCOMENDATIONS

One of the first things is the Web Application Gateway (WAG). Among other attributes and downsides it checks if the form inputs are within our expectation. With WAG the hidden fields and cookies are unmodified.

URL flow is according to the original design. WAG doesn’t help if we have free format text input.

We mentioned in the above chapter that concatenating can be an open door.

So there is a tip about that. Bind variables are valid in a LIKE clause and should be used. % and _ characters should be directly appended to the string rather than concatenating the SQL statement.

The following concatenation should not be used:

String name = request.getParameter("name"); conn.prepareStatement("SELECT id FROM users WHERE name LIKE '%" + name + "%'");

Rather multiple statements and a bind variable are required to properly create the SQL statement:

String name = request.getParameter("name"); name = query.append("%").append(name).append("%"); pstmt = conn.prepareStatement("SELECT id FROM users WHERE name LIKE ?"); pstmt.setString (1, name);

If an attacker cannot obtain the source code for an application, we also described that error messages become critically important for a successful attack.

Many Java applications do not return detailed error messages, but testing and analysis should be performed to determine if the application returns detailed error messages. Rather than returning detailed database error messages to the user, this information should be written to a log file.

The PL/SQL Gateway can be configured to display varying levels of error messages. The more information returned in an error message, the more useful the message is to an attacker.

All PL/SQL Gateway applications should be designed to return an application generated error page when an Oracle error is encountered rather than allowing the gateway to return an error message.

For Oracle databases, the only character at issue is a single quote. The simplest method is to escape all single quotes – Oracle interprets consecutive single quotes as a literal single quote. The use of bind variables and escaping of single quotes

should not be done for the same string. A bind variable will store the exact input string in the database and escaping any single quotes will result in double quotes being stored in the database (Paper, 2007).

We also need to maintain fixes. Those include: s/w patch level, make sure accounts have high standards for passwords and no more privileges than necessary, and work with developers or vendors to prevent attack.

Next, we need to limit database privileges. Starting from removing all privileges if possible from public accounts, which we know are open to public. Remove dba or sysadmin privileges from all accounts except the DBA's. Limit these privileges to DBA's only: "any" privileges, "create" privileges. Also remove access to xp_cmdshell in SQL Server. As for Account access matters, we need to lock the sys and system accounts in Oracle. Drop the scott account in Oracle. Give individuals accounts with specific or role-based privileges.

Maintain different passwords so we can potentially keep developers & testers out of production. Many apps store passwords in tables, so they should be encrypted. Requires programmatic changes, unless using Transparent Data Encryption, part of Oracle's Enterprise Edition, 10g.

As for SQL Server, Transact-SQL is a rich programming environment that allows interaction with an extremely large number of subcomponents and APIs. Some of these components and APIs will have buffer overflows and format string problems. These issues continue to be found on a regular basis and are unlikely to go away. So, we should not allow un-trusted users to submit arbitrary SQL; this obviously includes SQL injection attacks via custom applications.

Other good general defensive measures against Windows buffer overflows are to run SQL Server as a low - privileged account (not SYSTEM or a domain account), to ensure that the account that SQL Server is running as does not have the ability to run the command processor ('cmd.exe'). This will greatly mitigate the risk of 'reverse shell' and 'arbitrary command' exploits.

Also, ensure that the account that SQL server is running as has minimal access to the file system as this will mitigate the risk of 'file copy' exploits.

Going back to Oracle, encryption generally requires maintenance of keys. Here we need the use of strong passwords. Oracle 10g by default relies on OS authentication but can be forced to use password, which can be a good decision in certain cases.

We should be aware this could break any scripts that stop the listener – listener can be started without pwd.

One of the most popular hacker tricks is to inject a large amount of text into the listener, thereby causing it to abort. The database could be still up, but since the listener is down, no new connections can be established—which in effect is a "denial of service" attack (Nanda, May 2006). Along with the input validation, we should audit all the existing code and websites. For each field there should be defined the type of input it will allow. Then it should use filters to verify that only good input gets through and that it rejects any suspicious one. Also the Implementation of "known bad" filters for SQL reserved words and characters. Stored procedures could be used to interact with database and call procedures through a parameterized API. All input should be validated and all database users should run under the "least privilege" principle. You may need to define different roles, one for every type of query.

Also we should back up the database regularly, create a recovery plan because we may want to be able to recover in case of infection or data corruption.

Using databases gives us advantages, like data integrity and security. Having those attributes, database engines are targets on several levels. Database systems may be accessed through specific ports and there should be dynamic connections made towards them, so criminals have a start from there. To apply a defense strategy against SQL injection, we should consider comprehensive input validation, the use of a parameterized API, and to never compose query strings on an ad-hoc basis. In addition, a strong SQL Server lockdown is essential, incorporating strong passwords.

The awareness of SQL injection is increasing, although many products and applications are still vulnerable. So SQL injection is a threat and t is worth investing the time to fully understand it.

REFERENCES

[1] Anley, C. (2002). Advanced SQL Injection In SQL Server Applications , An NGSSoftware Insight Security Research (NISR) Publication ©2002 Next Generation Security Software Ltd

[2] Center, S. S. (2008). SQL Server 2008, Google Books Online

[3] Freeman, R. G.(2009) Oracle database 10g New Features. Google Books Online

[4] Knox, D. C. (2007). Effective Oracle Database 10g Security by Design, McGraw-Hill/Osborne ISBN:0072231300.

[5] McLaughlin, M.(2011) Pl/SQL Programming Oracle Database 11g, McGraw-Hill

[6] Micki Krause, H. F. (2009) Handbook of Information Security Management. CRC Press LLC.

[7] Nanda, A. (2006). Project Lockdown- A phased approach to securing your database infrastructure. Oracle technology network .

[8] Navathe, R. E. (2013) Fundamentals of Database Systems. The Benjamin/Commings Publishing, 3rd edition.

[9] Paper, O. (2007). Oracle9i Database Security for eBussiness.

[10] PAPER, W. (MARCH 2007). An Introduction to SQL Injection Attacks for Oracle Developers.

[11] Rahmel, D. (1997). Internet Systems security, CRC Press LLC.

[12] Ramakrishnan R., G. J.(2010) Database Management Systems,Second Edition. security, W. a. (n.d.). Acunetix.

[13] Sybase, S. w. (2014). Sybase web.sql™ Programmer's Guide. Last accessed: 24 December 2014 http://infocenter.sybase.com/help/topic/com.sybase.help.ocs_12.5.1.esqlcob/pdf/esqlcob.pdf

[14] Wiedman, B. (2012). Database Security (Common-sense Principles). Retrieved from Technorati: www.governmentsecurity.org Last accessed: 24 December 2014

Using Web Applications in Education

A. Babić, MBA*, S. Vukmirović Ph.D. **, Z. Čapko Ph.D.***

*University of Rijeka, Faculty of Economics, Department of computer science, Rijeka, Croatia
**University of Rijeka, Faculty of Economics, Department of computer science, Rijeka, Croatia
***University of Rijeka, Faculty of Economics, Department of computer science, Rijeka, Croatia

Abstract - Given that today's communication both in business and education takes place largely with the help of the Internet, the aim of this paper is to investigate how much we use the available web applications, tools, platforms and other technologies for an efficient and technologically enhanced communication.

In fact, our aim is to investigate changes in the development of information and communication technology, and adaptation of teaching methods and techniques in education. We have investigated the extent to which web tools are applied in everyday activities.

All teachers and students filled out the questionnaire to assess availability for web applications, their use and purpose. They were also asked to indicate whether the web tools are sufficiently used in teaching, whether their application has resulted in the expected efficiency, and what are the expectations when it comes to upgrading these tools.

In conclusion, there is the emphasis on the importance of monitoring progress in the development of technology, the need to master new and modern methods and techniques of teaching and learning.

1. INTRODUCTION

In this article, we investigate the role and significance of web applications in teaching and improvement possibilities, aimed at raising the quality of teaching in general. The authors believe that web applications are not used enough in the classroom and that it is necessary to increase the use of web applications in teaching in both a quantitative and qualitative way, and that is the opinion of both teachers and students. To prove their hypothesis, the authors used several scientific methods, including the questionnaire that played a major role. In addition, the authors used the methods of induction, and analysis and synthesis.

II. WEB APPLICATION DEFINITIONS

Web application can be defined as a set of the generated files that can be accessed through a web browser such as Google Crome, Mozilla Firefox, Internet Explorer, etc. that support html / xhtml / xml.

The difference between Client/Server applications and Web applications is that clients in a Client/Server system can send requests to the server via the user interface. However, a web application does not require an executable file (.exe), because all web applications are accessed via a web browser.

The advantage of web applications is that they are able to run on all the devices regardless of the operating system, while one disadvantage is the lack of sufficiently harmonized standards for HTML, CSS (CascadingStyleSheet) and DOM (DocumentObject Model) by web browser manufacturers.

Moreover, a business web application need not to be installed locally on a computer, because they are accessed in an identical way as other web sites without complex technical tasks and high cost of data processing and exchange.

The content of today's web sites is dynamic and interactive while in the past it was static without major advances in development. From the user's point of view, fast web-response time is critical.

Today's customers are very demanding and server response time need to be improved.

A. Using web applications

To describe in a true way how web applications are used and the importance of their use, and find out about

different ways in which users use the Internet, the authors conducted a survey of both students and teachers.

The participants ranged in age from 18 to 55 years. Out of 44 respondents, a sample of 21 people aged 18 to 25 years, a total of 10 respondents aged 26 to 35 years, 9 respondents aged 36 to 45 years, while there were only 4 respondents in a group aged 46 to 55 years.

The survey included 27 women and 17 men, of whom 12 teachers and 32 students.

The study was designed to investigate the purposes of using Internet involving research, information, entertainment (movies, music, computer games), communication, online shopping and banking, social networking, and business.

Given that most people today use the Internet, the findings were just as expected. The study found out that most of the respondents use the Internet every day, mainly for communication (84%), information (78%), entertainment and social networking (68%). Respondents pleaded Internet for business, research, online shopping and Internet banking or in less than 50 %.

E-Learning has become a common activity from the first grade all the way to the highest levels of education. E-Learning requires high technology and is yet to take hold in everyday teaching at all levels, mostly due to lack of infrastructure, inflexibility, lack of awareness and insufficient attention given to stimulate ICT education through lifelong learning programmes.

The process of education, or teaching methods and techniques involve dealing with additional problems in preparing both students and teachers for implementation of modern technology in the teaching process.

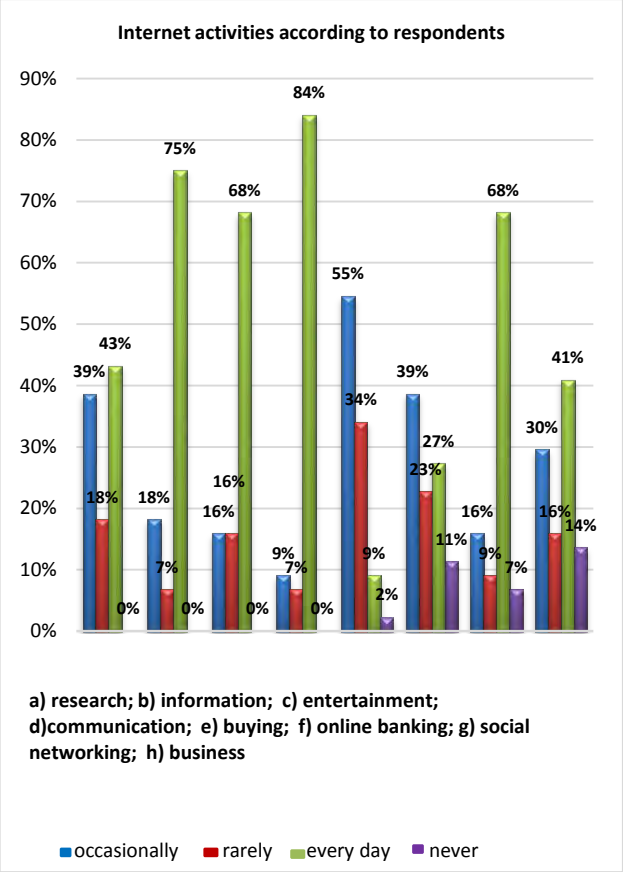


Figure 1: Internet activities according to respondents

ICT in education needs adequate computer hardware, a good network connectivity and sufficient number of workstations for undisturbed training of every individual.

Network and technical preconditions are required to use the available web applications aimed at improving the teaching process. The study included a set of web applications such as OneDrive (online collaboration), Mailchimp (marketing campaign), FormSite (digital questionnaires), Prezi (presentation software), Wordpress (web design), Piktochart (infographic design tool), and the respondents gave their opinion on each of them.

It is no coincidence that the study did not include Google tools of similar purposes which are the most popular and regularly used by majority of Internet users on a daily basis. The intention was to find out how they are acquainted with other types of web applications.

The authors specifically explored how well-informed are users about the suggested web applications, and their usage. The results are shown in Figures 2 and 3.

Slightly fewer respondents use the Internet for research purposes, or 19% on a daily basis, while 17% on an occasional basis. Also, the study revealed that 55% of respondents are occasional online shoppers. This is one indicator that web applications for online shopping have an important role when speaking of web applications usage.

Comparing Figures 2 and 3 reveals that most respondents do not use the proposed web applications in teaching methods and techniques. In other words, the respondents do not use any web application in a proportion of more than 50%. The situation is only slightly better when speaking of how well-informed are users about these tools.

The results of this study show that there is still room for improvement of education in this area. The results of this survey have been encouraging, with most respondents (over 86%) reporting that web applications are very useful. In most respondents there is a will and desire to use the available web applications. It is, therefore, necessary to take certain steps to become familiar with the available web applications and their technical characteristics.

It is certain that using web applications in teaching methods and techniques can speed up the teaching process and make it interactive and more fun. You cam see Figure 2. And 3.

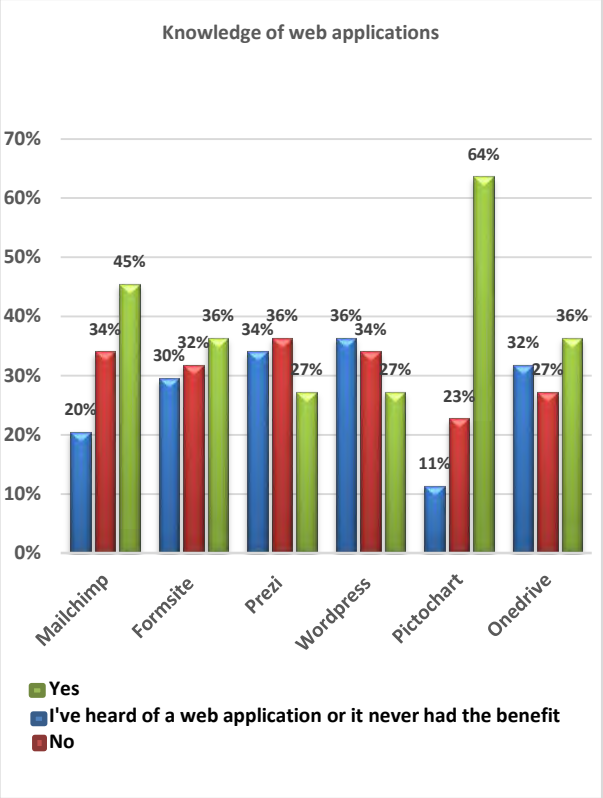


Figure 2. Knowledge of web applications

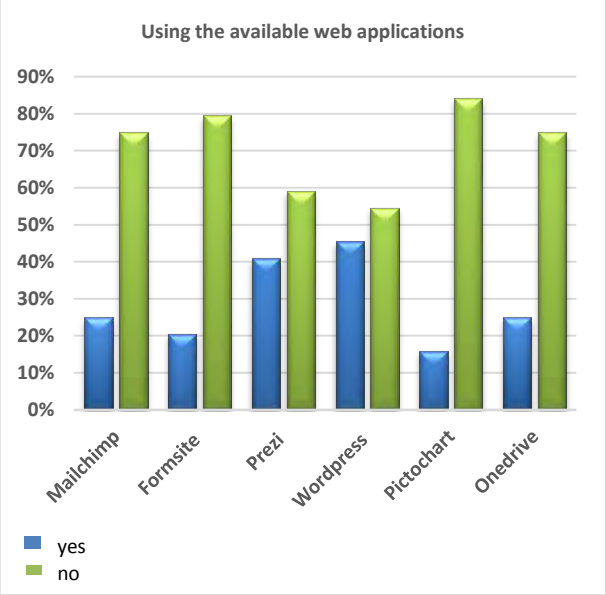


Figure 3. Using the available web applications

III. CONCLUSION

The results obtained in the survey that was conducted for the purposes of this article have proven the authors' hypothesis as explained in the introduction to be correct, and prompted the authors to suggest some recommendations for improvement of the teaching

process with the help of web applications. The objective of this research was to learn about the common obstacles, supports, and experiences as well as the tools used among early adopters of the Web as a teaching resource. Whereas most studies simply ask about online experience, time investments, and common complaints, this study attempts to understand some of the pedagogical tools and mechanisms that could benefit college faculty today.

There is some responses and comments of respondents:

- Providing computer hardware and software, network connectivity, and a training program that will encourage all participants in the teaching process to learn about the available web applications.
- Introducing a user-friendly linkage program (Merlin) - a system for e-Learning established as a platform for projects of e-Learning or course delivery studies using e-Learning technology, which is connected to the Higher Education Institutions Information System (ISVU), to set up an active calendar of lectures, exams and other activities for teachers and students.
- When using the Internet for teaching purposes, it is important that teachers regularly include new information and current news related to their course.
- Paying more attention to IT knowledge of students and improving technological infrastructure.
- Introducing a lifelong learning subject or program that will allow both students and teachers to become familiar with web applications and their capabilities.
- More interactive materials and use of technology in the classroom, such a web application would facilitate both students and teachers' activities allowing them to do tests, create case studies and store seminar papers all in one place with the goal of systematic monitoring of student performance.
- Installing web applications that will contain all material for a particular subject, integrating web applications in classroom learning environment and exams.

REFERENCES:

[1] http://www.mathos.unios.hr/wp/wp2009-10/P14_Web_aplikacije.pdf, M. Zekić-Sušac

[2] Sun MicroSystems, Getting Started with Web Applications, J2EE(TM)1.4Tutorial,http://java.sun.com/j2ee/1.4/docs/tutorial/doc/WebApp.html, 22.01.2008

[3] Phipps, R., & Merisotis, J. (1999). What's the difference?: A review of contemporary research on the effectiveness of distance learning in higher education.

[4] Web-based Education Commission (2000). The power of the Internet for learning: Moving from promise to practice. Washington, DC:U.S.Department of Education.

[5] Online teaching in an Online World, Curtis J. Bonk, Ph.D. President, CourseShare.com, Associate Professor, Indiana University

[6] A Typology of Web 2.0 Learning Technologies, Matt Bower, Macquarie University, Sydney, Australia,2 2015.

[7] http://moodle.srce.hr/2015-2016/

Qualitative Approach to Determining the Relevant Facets of Mobile Quality of Educational Social Web Applications

Tihomir Orehovački* and Snježana Babić **
*Juraj Dobrila University of Pula, Department of Information and Communication Technologies, Pula, Croatia
tihomir.orehovacki@unipu.hr
** Polytechnic of Rijeka, Business Department, Rijeka, Croatia
snjezana.babic@veleri.hr

Abstract – Being a convenient replacement for their desktop counterparts, social Web applications are nowadays widely employed in diverse aspects of human endeavour, including education. Current body of knowledge on assessment of social Web applications is composed of studies that were mainly focused on quantitative measurement of their quality when they were used in their primary setting. On the other hand, there is a lack of studies related to the evaluation of social Web applications when they are employed on mobile devices such as smartphones or tablets. This paper presents findings of the empirical study that was carried out with an objective to evaluate the mobile quality of social Web applications from qualitative perspective. For that purpose, the retrospective thinking aloud method was applied. Participants in the study were students from two different higher education institutions who completed predefined scenario of interaction with two social Web applications for collaborative writing: Google Docs and Zoho Docs. The most relevant facets of mobile quality in the context of social Web applications are uncovered and discussed.

I. INTRODUCTION

The expanding use of mobile devices such as smartphones or tablets in everyday life encourages the employment of mobile applications in the educational process and the development of a new educational paradigm known as "mobile learning" or "m-learning" [1]. One of the advantages of using mobile devices in educational settings is that they can easily communicate with other devices, which allows students to share data, files, and messages in the context of formal and informal collaborative learning activities at any time and from any place [2]. However, the use of applications on mobile devices is constrained by limited bandwidth and memory, small screen size, different screen resolutions, battery life, changing mobile context, etc. [3]. Moreover, Budiu [4] found that most commonly reported issues with the respect to the use of mobile applications, are concerning the unintentional operation of a touch screens, the use of a virtual keyboard, and lack of connectivity caused by differences in cellular networks coverage and Wi-Fi availability.

Compared to the design of desktop and Web applications, development of mobile applications requires a completely different approach. Namely, when designing a mobile application special attention should be, due to small screen size of a mobile device, paid to limitations in content display, site links, navigation buttons etc. As a result, the interface of mobile applications can become overloaded by content display, site links, navigation buttons, etc., and cause users' negative feelings and dissatisfaction with the mobile application [5].

The objective of this paper is to determine the relevant facets of mobile quality in the context of educational social Web applications. The use of social Web applications in the educational settings facilitates collaboration between stakeholders of the teaching process when performing certain teaching activities outside the traditional classroom. For instance, social Web application for collaborative writing such as Google Docs and Zoho Docs make it possible for students and teachers to work synchronously on the same documents, share, exchange and modify created artefacts with desktop, web and mobile counterparts of office suits [7].

The remainder of the paper is organized as follows. Next section presents an overview of relevant advances in the field. Research framework with brief description of all quality categories is provided in the third section. Research methodology is described in the fourth section. Results and discussion are reported in the fifth section. Conclusion of our work is given in the last section.

II. RELATED WORK

When the assessment of software quality is concerned, special attention should be paid to measuring specific dimensions of usability [3, 5, 8]. In the mobile ecosystem, usability refers to the degree to which a mobile application meets users needs (effectiveness, efficiency, and satisfaction) when applied in a specified context [5]. More specifically, users' satisfaction relates to the degree to which a mobile application fulfills users' desires and expectations. In that respect, perceived satisfaction affects users' acceptance of particular mobile application as well as users' continuance intentions. Taking the aforementioned into account, it is important to determine which factors contribute to users' satisfaction when interacting with a mobile application [3].

According to Bowen and Pistilli [9], the development of mobile applications can be divided into two categories: (a) designed to be "device-neutral" and delivered via

smartphone browser and (b) "native mobile apps" developed for specific mobile platforms such as iOS and Android. The same authors emphasize that the understanding of students' preferences and their ways of using mobile devices can be of assistance in designing and achieving success with mobile applications. The results of their research have shown that students prefer using "native apps" to a significantly larger extent than conducting mobile activities using a smartphone browser. Furthermore, Kim and Sundar [10] found that screen size determines users' acceptance of smartphones and positively influences perceived ease of use and attitude towards mobile device. Smartphones with large screens are more convenient to use, offer greater interactivity to users and enhance perceived affective quality of smartphones.

Current studies in the field indicate that specific prior skills and experiences have a significant impact on students' perceptions and their adoption of mobile technology. The results of a study conducted by Mac Callum and Jeffrey [11] revealed that students with advanced mobile technology skills are more likely to perceive the employment of mobile learning as easy and useful. The same authors come across with the finding that the level of basic ICT skills also plays a significant role in the intention to adopt mobile learning.

With an objective to explore usability of four popular mobile spreadsheet applications (OfficeSuite Viewer 6, Documents To Go, ThinkFree Online, and Google Drive) Chintapalli et al. [12] conducted an empirical study in which they compared the set forth applications with respect to following criteria: visibility; navigation, scrolling, and feedback; and interaction, satisfaction, simplicity, and convenience. The analysis of collected data revealed that applications designed for Android and iOS mobile operating systems were equally rated. However, it appeared that, compared to applications developed for iOS, Android applications suffer from shortcomings related to referencing data cells to their corresponding headers, viewing the spreadsheet as desired, and navigation.

The results of the research conducted by Shafiq et al. [13] have shown that only 23% of the users have fully adopted the features of their smartphones which indicates a problem related to the length of time that must be invested in order to perform a task smoothly. According to the same authors, there is a need for simplified lists of features, intuitive design that is fun and inspires, and decreased navigation process in terms of reduced number of touches which would result in optimized accessibility features, expedited interaction process, enhanced subjective performance and improved effectiveness of novices.

Based on the results of systematic literature review in the context of usability evaluation, Reis et al. [14] concluded that users' satisfaction was most commonly examined usability attribute which was followed by effectiveness, efficiency, learnability, errors, and memorability. Hoehle and Venkatesh [5] discovered that usability of mobile applications differs from usability of mobile devices. According to the results of their study conducted in the United States, usability of mobile application explains 47% of the variance in continuance intentions related to the use of mobile social media applications. The same authors also found that intention of continuous use of mobile social media applications is affected by following groups of indicators: (a) *application design* (branding, data preservation, instant start, orientation), (b) *application utility* (collaboration, content relevance, search) and (c) *user interface graphics* (aesthetic graphics, realism, subtle animation).

Current social Web applications designed for the use with mobile devices and therefore commonly referred to as "mobile Web 2.0" are applicable in the context of social constructivist learning pedagogies and have great potential for use on all levels and forms of education [15]. According to the results of study conducted by Cochrane and Bateman [15], critical factors that affect the acceptance of mobile Web 2.0 applications are most commonly related to features of a smartphone such as: touch screen, ease of interface use, screen size, portability - size, weight, video out, availability, Wi-Fi, cost, 3G, application availability, text entry, etc. On the other hand, in the context of examining factors which affect students' intentions to become loyal users of Web 2.0 applications designed for collaborative writing, Orehovački and Babić [16] found that perceived ease of use has significant influence on students' post-use satisfaction.

All the aforementioned implies that studies on user behavior and satisfaction when working with social Web applications on mobile devices are still in a deficit. Therefore, we initiate a research on examining quality of social Web applications meant for collaborative writing when employed in mobile environment. Details on employed research framework and methodology together with discussion of study findings are provided in following sections.

III. RESEARCH FRAMEWORK

Based on the initial set of quality attributes meant for measuring quality of Web 2.0 applications [8], proposed evaluation methodology [17], their validation on the representative sample of Web 2.0 applications meant for collaborative writing [18], mind mapping [19][20], diagramming [19][20], mashup tools [21], games [22], and mashups [23], and their refinement [9], a framework adapted to the context of evaluating mobile quality of social Web application was proposed [6]. In this paper, the adapted version of the conceptual model was enhanced with additional facets of quality which resulted in a research framework composed of 45 attributes assigned to six mobile quality categories.

System quality refers to attributes that measure the extent to which social Web application: provides various navigation mechanisms (navigability); has uniform interface structure, design, and terminology (consistency); is similar to previously used applications (familiarity); can be customized to meet users' needs (customizability); has implemented mechanisms that protect created artefacts from unauthorized use (security); operates properly with different types of devices and among different environments (compatibility); can exchange files with other applications and use files that were exchanged (interoperability); can adapt to the environment in which it is used (responsiveness); is usable on mobile devices (mobile-friendliness).

Service quality relates to attributes aimed for evaluating the degree to which social Web application: provides various forms of help to users (helpfulness); is available every time users need it (availability); facilitates management of created artefacts (artefacts management); contains mechanisms that prevent errors to emerge (error prevention); is dependable, stable, and bug-free (reliability); can quickly recover from errors and operational interruptions (recoverability); notifies users with appropriate and useful messages (feedback); supports teamwork (collaborativity); enables different types of communication among users (communicativity).

Content quality consists of attributes meant for the assessment of the extent to which artefacts created by means of the social Web application are: error-free, valid, and precise (correctness); complete, displayed clearly, and appropriately represented (coverage); unbiased, trustworthy, and verifiable (credibility); supplementable, modifiable, and updatable (timeliness); advantageous and impact users' decisions (value-added).

Performance refers to attributes that measure the extent to which the use of social Web application: enables users to execute tasks accurately and completely (effectiveness); enables users to quickly perform tasks (efficiency); responds promptly to users' actions (response time); is capable to operate under an increased or expanding workload (scalability); is usable within and beyond initially intended contexts of use (context coverage).

Effort relates to attributes dealing with the evaluation of the degree to which: the interaction with social Web application consumes small amount of physical and mental energy (minimal workload); social Web application is usable to people with the widest range of characteristics and capabilities (accessibility); users have full freedom in executing tasks by means of the social Web application (controllability); is simple to operate the social Web application (ease of use); is easy to become proficient in interacting with the social Web application (learnability); is simple to memorize how the social Web is used (memorability); the interaction with social Web application is unambiguous (understandability).

User experience refers to attributes meant for measuring the extent to which: the social Web application has visually appealing user interface (aesthetics); the social Web application is beneficial in the context of tasks execution (usefulness); the interaction with the social Web application holds the users' attention and stimulates their imagination (playfulness); users have positive perception about the use of social Web application (attitude towards use); the social Web application has met users' expectations (satisfaction); the social Web application arouses users' emotional responses (pleasure); the social Web application is distinctive among applications with the same purpose (uniqueness); users have the intention to continue to use the social Web application and recommend it to others (loyalty); the social Web application is popular (reputation).

IV. RESEARCH METHODOLOGY

A. Procedure

The study was conducted in controlled lab conditions and was constituted of two key parts: (1) interaction with two educational social Web applications aimed for collaborative writing on users' mobile devices and (2) evaluation of their perceived mobile quality in the form of retrospective thinking aloud (RTA) reports. RTA allowed respondents to complete the assignment first, and then describe their experiences of working with educational social Web applications.

As soon as participants arrived to the lab, they were welcomed and briefly informed about the mobile quality assessment study. At the beginning of the scenario performance session, forms containing a list of 46 representative steps of interaction were administrated to study participants. They were asked to finish all scenario steps twice – first with Google Docs and then by means of the Zoho Docs (both depicted in Figure 1). Once study participants had completed all scenario steps with aforementioned educational social Web applications, they had to provide a critical review related to their advantages and disadvantages and deliver it in the form of a written report. Data collected with the use of the RTA method was analyzed in two stages. Firstly, positive and negative comments related to particular Web application were extracted from the reports. Next, each comment was attached to the quality attribute whose scope it most closely fits. Drawing on the overall number of negative comments allocated, the most common problems related to the mobile quality of educational social Web applications were determined. Moreover, based on the sum of positive and negative comments attached in the context of both evaluated applications, relevance which users assign to particular attributes and categories of the mobile quality during their interaction with educational social Web applications was identified. Consequently, the attributes whose aspects were pointed out most frequently in users' reports are considered to be very important whereas the attributes whose facets were mentioned scarcely are supposed to be less relevant for evaluating the mobile quality of educational social Web applications. The same holds for the mobile quality categories. At the end of the study, respondents were debriefed, and thanked for their participation.

B. Apparatus

Majority of participants (80.86%) completed scenario of interaction with social Web applications on smartphones that run Android operating system whereas 79.01% of them used Google Chrome Web browser to complete specific scenario steps.



Figure 1. Screenshots of mobile versions of evaluated educational social Web applications for collaborative writing (left: Zoho Docs, right: Google Docs)

V. Results and Discussion

A. Participants

A total of 162 respondents (53.70% male, 46.30% female), aged 20.73 years (SD = 3.879) on average, took part in the study. At the time study took place, 74.07% of the sample were enrolled to one of the undergraduate study programs at Polytechnic of Rijeka (POLYRI) while remaining 25.93% were third-year undergraduate students of Informatics at the Department of Information and Communication Technologies, Juraj Dobrila University of Pula (UNIPU). Most of them (77.78%) were full-time students. All of them had been loyal consumers of mobile Internet on a daily basis. More specifically, 82.72% of students had been accessing the Internet via their smartphones more than three times a day while 59.26% of them have been spending more than ten hours a week on using the Internet via their smartphones. In most cases (88.89%), students had been employing their smartphones for accessing social networking sites like Facebook which was followed by other activities such as making and receiving calls (80.86%), Web browsing (71.60%), synchronous messaging and exchange of multimedia content by means of services such as Viber (69.14%), listening to music (66.05%), sending SMS messages (66.05%), taking photos (55.56%), mobile learning (35.80%), playing installed games (29.63%), mobile banking (25.93%), and taking and sharing selfies (19.14%). When the use of mobile Internet was considered, the students reported they have been employing it most frequently (88.89%) for social networking and subsequently for the purpose of synchronous messaging and exchange of multimedia content (73.46%), Web browsing (71.60%), watching videos on services such as YouTube (67.90%), accessing the educational content published on learning management system such as Moodle (56.79%), reading news on Web portals (51.23%), mobile learning (37.65%), accessing educational content located outside the learning management system (31.48%), mobile banking (24.07%), online shopping (19.75%), and playing online games (17.90%).

B. Findings

As shown in Table 1, data analysis uncovered that students had generated a total of3960 comments related to the advantages (59.09%) and disadvantages (40.91%) of the applied educational social Web applications. Taking into account results related to mobile quality categories, features of attributes meant for evaluating service quality (27.88%) and performance (25.35%) were reported most commonly during RTA sessions while characteristics of attributes designed for measuring effort (9.19%) and content quality (5.46%) were mentioned considerably less frequently. More specifically, most commonly reported aspects are related to the attributes designed for evaluating effectiveness (19.19%), availability (7.37%), artefacts management (7.27%), ease of use (5.66%), communicativity (4.24%), navigability (4.14%), content coverage (4.04%), customizability (3.94%), mobile-friendliness (3.94%), aesthetics (3.84%), and interoperability (3.43%). On the other hand, features of attributes meant for measuring consistency, feedback, accessibility, memorability, playfulness, attitude toward use, content credibility, and content value-added have not been reported even once neither in positive nor negative context.

The most common problems with which students were confronted when reaching the assignment solution were related to the dimensions of effectiveness (26.67%), content coverage (8.89%),mobile-friendliness (5.19%), availability (4.69%), customizability (4.44%), interoperability (4.44%), artefacts management (4.44%), navigability (3.95%), communicativity (3.95%), ease of use (3.46%), reliability (2.96%), scalability (2.96%), and compatibility (2.47%).On the other hand, positive experience students gained during the interaction with both social Web applications for collaborative writing was mostly associated with facets of effectiveness (14.02%), availability (9.23%), artefacts management (9.23%), ease of use (7.18%), aesthetics (5.13%), communicativity (4.44%), navigability (4.27%), collaborativity (3.93%), customizability (3.59%), satisfaction (3.42%), mobile-friendliness (3.08%), and error prevention (3.08%).The results of data analysis presented in Table 1 indicate that Google Docs (69.71%) has slightly better ratio of strengths and weaknesses than Zoho Docs (68.72%).

VI. Conlusion

Social Web applications for collaborative writing are due to their particularities nowadays widely applied in various aspects of human activity, especially education. It is therefore of great importance that they are accessible any time from any device that has a connection to the Internet. With an objective to examine to what extent two prominent representatives of this breed of social Web applications meet all relevant dimensions of mobile quality, a study was carried out. The analysis of collected qualitative data indicate that although there are mobile versions of collaborative editors, their usability is still far from those provided by their Web counterparts. The set forth primarily relates to the lack of functionalities required to complete a particular task, personalize created artefacts nor interact with peers who work on the same document, inability to operate properly in different environments (both mobile devices and mobile operating systems) nor to export created artefacts and use them in other editors without the

occurrence of errors in an artefact, and numerous bugs which were causing crashes on regular basis and loss of everything that was done up to that point. In that respect, findings reported in this paper should be used by practitioners as guidelines for developing mobile versions of social applications which quality would be equal to those of their Web versions. On the other hand, researchers can employ presented results as a background for future advances in the field.

Taking into account that work presented in this paper is part of an ongoing research, our future work will be focused on exploring the interplay of determined dimensions of mobile quality.

REFERENCES

[1] A. Botha, D. Van Greunen and M. Herselman. "Mobile human-computer interaction perspective on mobile learning". In Proceedings of the 6th annual international conference on computing and ICT research (Kampala, Uganda, 2010), pp. 3–18. 2010.

[2] S. Caballé, F. Xhafa, L. Barolli, "Using mobile devices to support online collaborative learning", Mobile information systems, vol. 6(1), pp. 27-47, 2010.

[3] A. O. Olubusola, "User satisfaction in mobile applications", available from: http://www.cs.bham.ac.uk/~rjh/courses/ResearchTopicsInHCI/2014-15/Submissions/ajayi--oluwande.pdf, [8.2.2016]

[4] R. Budiu, "Mobile User Experience: Limitations and Strengths", Nielsen Norman Group, April 19, 2015., available from: https://www.nngroup.com/articl es/ mobile-ux/, [8.2.2016]

[5] H. Hoehle, V. Venkatesh, "Mobile application usability: conceptualization and instrument development", Mis Quarterly, vol. 39(2), pp. 435-472, 2015.

[6] T. Orehovački, S. Babić, "Mobile Quality of Social Web Applications Designed for Collaborative Writing", In: 18th International Conference on Human-Computer Interaction (HCII), Toronto. Springer, Heidelberg, 2016.

[7] T. Orehovački, G. Bubaš, and A. Kovačić, "Taxonomy of Web 2.0 applications with educational potential", Transformation in teaching: Social media strategies in higher education, pp. 43-72, 2012.

[8] T. Orehovački, "Proposal for a Set of Quality Attributes Relevant for Web 2.0 Application Success". In: Proceedings of the 32nd International Conference on Information Technology Interfaces, pp. 319–326, IEEE Press, Cavtat, 2010.

[9] K. Bowen, M. D. Pistilli, „Student preferences for mobile app usage", Research Bulletin, Louisville, CO: EDUCAUSE Center for Applied Research, 2012., available from http://www. educause. edu/ecar, [8.2.2016]

[10] K. J. Kim, S. S. Sundar, "Does screen size matter for smartphones? Utilitarian and hedonic effects of screen size on smartphone adoption", Cyberpsychology, Behavior, and Social Networking, vol. 17(7), pp. 466-473, 2014.

[11] K. Mac Callum, and L. Jeffrey, "The influence of students' ICT skills and their adoption of mobile learning",Australasian Journal of Educational Technology, vol. 29(3), pp. 303-314, 2013.

[12] V. V. S. M. Chintapalli, W. Tao, Z. Meng, K. Zhang, J. Kong, and Y. Ge, "A Comparative Study of Spreadsheet Applications on Mobile Devices", Mobile Information Systems, 2016.

[13] M. Shafiq, M. Iqbal, J. G. Choi, Z. Rafi, M. Ahmad, W. Ali, and S. Rasool, "To What Extent System Usability Effects User Satisfaction: A Case Study of Smart Phone Features Analysis for Learning of Novice", In Design, User Experience, and Usability. User Experience Design for Diverse Interaction Platforms and Environments, Springer International Publishing, pp. 346-357, 2014.

[14] R. A. C. Reis, A. D. L. Fontão, L. L. Gomes, and A. C. Dias-Neto, „Usability Evaluation Approaches for (Ubiquitous) Mobile Applications: A Systematic Mapping Study", UBICOMM 2015 : The Ninth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, pp. 11-17, 2015.

[15] T. Cochrane, and R. Bateman, "Smartphones give you wings: Pedagogical affordances of mobile Web 2.0", Australasian Journal of Educational Technology, vol. 26(1), pp. 1-14, 2010.

[16] T. Orehovački, S. Babić, "Predicting students' continuance intention related to the use of collaborative Web 2.0 applications", In: Proceedings of the 23rd International Conference on Information Systems Development, pp. 112-122. Faculty of Organization and Informatics, Varaždin, 2014.

[17] T. Orehovački, "Development of a Methodology for Evaluating the Quality in Use of Web 2.0 Applications". In: P. Campos et al. (eds.) INTERACT 2011, Part IV. LNCS, vol. 6949, pp. 382–385. Springer, Heidelberg, 2011.

[18] T. Orehovački, "Perceived Quality of Cloud Based Applications for Collaborative Writing", In: J. Pokornyet al. (eds.) Information Systems Development – Business Systems and Services: Modeling and Development, pp. 575–586. Springer, Heidelberg, 2011.

[19] T. Orehovački, A. Granić, and D. Kermek, "Exploring the Quality in Use of Web 2.0 Applications: The Case of Mind Mapping Services". In: A. Harth, N. Koch (eds.) ICWE 2011 Workshops, LNCS, vol. 7059, pp. 266–277, Springer, Heidelberg, 2012.

[20] T. Orehovački, A. Granić, and D. Kermek, "Evaluating the Perceived and Estimated Quality in Use of Web 2.0 Applications", pp. 3039-3059, Journal of Systems and Software,vol. 86(12), 2013.

[21] T. Orehovački, T. Granollers, "Subjective and Objective Assessment of Mashup Tools", In: A. Marcus (ed.) DUXU, Part I, HCII 2014. LNCS, vol. 8517, pp. 340-351. Springer, Heidelberg, 2014.

[22] T. Orehovački, S. Babić, "Inspecting Quality of Games Designed for Learning Programming", In: P. Zaphiris, A. Ioannou (eds.) LCT, HCII 2015. LNCS, vol. 9192, pp. 620-631. Springer, Heidelberg, 2015.

[23] T. Orehovački, C. Cappiello, M. Matera, "Identifying Relevant Dimensions for the Quality of Web Mashups: An Empirical Study", In: 18th International Conference on Human-Computer Interaction (HCII), Toronto. Springer, Heidelberg, 2016.

[24] T. Orehovački, S. Babić, and M. Jadrić, "Exploring the Validity of an Instrument to Measure the Perceived Quality in Use of Web 2.0 Applications with Educational Potential", In: P. Zaphiris, A. Ioannou (eds.) LCT, Part I, HCII 2014. LNCS, vol. 8523, pp. 192-203. Springer, Heidelberg, 2014.

APPENDIX

TABLE 1. RESULTS OF THE RETROSPECTIVE THINKING ALOUD SESSIONS

	Google Docs		Zoho Docs	
	Pros	**Cons**	**Pros**	**Cons**
System Quality	**288**	**164**	**208**	**184**
Navigability	72	16	28	48
Consistency	0	0	0	0
Familiarity	24	0	36	8
Customizability	36	28	48	44
Security	16	4	20	0
Compatibility	32	32	20	8
Interoperability	28	48	36	24
Searchability	24	0	4	4
Mobile-friendliness	56	36	16	48
Service Quality	**380**	**156**	**396**	**172**
Helpfulness	36	0	16	12
Availability	100	40	116	36
Artefacts Management	100	44	116	28
Error Prevention	48	8	24	24
Reliability	20	0	4	48
Recoverability	0	0	0	4
Feedback	0	0	0	0
Collaborativity	48	16	44	4
Communicativity	28	48	76	16
Content Quality	**28**	**120**	**12**	**56**
Correctness	0	0	0	4
Coverage	12	108	4	36
Credibility	0	0	0	0
Timeliness	16	12	8	16
Value-added	0	0	0	0
Performance	**156**	**336**	**316**	**196**
Effectiveness	72	296	256	136
Efficiency	44	8	24	16
Response Time	20	0	0	28
Scalability	16	32	32	16
Context Coverage	4	0	4	0
Effort	**160**	**36**	**88**	**80**
Minimal action	12	8	0	4
Minimal memory load	0	4	0	8
Accessibility	0	0	0	0
Controllability	0	8	0	0
Ease of Use	100	16	68	40
Learnability	28	0	12	12
Memorability	0	0	0	0
Understandability	20	0	8	16
User Experience	**216**	**44**	**92**	**76**
Aesthetics	88	0	32	32
Usefulness	40	12	28	8
Playfulness	0	0	0	0
Attitude Toward Use	0	0	0	0
Satisfaction	56	16	24	16
Pleasure	4	0	0	0
Uniqueness	0	4	0	0
Loyalty	20	12	4	16
Reputation	8	0	4	4

Međukurikularni projekti u nastavi informatike u ekonomskoj školi – primjeri dobre prakse

Sonja Bulešić Milić *
* Ekonomska škola Pula, Pula, Republika Hrvatska
sonja.bulesic-milic@skole.hr

Sažetak - **U planu i programu informatike u srednjoj ekonomskoj školi postoji veliki raspon mogućnosti međupredmetne suradnje i provođenja projektne nastave. U članku će biti navedeni primjeri dobre prakse te poveznice (ako postoje) na aktivnosti koje su potaknule suradnju. Suradnja je ostvarena na sljedećim projektima: Eratostenov eksperiment, World Education Game - UNICEF, This is Our Time – Unite the Nations UNICEF, E-građani, Festival znanosti – Sunce, Znanstveni rad – kako i zašto? i ostalima. Za svaki od prikazanih projekata bit će naveden uzrast učenika, opis projekta i obrazovni ishodi te organizator i poveznica na projekt. Zajedničko je svim projektnim aktivnostima veća angažiranost i interes učenika za sve nastavne predmete obuhvaćene međukurikularnim projektom, te poticanje kreativnosti u rješavanju sveobuhvatnih projektnih zadataka**.

I. UVOD

Prema Nacionalnom okvirnom kurikulumu planiranje i ostvarivanje međupredmetnih tema pridonosi međusobnom povezivanju odgojno-obrazovnih područja i nastavnih predmeta u skladnu cjelinu, a njima se razvijaju različite temeljne kompetencije učenika. Jedna od međupredmetnih tema navedena u Nacionalnom okvirnom kurikulumu je i Uporaba informacijske i komunikacijske tehnologije. IKT* je izvrsna podloga za primjenu međupredmetnog učenja i poučavanja. Suradnja s nastavnicima ostalih predmeta pridonosi sposobnosti samostalnog učenja učenika i timskog rada te razvoju komunikacijskih sposobnosti, unaprjeđuje njihov pristup rješavanju problema i istraživanju. Korištenjem informacijsko komunikacijske tehnologije i njenom primjenom u međupredmetnim temama i projektima učenici dobivaju mogućnost istraživanja i komunikacije, stjecanje vještina razmjene ideja te timskoga rada.

II. ERATOSTENOV EKSPERIMENT

Projekt Eratostenovog eksperimenta pomoću scenarija istraživačkog učenja objavljen je na ISE portalu, a informacije o tomu nalaze se na stranicama Carneta. Nakon komunikacije s organizatorima te prijave na ISE portal u Ekonomskoj školi Pula pristupili smo organizaciji projekta, podjeli zadataka i definiranju ishoda. Eksperiment je proveden 22. rujna 2015., a zbog lošeg vremena ponovljen je 23. rujna 2015. Rezultati su objavljeni na stranicama projekta.

Suradnja je ostvarena s nastavnicima geografije i matematike. Učenici su 22. i 23. rujna (dan jesenske ravnodnevice) mjerili sjenu koju stvara štap od 1 m postavljen pod pravim kutom na ravnu površinu. Na satima informatike pretraživali su podatke i naučili o Erathostenu i njegovom genijalnom načinu mjerenja opsega Zemlje, a kreirali su i priloge na web stranici škole. Nastavnica geografije uputila ih je kako odrediti točne koordinate mjesta na Zemlji gdje se nalazimo i udaljenost do ekvatora, te točno vrijeme prema GMT-u jer se mjerenje trebalo izvršiti točno u podne prema Suncu. Učenici su na satima informatike odredili točne zemljopisne koordinate naše škole pomoću Google karata i udaljenost do ekvatora. Nastavnica matematike objasnila je, pokazala i uvježbala računanje opsega ako su poznata duljina štapa, duljina sjene, zemljopisna dužina i širina te udaljenost do ekvatora. Mjerenja su bilježena i izračun se vršio na satima infomatike u MS Excelu. Učenici su imali prilike surađivati i s ostalim školama u svijetu koje su sudjelovale u eksperimentu i tako razvijati svijest o planetu Zemlji kao našem zajedničkom domu čiju sudbinu svi mi dijelimo.

III. WORLD EDUCATION GAMES - UNICEF

Projekt WEG - World Education games je UNICEF-ov projekt u suradnji s 3P Learning, a održava se svake dvije godine (sljedeći je 2017.). To je najveće na svijetu online natjecanje učenika u matematici, pravilnom pisanju („spelling“) engleskoga jezika i znanosti s porukom „Ujedinimo svijet učenjem“ koji je imao i humanitarni karakter te je aktivnošću „School in a box“ prikupljeno 100.000 USD za UNICEF i pomoć u školovanju potrebitoj djeci u svijetu. Rezultati natjecanja nalaze se na službenoj web stranici projekta [6]. Nekoliko brojki: 2015. godine sudjelovalo je 166 zemalja s ukupno 6.308.021 natjecateljem koji su ukupno proveli 128.182.795 minuta u rješavanju i odgovaranju ukupno 133.973.097 pitanja i zadataka. Iz Hrvatske je sudjelovalo 35 škola s ukupno 2.840 natjecatelja, dok je primjerice u Italiji sudjelovalo 18 škola s 5.170 natjecatelja, a u Rumunjskoj 24 škole s 3.808 natjecatelja. (Svi podaci dostupni su na stranici rezultata natjecanja).

Natjecanje je trajalo od 1. listopada do 15. studenog 2015., a odvijalo se izuzetno zanimljivo za učenike. Naime, postoji 10 razina znanja u kojima se natjecalo iz matematike, te 3 razine pravilnog pisanja engleskog jezika. Učenike je sve potrebno prijaviti što se može i skupnom prijavom popisima učenika u Excelu. Svaki učenik dobije svoje korisničko ime i lozinku u obliku kartice osobne iskaznice sa WEG zaštitnim znakom. Nakon prve, ali i svake sljedeće prijave učenici mogu oblikovati svoj „avatar“, izgled portreta koji žele, pa tako

mogu odrediti boju kože, boju kose i frizuru, nose li naočale ili ne, oblik lica i nosa i slično. Nakon odabira teme i razine natjecanja pojavljuje se karta svijeta s pokazivačem u obliku globusa koji kruži i traži suparnike koji su upravo toga trenutka odabrali istu temu (matematika, engleski, znanost) i razinu natjecanja te vrši slučajni odabir do četiri natjecatelja koji ulaze u jednu natjecateljsku sobu. Cilj je što brže i točnije odgovoriti na pitanja iz odabranog tematskog područja. Tri pogrešna odgovora diskvalificiraju natjecatelja. Nakon jedne minute natjecanje je završeno i pojavljuju se točni i netočni odgovori te proglašava pobjednik! Moguće je i međusobno natjecanje unutar iste učionice ili škole. Učenici su izuzetno motivirani i rado sudjeluju i pomažu jedni drugima u rješavanju matematičkih zadataka.

Za rješavanje pravilnog pisanja engleskog jezika potreban je zvučnik, na zaslonu se pojavljuje rečenica kojoj nedostaje jedna ili dvije riječi, a glas izgovara najprije riječi koje nedostaju, a zatim cijelu rečenicu. Moguće je odabrati američki i britanski engleski izgovor. Potrebno je pravilno na engleskom jeziku napisati riječi koje nedostaju. U školi su za to natjecanje potrebne slušalice za svakog učenika, a pitanja su učenici rješavali i kod kuće.

Na završetku online natjecanja svaki je učenik mogao sa svojim korisničkim imenom i lozinkom ispisati diplomu s brojem „odrađenih“ pitanja iz matematike i engleskoga sa zaštitnim znakom UNICEF-a i 3P Learning tima.

Učenici su mogli uspoređivati svoje i znanje učenika iz dalekih zemalja poput Malezije, Pakistana, Meksika, SAD-a, Novog Zelanda, ali i Makedonije, Belgije, Estonije i drugih što je izuzetan poticaj za bolje savladavanje gradiva iz matematike i engleskoga jezika.

IV. THIS IS OUR TIME – UNITE THE NATIONS

"This is our time" je globalni komunikacijskih projekt za osnovne i srednje škole koji je pokrenuo UNESCO ASPnet (Associated Schools Project Network), a organizira se od 1995. godine.

Koncept vremena koji je nit vodilja kroz sve aktivnosti povezana je s mnogim područjima poput znanosti, sporta, računalstva, društvenih znanosti i umjetnosti, ekonomije, medicine i religije. U okviru projekta „This is our Time“, vrijeme ima dvojako značenje:

1. This is our Time: znanstveni fenomen vremena je univerzalan i beskonačan
2. This is our Time: vrijeme kao osoban doživljaj mladih ljudi: godišnja tema u svezi vremena

Aktivnosti u sklopu projekta povezane su s jednim ili oba značenja vremena u projektu, a mogu biti kratkoročne, dugoročne, mogu se odvijati online ili offline, te imati lokalni ili globalni karakter. Primjeri aktivnosti: blogerska suradnja razreda "Classroom twinning", videokonferencije licem-u-lice, rasprave i simulacije modela Ujedinjenih naroda, interaktivno učenje u 24-satnom istraživačkom natjecanju „Unite the Nations“ i istraživanje održivog razvoja putem umjetnosti (ove godine) u Suchitoto El Salvador.

Učenici Ekonomske škole sudjelovali su u aktivnosti „Unite the Nations“ – interaktivnoj igri u kojoj učenici trebaju dublje istražiti svoju vlastitu, ali i druge kulture. Učenici predlažu pitanja višestrukog odabira iz područja geografije, povijesti, društvenih znanosti, umjetnosti i sporta iz svoje zemlje. Može se reći kako je to ustvari test o vlastitoj domovini koji su kreirali učenici, a daju ga na rješavanje ostalima u igri. Potvrđena pitanja pojavljuju se na 24-satnoj igri, s time da učenici ne rješavaju pitanja o vlastitoj zemlji.

Unite the Nations je globalna, kratkoročna, online i offline aktivnost. Željeni ishod igre je povećanje svjesnosti o različitim kulturama i zemljama diljem svijeta međukulturalnom razmjenom informacija i znanja.

Pitanja osmišljavaju učenici na nastavnim satima iz zadanih predmetnih područja, a na satima engleskoga jezika prevode ista na engleski jezik. Nakon što organizator odobri pitanja, a tijekom igre koja traje 24 sata određenog datuma (od 00:00 UTC do 23:59 UTC), učenici ne odgovaraju na pitanja iz vlastite zemlje. Cilj je što brže i što točnije odgovorit na približno 250 pitanja iz raznih područja i raznih zemalja služeći se svim klasičnim, online, offline, pa i najneobičnijim izvorima informacija (naš primjer: telefonski poziv djedu). Učenicima nije problem, dapače neobično su zainteresirani doći u 1 sat noću u školu (to je prema našem vremenu 0:00 UTC vrijeme) i natjecati se. Najbolji pritom dobivaju priznanje i nagradu. Izuzetno poticajan kviz, povezuje učenike širom svijeta, potiče i promovira misao kako je Zemlja naš veliki i jedini dom, te da svaka zemlja ma kako mala bila ima povijest i nešto čime se diči.

V. E-GRAĐANI

Projekt je osmišljen nakon uspješno obavljenog domaćeg zadatka izrade prezentacija na teme koje su započinjale slovom „E“, a jedna od njih bila je i „E-građani“. Nakon dobre prezentacije projekta učenici završnih razreda, koji će uskoro upisivati fakultet, slati zamolbe za stipendiju i / ili studentski smještaj i ostalo, shvatili su prednost projekta koji omogućava brže, jeftinije i jednostavnije dobivanje potrebnih dokumenata.

Od takvog stava do realizacije bilo je tek nekoliko koraka – nastavnici ekonomske grupe predmeta i prava s učenicima su obradili tu temu i zajedno su popunili s internetske stranice ispisane obrasce. Nakon poziva u FINA-u, dobili smo termin posjete i za jedan školski sat učenici jednog razreda imali su svoje lozinke. Naučili su što je to verifikacijska poruka, zašto je važna i zašto je važno služiti se službenom e-poštom (za učenike je to ime.prezime@skole.hr koji osigurava CARNet).

Na satima informatike neki učenici su potvrdili svoje lozinke, a naučili i kako koristiti sustav.

Poučeni dobrim primjerom i zadovoljstvom učenika i roditelja (koji su bili upoznati s projektom i dali svoj pristanak) i ostali nastavnici organizirano su poveli učenike, pa su svi učenici 3. i 4. razreda Ekonomske škole Pula postali E-građani.

VI. FESTIVAL ZNANOSTI – SUNCE

U 2015. godini održan je Festival znanosti s temom Sunce. Uključili smo učenike drugih razreda koji su

izrađivali online prezentacije na prezi.com, a čije su poveznice bile javno objavljene na web stranici škole.

Isprva su teme bile vezane za Sunce kao središte našeg sustava s mnoštvom podataka o udaljenosti, masi, temperaturi i slično. Ali nakon što su se uključili i nastavnici ostalih predmeta „pojavile“ su se i druge zanimljive teme, primjerice: Sunce kao izvor D vitamina, Kako se pametno sunčati, Djeluju li zaista zaštitne UV kreme, Kako izabrati dobre sunčane naočale, Što je to UV indeks zračenja, Sunce kao izvor života, pa do Suncem protiv depresije - mogućnost turističke ponude u Istri. Poveznice za prezentacije objavljene su na web stranicama škole [7] gdje su ih još uvijek mogu pregledati učenici, nastavnici, pa i šira javnost.

Suradnja je ostvarena s predmetima: biologija, kemija, ekonomska grupa predmeta

VII. ZNANSTVENI RAD – KAKO I ZAŠTO?

Tematska cjelina koja obuhvaća klasično i online pretraživanje podataka o literaturi, pravilno navođenje literature u radovima, posjet Sveučilišnoj knjižnici u Puli i Nacionalnoj i sveučilišnoj knjižnici u Zagrebu namijenjena je učenicima 2., 3. i 4. razreda, a ostvaruje se u suradnji sa školskim knjižničarom.

Obrazovni ishodi obuhvaćaju snalaženje u klasičnim i online katalozima knjižnica, poznavanje pravila rada sveučilišnih knjižnica i vrste građe koju nude, poznavanje mogućnosti online baza podataka sa stručnim člancima u punom tekstu dostupnim u PDF formatu (i što je PDF format), zatim zavičajna povijest i kultura, uključujući i teme iz građanskog odgoja.

Učenici uvježbavaju pretraživanje literature i svoja znanja koriste priliko pisanja referata iz informatike i drugih predmeta, kao i za pisanje završnog rada.

Upisom fakulteta učenici se i dalje javljaju sa zahvalom za naučeno, što predstavlja zaista veliko zadovoljstvo i poticaj za daljnje usavršavanje tematske cjeline.

VIII. ZAKLJUČAK

Informatika svojim širokim obuhvatom tema predstavlja idealnu međurukikularnu kariku koja može i treba povezivati vlastita i formativna znanja ostalih predmeta s praktičnim znanjima. Postoje mnogobrojne mogućnosti projektne nastave i obuhvata tema iz raznih područja odgoja i obrazovanja. Informacijsko-komunikacijska tehnologija treba biti upravo to – karika za povezivanje raznih i raznolikih znanja u cjelinu koju će učenici lakše prihvatiti uz modernu tehnologiju koju izuzetno spretno koriste, no ponekad ih treba usmjeriti prema korisnim i odgojno-obrazovnim temema.

LITERATURA

[1] http://www.azoo.hr/images/stories/dokumenti/Nacionalni_okvirni_kurikulum.pdf (15. 2. 2016.)

[2] http://www.carnet.hr/ise/eratostenov_eksperiment_2015 (3. 9. 2015.)

[3] http://eratosthenes.ea.gr/en/content/eratosthenes-online-lesson-contest (15. 9. 2015.)

[4] http://www.3plearning.com/worldeducationgames/unicef/ (10. 9. 2015.)

[5] www.worldeducationgames.com (10. 9. 2015.)

[6] http://west.cdn.mathletics.com/Weg-2015/WEG_Results_2015_Final.pdf (14. 10. 2015.)

[7] http://ss-ekonomska-pu.skole.hr/nastava/informatika (14. 4. 2016.)

Tradicionalni ili hibridni model nastave računalstva

*Martina Sertić, **Krešimir Šolić
*Industrijsko-obrtnička škola Slavonski Brod, Republika Hrvatska, sertic_79@yahoo.com
**Medicinski fakultet Osijek, Republika Hrvatska, kresimir@mefos.hr

Sažetak – Cilj rada je bio dokazati kako primjena elektroničkog učenja u obrazovanju omogućava veću povezanost teoretskog znanja i praktične primjene naučenog. Pretpostavilo se da bi hibridni model učenja trebao biti učinkovitiji od tradicionalnog modela učenja. Provedeni eksperiment dokazao je održivost hipoteze po kojoj elektroničko učenje poboljšava praktičnu primjenu naučenog i povećava razinu znanja. Iz analize podataka istraživanja došlo se do zaključka da korištenje hibridnog modela učenja u nastavi informatike povećava i pospješuje znanje učenika te povećava razinu znanja i zadovoljava učenikove obrazovne mogućnosti.

I. UVOD

Elektroničko učenje rastući je trend na mnogim visokim učilištima i obrazovnim institucijama. Učenike i nastavnike privukle su mnoge prednosti elektroničkog učenja, kao što su: fleksibilnost učenja bilo gdje, u bilo koje vrijeme i to na individualnoj razini. Pitanje učinkovitosti elektroničkog učenja treba postaviti prije usvajanja elektroničkog učenja u velikim razmjerima, kao nečeg što ide ukorak s budućnošću. Znanje modrenog čovjeka svakim danom postaje sve veće i to ponajviše zahvaljujući razvoju tehnologije o kojoj ovisimo.

Obrazovni sustav, kao i svaki drugi sustav modernih društava, postao je dio aparata znanja koji u tehnologiji vidi budućnost i moć naše civilizacije. Dio tog sustava su učenici, nastavnici, znanstvenici, službenici. Obrazovanje je postalo dio ljudskih života, od rođenja do smrti, s jednim osnovnim ciljem, a to je učinkovitost u onome što se radi, proizvodi i nudi.

II. UČENJE

Postoji više teorija o učenju, a među njima se izdvajaju biheviorizam, kognitivizam, konstruktivizam i humanizam. Sve ove teorije se međusobno nadopunjuju, a u praksi nerijetko i preklapaju. Detaljnim proučavanjem načina na koji ljudi uče, ustanovljeno je da različite osobe imaju i različite stilove učenja.[1]

Neki učenici više vole:

- razgovor s nastavnicima,
- razmjenu iskustava,
- aktivno sudjelovanje u učenju,
- postavljanje konkretnih pitanja,
- dok neki jednostavno slušaju predavanja nastavnika i na taj način uče. [2]

Prema istraživanjima, većina učenika će kombinirati nekoliko načina učenja, a dokazano je da je takav način i najuspješniji.[1]

Moguće prednosti elektroničkog učenja su višestruke:

- može objediniti sve teorije učenja i sve stilove učenja,
- potaknuti želju za kvalitetnijim učenjem,
- omogućiti korištenje različitih metoda,
- potaknuti korištenje različitih medija.[3]

A. Tradicionalna nastava

Kod tradicionalnog oblika nastave prisutni su određeni elementi koji ne postoje kod obrazovanja utemeljenog na uporabi informacijsko-komunikacijske tehnologije.[4]

Neki od elementa koji su vrlo bitni kod tradicionalnog oblika nastave, a ne susrećemo ih kod elektroničkog obrazovanja su:

- u tradicionalnoj verziji nastave učenici mogu izravno komunicirati s predavačem,
- u ovakvom procesu nastave razvija se i socijalizacija, učenici međusobno razgovaraju, upoznaju se i postavljaju pitanja jedni drugima. [4]

B. Hibridna nastava

Hibridna nastava (engl. *blended learning, hybrid learning)* je izvođenje obrazovnih programa djelomično *online* uz pomoć Interneta, a djelomično na tradicionalan način. Ovaj oblik elektroničkog učenja danas je sve popularniji i ponajviše se koristi u poslovnom okruženju.[5]

Vrijeme koje se kod tradicionalnog učenja provodi u učionici znatno je smanjeno, ali nije sasvim eliminirano.

Cilj hibridnog modela učenja je povezivanje najboljih osobina tradicionalnog učenja s najboljim elementima *online* učenja. Na ovaj način želi se stvoriti proces aktivnog učenja i poboljšati kvalitetu rada.[5]

Glavne prednosti hibridnog modela učenja u odnosu na tradicionalno učenje su višestruke:

- učenje prilagođeno učeniku,
- interaktivne sadržaje,
- ujednačena razina znanja kao moguća opcija kod većine polaznika,
- fleksibilnost programa pruža i fleksibilniji pristup predavača,
- uvijek dostupni sadržaji za polaznike,
- mogućnost korištenja različitih tehnologija.[6]

Glavni nedostatci hibridnog modela učenja u odnosu na tradicionalno učenje su:

- nema individualnog kontakta među sudionicima,
- polaznici koji nisu navikli na *online* rad,
- velika ulaganja u opremu,
- velika ulaganja u *online* materijale,
- nedostatak međuljudske interakcije. [6]

III. MJERENJE ZNANJA

A. Načini ocjenjivanja

U ovom djelu rada prikazana su dva načina mjerenja rezultata na pismenom testu. Prvi način je postupak mjerenja, tj. pretvaranja broja bodova u ocjene prema unaprijed utvrđenim pravilima. Drugi postupak je jedan od načina naknadnog utvrđivanja tih mjerila tj. pravila.

Korištenjem prvog načina, postupkom utvrđivanja broja bodova unaprijed, možemo izmjeriti dobivene rezultate na dva načina. Ako pri ispravljanju pismenog ispita u zadatcima bodujemo korake i mjerimo svaki važan međurezultat koji potom donosi određeni broj bodova, tada će se primijeniti ljestvica prikazana tablicom I. Analizom tablice I. može se zamijetiti da je za ocjenu dovoljan (2) potrebno ostvariti najmanje 50% bodova, dok je za ocjenu odličan (5) potrebno ostvariti 90% od svih mogućih bodova.[7]

Tablica I. Pretvaranje broja bodova u ocjene bodujući međurezultate

Broj bodova u postotku (%)	manje od 50%	od 50 do 63%	od 64 do 76%	od 77 do 89%	90 i više %
Ocjena	Nedovoljan (1)	Dovoljan (2)	Dobar (3)	Vrlo dobar (4)	Odličan (5)

Idući primjer postupka jest utvrđivanje broja bodova unaprijed, a provodi se na način da se boduju samo točni odgovori, tj. točna rješenja, bez bodovanja međurezultata. Na ovaj način rezultate, tj. odgovore na postavljena pitanja vrednujemo "*ili je nešto točno ili je netočno*". Kod ovog načina bodovanja moguće je primjenjivati blaži kriterij prikazan određenom ljestvicom u tablici II. [8]

Tablica II. Ljestvica pretvaranja broja bodova u ocjene bodujući "ili je točan ili je netočan"

Broj bodova u postotku (%)	manje od 40%	od 40 do 55%	od 56 do 70%	od 71 do 85%	86 i više %
Ocjena	Nedovoljan (1)	Dovoljan (2)	Dobar (3)	Vrlo dobar (4)	Odličan (5)

Uz prikazane dvije ljestvice za pretvaranje postignutih bodova u ocjene svakako je potrebno navesti sljedeću napomenu. Navedene bodovne granice za pojedinu ocjenu samo su orijentacijske, jer davanje ocjena nije statistički postupak. Kod mjerenja znanja, tj. bodovanja na testu, najbolje je da profesor napravi ljestvice za pretvaranje broja bodova u ocjene, uzimajući u obzir sve čimbenike poput složenost zadatka i programa, uvjeta rada i opreme s kojom se radi, težinu pitanja u ispitnom materijalu, te karakteristike razrednog odjeljenja.[8]

Samim postupkom ispravljanja pismenog ispita i analizom rezultata moguće je provjeriti primjenjivost i efikasnost. Analizom rezultata moguće je zaključiti je li ispitni materijal bio primjeren ili pak prelagan ili pretežak. Kod približno normalne razdiobe analiziranih podataka ispitni materijal je primjeren određenoj razini znanja. [7]

Osim gore navedenih postupaka izračunavanja ocjena iz broja bodova, moguće je izračunati, tj. izmjeriti bodove jednim od načina naknadnog utvrđivanja mjerila za ljestvicu ocjena. Pri tome se koriste vrijednost *aritmetičke sredine* (u oznaci X) i *standardne devijacije* (u oznaci σ).

Prvo se iz određenih skupina istraživanja izračuna aritmetička sredina, a zatim i standardna devijacija i to ovim postupkom:

- za ocjenu nedovoljan (1): od 0 do ($\bar{X}$ - 1.5σ)
- za ocjenu dovoljan (2): od ($\bar{X}$ - 1.5σ) do ($\bar{X}$ - 0.5σ)
- za ocjenu dobar (3): od ($\bar{X}$ - 0.5σ) do ($\bar{X}$ + 0.5σ)
- za ocjenu vrlo dobar (4): od ($\bar{X}$ + 0.5σ) do ($\bar{X}$ + 1.5 σ)
- za ocjenu odličan (5): Iznad ($\bar{X}$ + 1.5σ). [8]

Koja je od navedenih metodologija najpouzdanija i najefikasnija, teško je odrediti. Primjena bilo koje od

navedenih metodoloških postupaka mjerenja broja bodova ovisi o više čimbenika. [8]

"Kada su rezultati nekog ispita znanja pravilno raspoređeni, onda kažemo da se radi o normalnoj distribuciji, a ispitni materijal je primjeren učenikovom znanju. Ako ispitni materijal vjerno odražava zadani nastavni program, onda možemo zaključiti da je i taj program primjeren učeničkim mogućnostima".[9]

B. Postupak mjerenja znanja

U istraživanju prikazanom u ovom radu koristio se postupak naknadnog utvrđivanja mjerila za skalu ocjena.

U ovom načinu mjerenja korištena je *aritmetička sredina* (u oznaci *X*) i *standardna devijacija* (u oznaci σ).

Koristeći se izračunatim podacima, došlo se da skale ocjena koja je prikazana u tablici III.

Formule koje su korištene za izračun postotnih vrijednosti za pojedinu ocjenu po broju ostvarenih bodova kod ispitanika su:

- za ocjenu nedovoljan (1): od 0 do ($\bar{X}$ - 1.5σ)
- za ocjenu dovoljan (2): od ($\bar{X}$ - 1.5σ) do ($\bar{X}$ - 0.5σ)
- za ocjenu dobar (3): od ($\bar{X}$ - 0.5σ) do ($\bar{X}$ + 0.5σ)
- za ocjenu vrlo dobar (4): od ($\bar{X}$ + 0.5σ) do ($\bar{X}$ + 1.5 σ)
- za ocjenu odličan (5): Iznad ($\bar{X}$ + 1.5σ)

Koristeći se navedenim postupkom, došlo se i do skala ocjena koje su prikazane u tablicama IV, V i tablici VI.

C. Provjera znanja

Provjera znanja provedena je testiranjem učenika i to dva puta. Na početku predavanja kroz predtest i na kraju predavanja kroz posttest. Učenicima su postavljena pitanja iz nastavne cjeline Microsoft Excel. Test se sastojao od 34. pitanja. Testom je obuhvaćena srednjoškolska populacija učenika.

Uzorak koji je uzet iz srednjoškolske populacije bili su učenici prvih i drugih razreda koji su pohađali nastavu računalstva. Broj žena i muškaraca u istraživanju bio je podjednak. Veličina ukupnog uzorka bila je n=90. U kontrolnoj skupini veličina uzorak bila je n=45, a u eksperimentalnoj skupini veličina uzorka bila je n=45.

Starosna dob svih ispitanika bila je 15,63 ± 0,57 godina (aritmetička sredina ± standardna devijacija).

Prosječna dob učenika bila je 16 godina, a u istraživanju su podjednako bila zastupljena oba spola (muških ispitanici = 48,88%, ženski ispitanici = 51,11%).

Prije početka eksperimenta provedeno je testiranje predznanja u eksperimentalnoj i kontrolnoj skupini, a na završetku istraživanja provjereno je stečeno znanje u obje skupine.

Nakon završenih predavanja, učenici su pristupili testiranju znanja koje je provedeno kroz kontrolnu i eksperimentalnu skupinu. Obje skupine rješavale su isti tip zadataka objektivnog tipa kao i u predistraživanju.

Na taj način utvrđena je veća učinkovitost elektroničkog učenja (hibridnog modela) s obzirom na tradicionalnu nastavu.

Provedeni eksperiment dokazao je održivost hipoteze po kojoj elektroničko učenje poboljšava praktičnu primjenu naučenog i povećava razinu znanja.

D. Primjeri zadataka testa znanja

Neki od zadataka testa znanja izdvojeni su i prikazani su u ovom djelu rada.

Zadatak 4.

Definiraj i objasni navedene pojmove:

a) Funkciju
b) Napiši primjer funkcije u Excelu
c) Apsolutnu adresu ćelije
d) Navedi primjer apsolutne adrese ćelije
e) Relativnu adresu ćelije
f) Navedi primjer relativne adrese ćelije

Zadatak 8.

U ovoj rečenici krije se pogreška. Pronađi pogrešku i na crti ispravno napiši rečenicu.

Program Excel sastoji se od slajdova i u njemu možemo raditi grafikone i potrebne proračune.

_____

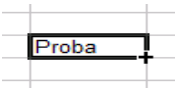
_____

Zadatak 12.

Što označavaju ove ikone?



_____



_____



_____

Zadatak 17.

Funkcija POWER spada u logičke funkcije programa MS Excel.

DA NE_____

Zadatak 19.

Funkcija MOD spada u matematičke funkcije programa MS Excel.

DA NE_____

Zadatak 28.

U programu Excel svaka formula započinje znakom _____ iza toga znaka dolazi _____

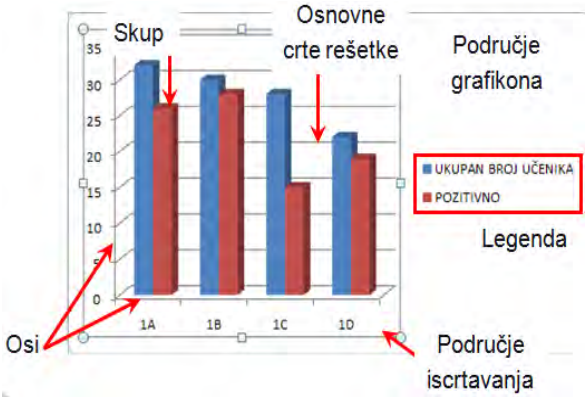
Zadatak 30.

Filtrirati podatke u bazi podataka znači:

a) poredati podatke po nekom redoslijedu
b) poredati podatke po abecedi
c) dodati određeni podatak u skupinu podataka
d) dodati i poredati podatke po abecedi
e) izdvojiti određeni podatak iz skupine podataka

Zadatak 33.

Pravilno označi elemente grafikona.



E. Rezultati mjerenja znanja

Analizom rezultata testa znanja za oba istraživanja i za sve skupine dobiveni su idući rezultata koji su u daljnjem tekstu prikazani tablično i grafički. (Tablica III, IV, V, VI i VII, te grafikon I).

Tablica III. Ostvarene vrijednosti na testu za kontrolnu skupinu predistraživanja

Bodovi u postotcima	Ocjena
0,00% - 11,33 %	Nedovoljan (1)
11,34 % - 27,12 %	Dovoljan (2)
27,13 % - 42,91 %	Dobar (3)
42,92 % - 58,71 %	Vrlo dobar (4)
58,72 % – sve ostalo	Odličan (5)

Tablica IV. Ostvarene vrijednosti na testu za eksperimentalnu skupinu predistraživanja

Bodovi u postotcima	Ocjena
0,00% - 10,31 %	Nedovoljan (1)
10,32 % - 27,06 %	Dovoljan (2)
27,07 % - 43,81 %	Dobar (3)
43,82 % - 60,56 %	Vrlo dobar (4)
60,57 %– sve ostalo	Odličan (5)

Tablica V. Ostvarene vrijednosti na testu za kontrolnu skupinu glavnog istraživanja

Bodovi u postotcima	Ocjena
0,00% - 12,04%	Nedovoljan (1)
12,05% - 31,65%	Dovoljan (2)
31,66% - 51,27%	Dobar (3)
51,28% - 70,88%	Vrlo dobar (4)
70,89% – sve ostalo	Odličan (5)

Tablica VI. Ostvarene vrijednosti na testu za eksperimentalnu skupinu glavnog istraživanja

Bodovi u postotcima	Ocjena
0,00% - 15,69 %	Nedovoljan (1)
15,70 % - 39,36 %	Dovoljan (2)
39,37 % - 63,03 %	Dobar (3)
63,04 % - 86,70 %	Vrlo dobar (4)
86,71 % – sve ostalo	Odličan (5)

Poznato je i dokazano kroz razna istraživanja koja se bave analizom podataka inteligencije svjetske populacije da većina ljudi spada u prosječnu skupinu, dok je manji broj ljudi ili visoko inteligentan ili ispodprosječno inteligentan. Kada bi bila korištena Gaussova skala mjerenja dobili bi otprilike slične rezultate.

Korištenjem postupka naknadnog utvrđivanja mjerila za skalu ocjena prema aritmetičkoj sredini i standardnoj devijaciji, te iz te skale izračunom dobivenih postotnih i kasnije standardnih vrijednosti bodova za svaku pojedinu ocjenu došlo se do završne faze raspodjele dobivenih rezultata u tri kategorije znanja.

Prva kategorija za sva istraživanja i skupine su nadprosječni učenici, druga kategorija su prosječni učenici i treća kategorija su ispodprosječni učenici.

Zadnja tablica ovog poglavlja predstavlja usporedbu svih podataka istraživanja. U njoj je moguće iščitati podatke za MAX i MIN vrijednosti oba istraživanja i svih skupina istraživanja po kategorijama nadprosječni, prosječni i ispodprosječni učenici (Tablica VII).

Tablica VII. Prikaz vrijednosti svih istraživanja

	Nadprosječni učenici	Prosječni učenici	Ispodprosječni učenici
MAX-k1-	63	46	9
MIN-k1-	48	12	2
MAX-e1-	67	49	9
MIN-e1-	50	21	2
	Nadprosječni učenici	**Prosječni učenici**	**Ispodprosječni učenici**
MAX-k2-	69	58	10
MIN-k2-	60	16	1
MAX-e2-	80	71	8
MIN-e2-	72	13	3

U tablici VII prikazane su MAX i MIN vrijednost u provedenim istraživanjima mjerenja učinkovitosti elektroničkog učenja u nastavi računalstva.

U prikazanoj tablici VII vidljive su ove vrijednosti pod oznakom:

MAX-k1- koja označava najveći ostvareni rezultat u kontrolnoj skupini predistraživanja za nadprosječne učenike, prosječne učenike i ispodprosječne učenike.

MIN-k1- koja označava najmanji ostvareni rezultat u kontrolnoj skupini predistraživanja za nadprosječne učenike, prosječne učenike i ispodprosječne učenike.

MAX-e1- koja označava najveći ostvareni rezultat u eksperimentalnoj skupini predistraživanja za nadprosječne učenike, prosječne učenike i ispodprosječne učenike.

MIN-e1- koja označava najmanji ostvareni rezultat u eksperimentalnoj skupini predistraživanja za nadprosječne učenike, prosječne učenike i ispodprosječne učenike.

MAX-k2- koja označava najveći ostvareni rezultat u kontrolnoj skupini glavnog istraživanja za nadprosječne učenike, prosječne učenike i ispodprosječne učenike.

MIN-k2- koja označava najmanji ostvareni rezultat u kontrolnoj skupini glavnog istraživanja za nadprosječne učenike, prosječne učenike i ispodprosječne učenike.

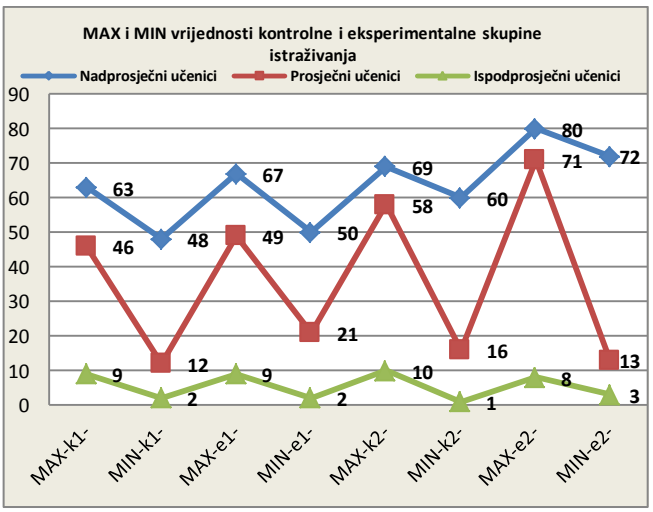
MAX-e2- koja označava najveći ostvareni rezultat u eksperimentalnoj skupini glavnog istraživanja za nadprosječne učenike, prosječne učenike i ispodprosječne učenike.

MIN-e2- koja označava najmanji ostvareni rezultat u eksperimentalnoj skupini glavnog istraživanja za nadprosječne učenike, prosječne učenike i ispodprosječne učenike.

Gore navedeni podatci u tablicama istraživanja za obje skupine kontrolnu i eksperimentalnu prikazani su i grafikonom I.

Iz grafikona se mogu iščitati podatci za MAX i MIN ostvareni broj bodova za svaku od tri kategorije. Na osi y koordinatnog sustava grafikona prikazani su ostvareni bodovi na testu, a na osi x MAX i MIN funkcije vrijednosti za kontrolnu i eksperimentalnu skupinu glavnog istraživanja.

Grafikon I. MAX i MIN vrijednosti za kontrolnu i eksperimentalnu skupinu svih istraživanja



Detaljnom analizom grafikona I vidljivo je da plava linija grafikona predstavlja nadprosječne učenike, crvena prosječne učenike, a zelena ispodprosječne učenike.

Najveća izlazna vrijednost kod plave linije je u točci 80 koja predstavlja najbolji rezultat tj. najbolje riješeni test. Ova vrijednost prema osi x grafikona spada pod oznaku MAX-e2- tj. maksimalnu vrijednost glavnog istraživanja eksperimentalne skupine. Vrijednost 69 sa oznakom MAX-k2- također se nalazi na plavoj liniji i predstavlja najveći ostvareni rezultat, najbolje riješeni test glavnog istraživanja kontrolne skupine.

Iz ove dvije vrlo značajne vrijednosti za istraživanje potvrđena je valjanost postavljene hipoteze ovog rada po kojoj elektroničko učenje poboljšava praktičnu primjenu naučenog i povećava razinu znanja.

Ovdje je potrebno još istaknuti da su vrijednosti na ostalim linijama očekivane i za MAX i MIN oznake oba istraživanja i obje skupine u prosječnoj i ispodprosječnoj kategoriji.

Vrijednosti zelene linije koja predstavlja ispodprosječne učenike su podjednake za sva istraživanja, a vrijednosti crvene linije prosječnih učenika, su najviše kod glavnog istraživanja eksperimentalne skupine 71 bod i 58 bodova kod kontrolne skupine glavnog istraživanja.

Iz analize podataka istraživanja došlo se do zaključka da korištenje hibridnog modela učenja u nastavi informatike povećava i pospješuje znanje učenika te povećava razinu znanja i zadovoljava učenikove obrazovne mogućnosti.

IV. ZAKLJUČAK

Moderno suvremeno doba nudi slobodnu i brzu komunikaciju, dostupnost i umreženost, aplikacije za svaki problem koji treba riješiti. Uči ljude kako efikasno i efektivno razmišljati u domeni učinkovitosti i brzog učenja, pronalaženja rješenja i povećavanja dobiti, svladavanja problema na najjednostavniji i najproduktivniji način.

Efikasnost je ono što traži svaki današnji sustav, pa tako i obrazovni. Obrazovni sustav koji školuje mlade ljude i uči ih kako pravilno razmišljati i djelovati u društvu znanja.

Kroz provedeno istraživanje dokazano je da primjena informacijsko-komunikacijske tehnologije u nastavi modelom hibridnog učenja povećava i pospješuje znanje učenika. Korištenjem navedenog modela učenja obrazovanje postaje za jedan korak bliže težnjama mladih ljudi današnjice.

Društvo znanja i društvo tehnike glavni je cilj i hrvatskog obrazovnog standarda. Cilj u kojem i male ekonomije poput naše žele postati dio globalnog svijeta, dio Europske unije kojoj prostorno pripadamo, iz koje pokušavamo učiti i posuditi ono što bi nas učinilo boljim i konkurentnijim na svjetskom tržištu znanja. Tržištu koje inzistira na povećanju praktičnih sposobnosti i primjene naučenog u školama, fakultetima i ostalim obrazovnim institucijama.

LITERATURA

[1] A. Slavić, "Semantički web, sustavi za organizaciju znanja i mrežni standardi." Zagreb: Zavod za informacijske studije, 2005.

[2] N. Breslauer, "Obrazovanje uz pomoć informacijsko-komunikacijskih tehnologija."
Čakovec: Međimursko veleučilište, 2011.

[3] N. Buzzetto-More i A. J. Alade, "E-nastava i e-procjena znanja." Edupoint, 7(53). Zagreb, 2007, pp. 28-32.

[4] T. Burger, "Percepcija i stavovi studenata prema učenju putem Interneta." Edupoint 3(16). Zagreb 2006.
Preuzeto s http://edupoint.carnet.hr/casopis/broj-16/clanak-01/burger.pdf.

[5] C. J. Bonk i C. R. Graham, Handbook of blended learning: "Global Perspectives, local designs." San Francisco, CA: Pfeiffer Publishing, 2006.
Preuzeto s
http://www.click4it.org/images/a/a8/Graham.pdf.

[6] J. Ledić i V. Mornar, "iProjekt metodika nastave informatike online." Katedre za e-obrazovanje Filozofskog fakulteta u Rijeci, Zavod za primijenjeno računarstvo Fakulteta elektrotehnike i računarstva u Zagrebu, 2010.
Preuzeto s
http://ahyco.uniri.hr/portal/OProjektu.aspx.

[7] V. Zorić, "Provjera i ocjenjivanje znanja učenika u nastavi filozofije u srednjoj školi." Život i škola, 20 (56), 2008, pp. 113-126. Preuzeto s:
http://hrcak.srce.hr/36906.

[8] D. Brajković i S. Kadum, "Praćenje, provjeravanje i ocjenjivanje učenika u nastavi." Metodički obzori, 2 (2), 2007, pp. 36-51.
Preuzeto s:
http://hrcak.srce.hr/index.php?show=clanak&id_clanak_jezik=30415.

[9] I. Furlan, "Upoznavanje, ispitivanje i ocjenjivanje učenika." Zagreb: Pedagoško-književni zbor, 1966.

Provjere znanja pomoću Classroom Managera u učionicama budućnosti

Milan Korać, dipl. ing.
Tehnička škola Ruđera Boškovića, Zagreb
Getaldićeva 4, Zagreb, Republika Hrvatska
Telefon: 01-2371 061, e-mail: milan.korac@zg.t-com.hr

Sažetak - Dio škola u Republici Hrvatskoj je opremljen pametnim učionicama odnosno učionicama budućnosti. Opremljene su tabletima uz programsku potporu *HP Classroom Managera (CMS)*. Učionice se ne koriste u svim dostupnim mogućnostima zbog nepoznavanja mogućnosti *CMS* sustava od strane nastavnika. Jedna od vrlo praktičnih primjena navedenog sustava je široka lepeza provjera znanja učenika u svim oblicima, od izrade i provedbe različitih testova do interaktivne provjere nastavnikovim pitanjima i odgovorima učenika u različitim oblicima pomoću tableta. Također je dostupno i anketiranje učenika. Cilj ovog rada je detaljan opis svih mogućnosti provjera znanja učenika pomoću *CMS-a* kao i opis postupaka izrade i provedbe ispita. U radu su opisana i iskustva iz Tehničke škole Ruđera Boškovića na strukovnim računalnim predmetima.

I. UVOD

CMS omogućuje nastavnicima povećanje učinkovitosti prijenosa znanja korištenjem tehnologije. Rješenje sadrži niz alata i mogućnosti koje su razvijene u suradnji upravo s nastavnicima kako bi se pojednostavilo korištenje te omogućila što bolja integracija u nastavu. Nastavnici i njihovi asistenti mogu upravljati uređajima učenika i voditi ih kroz nastavni sadržaj s jednog mjesta tj. sa svojeg centralnog uređaja. Ujedno mogu pomoći učenicima ostati fokusirani na predavanje uz pomoć niza funkcionalnosti poput nadzora ekrana učeničkih uređaja, kontrole korištenja aplikacija, internet stranica, ispisa i drugih funkcionalnosti. Kao pomoć svladavanju sadržaja nastavnici i učenici mogu kreirati studentski dnevnik u digitalnom obliku te koristit čitav niz alata za ciljano testiranje učenika kao što su kviz znanja, ankete učenika, pitanja i odgovori i dr. školama.

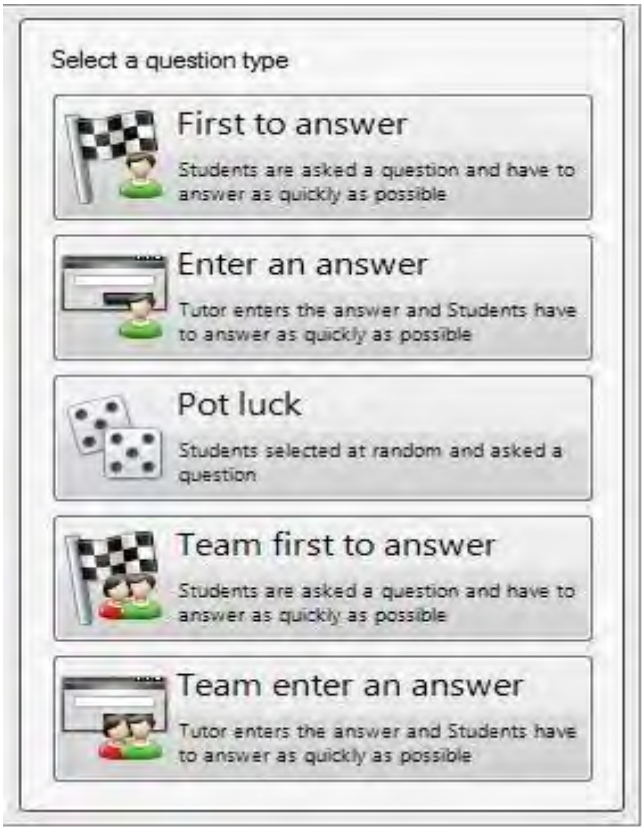
CMS osim upravljanja učionicom, kao jedinstveno rješenje za upravljanje na nivou škole, daje mogućnost i tehničarke konzole za mrežne i sistemske administratore kako bi po potrebi pružali podršku korisnicima i upravljali uređajima cijele škole. Tehničarska konzola omogućuje pogled na sva računala u jednom ekranu, nadzor aplikacija i Interneta na svim uređajima, generiranje izvještaja o hardveru i softveru koji se koristi na uređajima, kopiranje podataka na određene ili sve uređaje i dr. *CMS* rješenje omogućuje što jednostavnije korištenje Microsoftovih alata, platformi i servera za potrebe učenja. Cilj je izgraditi zajednicu predavača koji mogu razmjenjivati iskustva i materijale, kako bi podržali korištenje ove tehnologije.

U školama se učestalo mogu naći različita miješana okruženja u kojima se nalaze stolna računala, prijenosna računala i tableti što otežava kreiranje plana nastave i nastavnog sadržaja te uporabu opreme tijekom nastave. Uz navedene izazove, učenici donose svoje uređaje koje bi žele koristiti za praćenje nastavnog sadržaja što stvara dodatne probleme. *CMS* predstavlja jedinstveno i centralizirano rješenje za sve vodeće platforme tj. operacijske sustave (Windows, Android, Mac, iOS, Linux, Chrome OS i Kindle).

CMS je izgrađen na funkcionalnostima koje pomažu stvoriti najbolji način prijenosa znanja na učenike. Rješenje je jedinstveno jer objedinjuje učenički digitalni dnevnik zajedno s funkcionalnošću pitanja i odgovora za brzo i učinkovito ispitivanje i ocjenjivanje učenika.

II. OCJENJIVANJE: PITANJA I ODGOVORI

Funkcionalnost *Pitanja i odgovori* (Slika 1 Pitanja i odgovori) je namjenski kolaboracijski alat koji nastavnicima omogućava davanje naglaska na ključne točke predavanja kao i mjerenje i ocjenu razumijevanja nastavnog sadržaja učenika još tijekom samog predavanja.



Slika 1 Pitanja i odgovori

Samo rješenje prati najbolju praksu za kontinuiranom provjerom znanja omogućavajući nastavnicima usmeno postavljanje pitanja učenicima kako bi provjerili razumijevanje nastavne materije, uključivanje novih načina ispitivanja (npr. *„basketball questioning"*) te praćenje nagrada za pojedince i po potrebi za grupe učenika.

Načini ispitivanja mogu biti sljedeći:

A. Najbrži odgovor

Nakon postavljenog pitanja učenicima, nastavnik definira vrijeme potrebno za odgovor učenika. Nakon isteka vremena za razmišljanje, učenici koji znaju odgovor prijavljuju se putem svojih uređaja dok nastavnik na svom uređaju prati redoslijed prijava. Nastavnik zatim odabire učenike koji su se najbrže prijavili te postavlja pitanje učeniku koji se prvi prijavio. Odgovor može biti odmah ocijenjen kao točan i netočan. Po potrebi, pitanje o točnosti odgovora može biti proslijeđeno sljedećem prijavljenom učeniku. Nakon točnog odgovora učenik ili učenici mogu biti nagrađeni zvjezdicom ili im prethodno zarađena zvjezdica može biti oduzeta.

B. Unesi odgovor

Nastavnik u sustav za upravljanje nastavom unosi odgovor na pitanje koje će biti postavljeno učenicima. Nakon postavljenog pitanja, učenici na svojim uređajima unose odgovor. Rezultat ispravnosti odgovora prikazuje se u realnom vremenu na nastavničkom uređaju. Učenik ili učenici mogu biti nagrađeni zvjezdicom ili im prethodno zarađena zvjezdica može biti oduzeta.

C. Nasumični odabir

Sustav nasumično odabire učenike kojima nastavnik zatim može postaviti pitanje.

D. Grupni rad

Učenici se grupiraju kako bi im se omogućio timski rad ali i timsko natjecanje za nagradu - zvjezdicu. Članovi grupa mogu biti nasumično odabrani ili se ostavlja učenicima mogućnost prijave u grupu čijim članom žele biti. Svaki učenik unosi svoj odgovor na postavljeno pitanje, a pobjednik je grupa koja ima najviše točnih odgovora što je i vidljivo u realnom vremenu na nastavničkom uređaju.

E. Ocjenjivanje kolega

Učenike se potiče dati svoje mišljenje na odgovor koji je na netom postavljeno pitanje dao njihov kolega. To čine odabirom točnog ili netočnog odgovora na svojem uređaju. Nastavnik na svom uređaju vidi rezultate u realnom vremenu.

Postavljanje pitanja na satu pomoću *HP Classroom Managera* bilo bi korisno kod jednostavnijih odgovora i kada bi se na brzinu htjelo provjeriti znanje svih ili određenih učenika na postavljeno pitanje. Nedostatak je vrijeme potrebno da se to postavi.

Prednost kod ovakvih provjera je da se na ovaj način mogu aktivirati svi učenici bez javljanja. To je dobar način realne provjere znanje zato što neki učenici često znaju odgovor ali se ne žele javiti jer su povučeni ili sramežljivi. Također, ovaj način je zabavniji i više motivirajući za učenike.

III. ANKETE UČENIKA

Jednostavno anketiranje učenika omogućava brz i pouzdan način provjere usvojenosti znanja učenika vezanih za odrađeni nastavni sadržaj. Ankete se mogu kreirati koristeći postojeće predloške, odnosno već unaprijed definirana pitanja i odgovore što omogućava brzo postavljanje pitanja te trenutni uvid u sve odgovore i sažetak / izvještaj za cijeli razred. Shodno dobivenim rezultatima moguće je kreirati grupe učenika. Jednako tako, rezultate je ankete moguće prezentirati cijelom razredu.

Ankete su korisne nastavnicima za dobivanje povratne informacije od učenika za određena predavanja, lekcije i sl. Iako se to također može dobiti jednostavnim usmenim postavljanjem pitanja i čekanja učenika da podignu ruke za ono s čime se slažu, uvijek postoje učenici koji je sudjeluju iz inih razloga. Ovaj način ispitivanje učenika daje stvaran stav učenika bez ikakvih pristranosti ili utjecaja ostatka razreda.

IV. TESTOVI ZNANJA I KVIZOVI ZA UČENIKE

Aplikacija za kreiranje testova znanja omogućava kreiranje pitanja i odgovora koja mogu uključivati tekst, slike, audio i video. Sama se pitanja pohranjuju u bazu podataka koju je naknadno moguće dijeliti. Pitanja i njihovi odgovori mogu biti raznoliko koncipirani. Na raspolaganju je osam različitih stilova; od pitanja s mogućnošću odabira jednog ili više odgovora, pitanja s mogućnošću uparivanja odgovora, pitanja s mogućnošću jednostrukih i višestrukih točno - netočno odgovora do postavljenih pitanja s mogućnošću davanja odgovora u pravilnom redoslijedu.

Nastavnik prilikom kreiranja sam određuje granice za određene ocjene (npr. preko 75% = odličan). Sam je napredak i uspješnost učenika vidljiv u realnom vremenu na nastavničkom računalu odnosno pametnoj ploči. Testovi znanja kreću i završavaju u isto vrijeme tako da svi učenici imaju isto vrijeme za rješavanje.

Na kraju testa nastavnik može prikazati pojedinačni rezultat testa svakom učeniku i po potrebi prikazati točne odgovore kako bi učenici shvatili gdje su pogriješili te ujedno ponovo utvrdili taj dio gradiva.

Za početak izrade ispita potrebno je pokrenuti *Test Designer* (Slika 2 Sučelje za odabir tipa pitanja), alat koji omogućuje izradu ispita s minimalnim naporom. Ispiti uključuju proizvoljan broj pitanja (Slika 3 Sučelje za kreiranje pitanja), koja mogu uključivati tekst, slike, audio i video sadržaj.

Nakon dizajniranja i objavljivanja ispita kao cjeline, koristi se konzola za testiranje. Pomoću nje se provodi sam ispit sa mogućnošću podešavanja svih potrebnih parametara kao što su dužina trajanja ispita, vrijeme za razmišljanje, način generiranja redoslijeda ponuđenih odgovora i sl.

Odgovori učenika se prate u realnom vremenu i odmah je moguće dijeliti povratne informacije.

Test dizajner može biti i samostalno instaliran na operacijski sustav.

Nastavnik može stvoriti bazu koja sadrži pitanja, audio, video i slikovne datoteke. Pomoću nje je lako i jednostavno kreirati i dizajnirati nova pitanja i ispite. Navedeni sadržaj jednostavno se može dijeliti s ostalim nastavnicima.

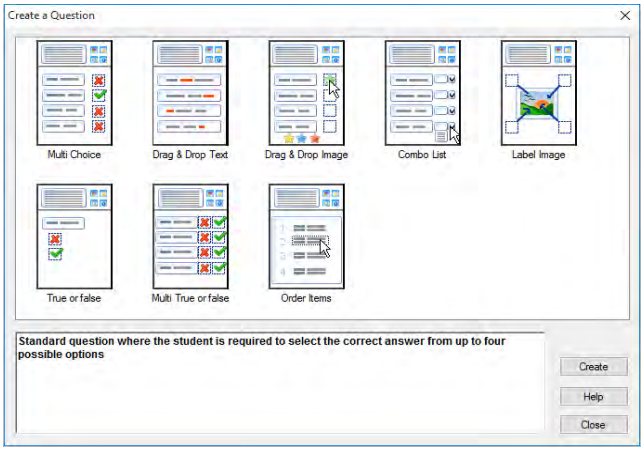
Kreiranjem bodovnog praga mogu se pratiti rezultati i napredak učenika u realnom vremenu, čak i pitanje po pitanje.

Testovi se automatski ocjenjuju te su rezultati dostupni odmah u obliku preglednog izvješća koje sadrži uspjeh kao i mogućnost uvida u točne i netočne odgovore.

Tipovi pitanja:

- *Multi Choice (Višestruki odabir)*

- *Drag & Drop Text (Povuci i ispusti tekst)*
- *Drag & Drop Image (Povuci i ispusti sliku)*
- *Combo list (Kombo lista)*
- *Label Image (Označi sliku)*
- *True or false (Točno ili netočno)*
- *Multi True or false (Višestruko točno ili netočno)*
- *Order items (Poredaj predmete)*



Slika 2 Sučelje za odabir tipa pitanja



Slika 3 Sučelje za kreiranje pitanja

Tipovi pitanja su po obliku i načinu rješavanja identični onima koje sadrže druge aplikacije za provjeru znanja pomoću računala, kao npr. *QuizMaker*.

Sučelje za izradu pitanja (nastavnikovo sučelje) se razlikuje od sučelja koje vidi učenik tijekom provjere znanja.

Primjer na tipu pitanja *Drag & Drop Image* (Slika 4 Sučelje koje vidi učenik tijekom ispita).



Slika 4 Sučelje koje vidi učenik tijekom ispita

Provjera znanja pomoću *HP Classroom Managera* je izuzetno praktična nastavnicima sa više aspekata. Taj oblik provjere primjenjiv je za: vrednovanje znanja učenika, pomoć učenicima pri učenju, provjeru predznanja učenika.

Nastavnici i učenici u realnom vremenu stječu uvid u razinu znanja učenika.

Jednom kreirana baza pitanja i ispita je višekratno upotrebljiva. Pitanja i odgovori se mogu generirati slučajnim postupkom. Nažalost, primjenjivost nije jednaka kod svih predmeta. Mogućnost odabira između čak 8 različitih tipova pitanja zahtjeva kreativnost nastavnika u kreiranju pitanja.

V. OSTALI OBLICI PROVJERE ZNANJA POMOĆU CMS-A

HP Classroom Manager (CMS) pruža vrlo široku lepezu mogućnosti komunikacije nastavnika i učenika pomoću računala. Dio tih mogućnosti komunikacije nastavnika i učenika pomoću računala može se iskoristiti i za provjeru znanja.

Jedna od tih mogućnosti je Slanje i sakupljanje datoteka *(Send/Colect)*.

Datoteke se mogu pripremiti te slati učenicima, isto tako datoteke se mogu skupiti s računala učenika. Datoteke i mape mogu se poslati na više računala učenika odjednom jednim jednostavnim postupkom. Na dokumentima koji se pošalju učenicima mogu biti ispisana pitanja, zadani zadaci, zatraženo objašnjenje. Nakon što učenici odgovore ili napišu zatraženo, pospreme svoj uradak pod istim imenom i na isto mjesto. Datoteke svih učenika se automatski prikupljaju na nastavnikovo računalo sa traženim odgovorima. Ako učenik ne pospremi napisane odgovore, datoteka neće biti pokupljena.

Mogućnost koja se poziva alatom *Communicate* prevodi se često kao *Čavrljanje i slanje poruka*. Odabiru se učenici koje nastavnik želi uključiti u diskusiju i odabirom spomenutog alata na zaslonu nastavničkog računala i učeničkih tableta prikaže se okvir za diskusiju.

Korištenje virtualne ploče (veliki interaktivni zaslon) također može poslužiti u postupku provjere. Virtualna ploča je sastavni dio pametne učionice i omogućava crtanje dijagrama, dodavanje bilješki na grafove i slike pa čak i davanje dopuštenja učeniku sudjelovanja u navedenim aktivnostima na ploči. Sve je vidljivo na pametnoj ploči koja je ispred svih učenika (obično na mjestu klasične ploče).

Odabirom alata *File Transfer* omogućava se slanje gotovog kviza učenicima ili drugih materijala za rješavanje odnosno provjeru znanja. Prije samog slanja u gornjoj traci određujemo destinaciju učenika gdje će se poslati materijal. Nastavnik ima mogućnost odabira kome će poslati materijal (Svim učenicima, pojedinačnom učeniku ili skupini učenika). Slanje poruke učeniku također može biti oblik provjere znanja. Učeniku se prikaže poruka koja može biti pitanje ili upute ili upozorenje (ovisno o sadržaju). Učenik poruku vidi ali na nju ne može odgovoriti. Odgovara usmeno ako je postavljeno pitanje.

Jedna od mogućnosti je i snimanje zaslona učenikovog tableta. Moguće je snimiti sliku zaslona učenikovog tableta ili računala i spremiti ju. Uz sliku zaslona spremiti će se ime učenika, *Login ID* učenika, ime računala, datum i vrijeme. Ova mogućnost je praktična kod provjere određenih postupaka unutar neke aplikacije.

VI. ISKUSTVA U PRIMJENI PROVJERE ZNANJA POMOĆU CMS-A

U Tehničkoj školi Ruđera Boškovića instalirana je pametna učionica. Učionica je opremljena s 30 tablet računala i pametnom pločom. Opremljena je u partnerstvu tvrtki Microsoft, Hewlett-Packard i STORM Computers, a na a zajedničku inicijativu CARNeta, Ministarstva znanosti i obrazovanja te Tehničke škole Ruđera Boškovića. Učionica se postupno uvodi u primjenu zbog nedovoljnog poznavanja CMS sustava od strane nastavnika. Od svih pobrojanih mogućnosti *HP Classroom Managera*, nastavnici su počeli sa korištenjem pametne interaktivne ploče na edukacijskim simulatorima (*EdSim, Logisim*) kao i *Office* aplikacijama. Nakon toga se krenulo sa pripremom i provedbom ispita.

Korištenjem interaktivne ploče i tableta zamijećeno je da je klasična organizacija školskog sata nije u potpunosti primjenjiva u takvom obliku podučavanja. Potrebno je nešto mijenjati u tom smislu.

Kako se u Školi već godinama koriste računala za provjeru znanja pomoću aplikacije *Free Quiz Maker*, korištenje *CMS-a* za provjeru znanja je odmah privuklo pozornost nastavnika zbog sličnosti tipa pitanja, organizacije školskog sata i provedbe ispita. Taj prijelaz i za nastavnike i za učenike koji su koristili *Quiz Maker* nije bio težak.

Učenicima, koji gotovo redovito koriste tablete i pametne telefone sam rad i postupci tijekom ispita uopće ne predstavljaju problem. Također, sučelje koje je namijenjeno učenicima tijekom samog ispita je vrlo intuitivno i tu također nije bilo problema. Određenih manjih poteškoća je bilo zbog rezolucije tableta i nedovoljne preciznosti dodira prstom na određene sitnije objekte na zaslonu tableta.

Kod pitanja koja zahtijevaju složeniji matematički izračun, a u ispitu se odabire točan rezultat među više ponuđenih, učenicima je omogućeno da na dodatnom papiru izračunaju traženo i na osnovu tog izračuna odaberu rezultat.

Određeno gradivo, karakteristično za tehničke strukovne škole, teško je provjerivo pomoću dostupnih pitanja u *Test Designeru*. Na primjer,

- pisanje programa ili programskih sekvenci
- crtanje vremenskih dijagrama
- crtanja struktura tehničkih sustava

Učenici sa zadovoljstvom prihvaćaju takav oblik provjere znanja.

U Tehničkoj školi Ruđera Boškovića provjera znanja pomoću računala se primjenjuje za provjeru predznanja učenika prije laboratorijskih vježbi na strukovnim računalnim predmetima, za vrednovanje znanja učenika na strukovnim predmetima.

VII. ZAKLJUČAK

HP Classroom Manager pruža vrlo veliki broj mogućnosti planiranja, organiziranja i izvođenja nastavnog procesa u razredu. Jedna od mogućnosti je i provjera znanja učenika u svim oblicima uz pomoć računala.

Oblici provjere znanja pomoću *CMS-a* mogu se uvjetno podijeliti na formalne i neformalne oblike.

U formalne oblike spadaju alati za izradu i provedbu testova sa velikim brojem tipova pitanja (*Test Designer*). Pri tome obliku provjere učenici odgovaraju isključivo pomoću računala odnosno tableta. U te oblike svrstavamo i alate Pitanja i odgovori koji omogućuju kombinaciju primjene računala, usmenog pitanja i usmenog odgovora ili odgovora pomoću računala.

Kako bi se odradila kvalitetna priprema i provedba ispita, potrebno je proučiti sve mogućnosti *CMS*-a vezane za organizaciju i planiranje nastavnog sata u pametnim učionicama, kao i sve mogućnosti provjere znanja pomoću *CMS-a*.

Konačan rezultat su baze pitanja i ispita koje bi bile primjenjive u svim školama koje imaju iste planove i programe, odnosno predmete.

Za što jednostavnije i bolje postizanje konačnih rezultata potrebno je uključiti stručno razvojne i pedagoške službe u školama i na razini Agencija i MZOS-a za sustavno organiziranje nastavnog sata u pametnim učionicama.

Svi oblici komunikacije nastavnika i učenika pomoću *CMS-a* mogu se iskoristiti i prilagoditi provjeri znanja. Iako ti alati za komunikaciju nisu primarno namijenjeni provjeri znanja, kreativnošću nastavnika mogu se vrlo učinkovito iskoristiti za taj oblik nastavnog procesa.

LITERATURA

[1] http://www8.hp.com/us/en/classmanager/overview.html
[2] http://h20331.www2.hp.com/hpsub/cache/hp_classroom_manager_quick_start_guide.pdf
[3] http://h20195.www2.hp.com/V2/GetPDF.aspx/4AA4-5622ENA
[4] https://ssl.www8.hp.com/h41268/live/index.aspx?qid=19533
[5] http://h10032.www1.hp.com/ctg/Manual/c02779646
[6] https://www.g2crowd.com/products/hp-classroom-manager/competitors/alternatives
[7] http://www.pepperglobal.com/money-on-the-table/Pepper_HP_Case_Study.pdf
[8] http://www.manualguru.com/hp/classroom-manager/software-guide

IZVOR SLIKA

[1] Aplikacija HP Classroom Manager

Primjena e-učenja u hrvatskom vojnom obrazovanju

Darko Možnik
Ministarstvo obrane Republike Hrvatske
Hrvatsko vojno učilište „Dr. Franjo Tuđman“
Ilica 256B, Zagreb, Republika Hrvatska
Telefon: 01-3784 613 Fax: 01-3784 210 E-mail: dmoznik@xnet.hr

Sažetak - Rad opisuje i objašnjava teorijski koncept i praktične primjene odnosno primjere e-učenja u obrambenim i vojnim sustavima u području vojnog obrazovanja koji se u RH organizira, priprema i provodi na Hrvatskom vojnom učilištu „Dr. Franjo Tuđman“.

Navode se motivi, prednosti i nedostaci, razlozi i mogućnosti tj. sposobnosti e-učenja od teorijskog koncepta (kroz planiranje, organiziranje i nadzor procesa e-učenja) do konkretnih primjera primjene uspješnih procesa e-učenja u području vojnog obrazovanja.

Navedene su definicije, oblici, elementi i tehnologije e-učenja jer je to jedan od najbržih i najjednostavnijih načina stjecanja novih znanja, vještina i sposobnosti odnosno učenja bilo kada i bilo gdje.

Za e-učenje se u NATO savezu koriste pojmovi ADL (engl. *Advanced Distributed Learning*) i JADL (engl. *Joint Advanced Distributed Learning*). Navedeni su primjeri nekih od razvijenih ADL tj. online tečajeva, modula i kolegija u području vojnog obrazovanja za potrebe MORH-a i OS RH.

Navedeni su LMS sustavi za upravljanje učenjem: ILIAS koji se koristi u zemljama članicama NATO saveza, te MOODLE i MERLIN koji se koriste u akademskoj zajednici RH.

Navedeni su programi ERASMUS (engl. *European Community Action Scheme for the Mobility of University Students*) , kao i inicijativa i program ERASMUS *militaire* poznat po kratici EMILYO (engl. *Exchange of Military Young Officers*) - razmjena kadeta i mladih časnika među vojnim akademijama zemalja članica EU.

U tijeku je proces transformacije HVU „Dr. Franjo Tuđman“ u visoko obrazovnu i znanstveno-istraživačku ustanovu MORH-a i OS RH, kao i sastavnicu akademske zajednice RH i punopravnu aktivnu članicu europskog vojnog obrazovanja.

Tijekom navedenog procesa implementiraju se sadašnje i buduće mogućnosti i sposobnosti primjene e-učenja u području vojnog obrazovanja kroz projekt *„Uvođenje novih načina usvajanja i održavanja znanja“*.

Na kraju su navedena zaključna razmatranja.

I. UVOD

U RH se za potrebe obrambenih i vojnih sustava u području vojnog obrazovanja na Hrvatskom vojnom učilištu „Dr. Franjo Tuđman“ organizira, priprema i provodi i e-učenje.

E-učenje i *učenje na daljinu* primjenjuju se u obrambenim i vojnim sustavima zemalja članica NATO saveza i EU. U obrambenim i vojnim sustavima zemalja članica NATO saveza *e-učenje* se koristi i primjenjuje pod nazivima: JADL (engl. *Joint Advanced Distributed Learning*) i ADL (engl. *Advanced Distributed Learning*).

O*pen source* sustavi za upravljanje učenjem tj. LMS-ovi (engl. *Learning Management Systems*) **MOODLE** (engl. *Modular Object-Oriented Dynamic Learning Environment*) i **MERLIN** najviše koriste akademske zajednice, a u obrambenim i vojnim sustavima zemalja članica NATO saveza i EU koristi se, uz navedene sustave, i **ILIAS** (njem. *Integriertes Lern Informations und Arbeitsko-operations System*). U MORH-u i OS RH u sustavu vojnog obrazovanja za *e-učenje* se koriste LMS sustavi MOODLE, MERLIN i ILIAS kao platforme i skupovi alata.

Također, na HVU „Dr. Franjo Tuđman“ provodi se projekt “*Uvođenje novih načina usvajanja i održavanja znanja*” odnosno *e-obrazovanje u OS RH i MORH-u.*

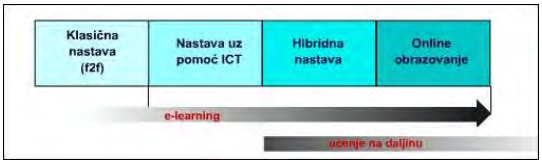
Određeni pripadnici OS RH i djelatnici MORH-a, a poglavito nastavnici Hrvatskog vojnog učilišta “Dr. Franjo Tuđman” osposobljeni su vlastitim resursima, učinkovitim korištenjem LMS-ova MOODLE, MERLIN i ILIAS, razvijati i primjenjivati *online* ADL tečajeve i module za obrambene i vojne potrebe.

II. E-UČENJE I UČENJE NA DALJINU

A. E-učenje

E-učenje je *vrsta komunikacijskog kanala*, kanala kroz koji se odvija učenje.

Postoje dvije skupine definicija *e-učenja* i to: 1. skupina "tehničkih" definicija stavlja naglasak na tehnologiju (na "*e*" dio u imenu). Primjer jedne od "tehničkih" definicija glasi: *"E-obrazovanje* je bilo koji oblik učenja, poučavanja ili obrazovanja koji je potpomognut uporabom računalnih tehnologija, a posebno računalnih mreža temeljenih na Internet tehnologijama.*"* i 2. skupina "pedagoških" definicija ima naglasak na obrazovanju, tj. na učenju i poučavanju ("*learning*" dio u imenu). Jedna od takvih definicija je: *"E-learning* je interaktivan ili dvosmjeran proces između nastavnika i učenika uz pomoć elektroničkih medija, pri čemu je naglasak na procesu učenja, dok su mediji samo pomoćno sredstvo koje upotpunjuje taj proces [1].“ *E-učenje* kao oblik obrazovanja postoji na više razina: kao potpuno samostalan oblik, ali i kao sastavni dio ili nadopuna klasičnog obrazovanja. Proširena klasifikacija ili "vremenska crta" *e-obrazovanja* prikazuje obrazovanje kao kontinuum e-učenja.



Slika 1. Prikaz “kontinuum”a *e-učenja*

Na slici 2. prikazan je odnos *e-učenja*, *učenja na daljinu i online obrazovanja*. Iako se *e-learning* i *učenje na daljinu*

često izjednačuju, nije riječ o istim oblicima obrazovanja: postoje vrste *e-obrazovanja* koje se ne odvijaju online, a isto tako postoje i oblici *učenja na daljinu* koji ne koriste ICT. Među glavnim prednostima *e-učenja* su: a) vremenska i prostorna fleksibilnost; b) interakcija (komunikacija) između polaznika i nastavnika, koja se odvija putem računala (na primjer: e-mail, forumi) te je često neposrednija i intenzivnija nego komunikacija u razredu; c) komunika-cija i grupni rad na zajedničkim projektima između polaz-nika međusobno; d) programi *e-obrazovanja* su dostupni putem Interneta nevezano za fizičku lokaciju na kojoj se polaznici nalaze; e) programe *e-obrazovanja* mogu putem Interneta pohađati polaznici koji zbog svog posla ili drugih obveza ne mogu nazočiti tradicionalnoj nastavi u vrijeme u koje se ona odvija; f) materijali pripremljeni za *e-obrazovanje* relativno se lako obnavljaju, prilagođavaju i ponovo koriste za različite grupe polaznika; g) korištenje *interaktivnih sadržaja za učenje* i različitih *medija* za prezentiranje sadržaja te *dostupnost* sadržaja 24 sata *online*.



Slika 2. Odnos *e-učenja* i *učenja na daljinu*

B. Učenje na daljinu

Učenje na daljinu (engl. *distance learning, distance education*) je obrazovanje ili obuka koja se nudi učenicima na različitim mjestima odnosno fizički udaljenima od učitelja ili izvora informacija. U praksi je učenje na daljinu puno složenije od ove definicije jer uključuje koriš-tenje novih tehnologija i novih interaktivnih nastavnih metoda. Termin učenje na daljinu u novije vrijeme se zamjenjuje terminom online učenje, čime se naglašava da je riječ o posebnom obliku e-learninga ili e-obrazovanja.

Department of Defense (DoD) je uvidio da sa konstantnim brzim i neplaniranim promjenama lokacija vojnih snaga ne može omogućiti konstantnu potrebu za edukacijom u novim tehnologijama. Pošto je vremenski interval razmještaja postrojbi kao i vrijeme implementiranja novih tehnologija kraće sve više pojavljuje se potreba za edukacijom *"right time, right place"*, gdje je vrijeme često ***sada***, a mjesto ***bilo gdje*** na planeti. Dolazi se do zaključka da u informacijskom dobu vojska mora biti spremna odgovoriti na nove prijetnje ***odmah i sada*** sa najnovijim tehnologijama. Kao odgovor na takve izazove razvijalo se i dalje se razvija *e-učenje* odnosno JADL i ADL čija je svrha postavljanje okvira koji će omogućiti *DoD* službenicima (od ratnika - do tajnika) pristup: a) kvalitetnoj edukaciji i treningu prema individualnoj potrebi, b) relativno mali troškovi i c) kada god i gdje god to bude potrebno. Kao jedna od najvećih prednosti *učenja na daljinu* smatra se to što omogućuje **stalno učenje** (engl. *lifelong learning*) i **profesionalno usavršavanje**.

III. KONCEPT, UPRAVLJANJE PROCESIMA I TEHNOLOGIJE E-UČENJA

a) KONCEPT E-UČENJA

Danas je sveprisutna potreba ali i želja da se unaprijedi kvaliteta i dostupnost različitih obrazovnih programa. Pri tome je obrazovanje postalo *cjeloživotnim procesom* i potrebom, pa se briga o dostupnosti i kvaliteti odnosi na sveukupni ciklus *cjeloživotnog obrazovanja*. E-učenje je proces obrazovanja (proces učenja i poučavanja) uz uporabu informacijske i komunikacijske tehnologije, koja pri tome doprinosi unaprjeđenju kvalitete toga procesa i kvalitete ishoda obrazovanja.

E-učenje omogućava: učenje prilagođeno individualnim potrebama studenta (bilo kada, bilo gdje), aktivno sudjelovanje i aktivnu ulogu studenata, aktualnije i ažurnije obrazovne materijale, globalni pristup različitim obrazovnim sadržajima, smanjenje „pritiska“ na fizičke resurse (omogućavajući virtualne učionice, virtualne laboratorije), više komunikacije nastavnik-student i student-student, timski rad, projektni rad nastavnika i studenta, bolju iskorištenost vremena (za kvalitetnije sadržaje) pri kontaktima nastavnik-student, samostalnu provjeru znanja i kontinuirano mjerenje i praćenje napretka studenta [5].

To je učenje i poučavanje koje je usmjereno na studente.



Slika 3. Model učenja usmjeren na studenta

b) UPRAVLJANJE PROCESIMA E-UČENJA

Model upravljanja procesom e-učenja se sastoji od 3 faze: *planiranje, organiziranje i kontroliranje* procesa e-učenja. Proces upravljanja e-učenjem napreduje kroz slijedeće faze: *planiranje* procesa e-učenja (izrada operativnog plana odnosno scenarija e-učenja), *organiziranje* odnosno implementacija procesa e-učenja (uz provedbu scenarija e-učenja u praksi), *kontroliranje* procesa e-učenja (tj. vrednovanje različitih aspekata procesa) i *poboljšanje procesa i platforme*. Osnovni princip ovog procesa je *ponavljanje* - jednom kada se potvrde ili odbace početne pretpostavke kroz fazu vrednovanja, ciklus se ponavlja uz primjenu poboljšanja [6].

PLANIRANJE procesa e-učenja je iznimno bitan za cjelokupni uspjeh e-učenja, a najčešće obuhvaća aktivnosti od analize suvremenih trendova u obrazovanju i identificiranje karakteristika polaznika tečajeva za e-učenje; zatim osmišljavanje okoline za učenje koja odgovara sadržaju, ciljanoj publici i tehničkoj platformi te odabira odgovarajuće platforme za e-učenje; pa sve do formuliranja očekivanih ishoda e-učenja i izrade operativnog plana odnosno scenarija e-učenja. Cilj faze planiranja procesa e-učenja je osmišljavanje ishoda i scenarija e-učenja prema određenom kontekstu. Ishodi učenja predstavljaju instru-

ment razvoja plana i programa, te provjere njegove konzistentnosti. Na razini programa preporučeno je definirati 15 do 30 ishoda učenja, na razini predmeta 5 do 10, a na razini nastavne cjeline 5 do 15 ishoda učenja [7].

ORGANIZIRANJE osnovnog sadržaja e-učenja

LMS sustavi omogućuju polaznicima da kolegije istražuju nelinearno, a isto tako i nastavnicima omogućuju izradu tečajeva na fleksibilan način, nelinearno. Redoslijed dodavanja materijala i funkcionalnosti u sam kolegij ovisi isključivo o nastavniku. Ako nastavnik i ne zna gdje započe-ti, npr. nije siguran treba li organizirati kolegij oko tjed-nog ili tematskog oblika, treba ipak odabrati jedan oblik i početi s dodavanjem materijala. U obrazovnom okruženju gdje se koriste LMS-ovi nastava se može odvijati i u potpunosti *online*.

KONTROLIRANJE napretka i ponašanja polaznika

Pri kontroliranju je nužno postaviti standarde mjerenja, izmjeriti performanse, usporediti performanse sa standardima te poduzeti potrebne korektivne akcije u svrhu poboljšanja cjelokupnog procesa e-učenja. Svrha faze kontroliranja e-učenja može se najbolje izraziti kao promatranje i procjena cjelokupnog procesa e-učenja i očekivanih rezultata u odnosu na postavljene ciljeve.

c) TEHNOLOGIJE E-UČENJA

Postoji više tehnologija obrazovanja i obuke primjenom e-učenja, a prema NATO-voj podjeli, to su:

1. združeno napredno distribuirano učenje - JADL (engl. *Joint Advanced Distributed Learning*) i napredno distribuirano učenje - ADL (engl. *Advanced Distributed Learning*),
2. učenje pomoću ICT-a odnosno CBT (engl. *Computer-Based Training*),
3. „uronjeno“ učenje (engl. *Immersive learning*),
4. mobilno učenje (engl. *M-learning*),
5. suradničko ili kolaborativno učenje (engl. *Collaborative learning*) i
6. kombinirano ili hibridno učenje (engl. *Blended learning*) [5].



Slika 4. Tehnologije e-učenja prema NATO-u

1. JADL i ADL

a) JADL



Slika 5. Oznaka za JADL

b) ADL



Slika 6. Oznaka za ADL

ADL (engl. *Advanced Distributed Learning*) proizvod je *ADL inicijative* uspostavljene 1997. godine u Ministarstvu obrane SAD-a za standardizaciju i modernizaciju upravljanja „treninzima“ i „edukacijama“ te načinima poučavanja. ADL koristi strukturne i suradničke (engl. *Collaboration*) metode iz industrije, akademskih zajednica i vlada koji razvijaju standarde, alate i sadržaje učenja.

Zemlje članice NATO saveza i EU mogu koristiti izrađene obrazovne on-line tečajeve za različita područja u obrambenim i vojnim sustavima.

Više o JADL i ADL nalazi se na: https://jadl.act.nato.int

2. CBT (engl. *Computer Based Training*)

CBT je obrazovni ili obučni tečaj dostavljen direktno na studentsko računalo putem CD-ROM-a ili nekog drugog uređaja za pohranu. On se razlikuje od ADL-a po tome što se ne isporučuje preko weba ili putem sustava za upravljanje učenjem.



Slika 7. Oznaka ili logo za CBT

3. „URONJENO“ UČENJE (engl. *Immersive learning*)

„Uronjeno“ (engl. *Immersive*) učenje koristi računalno simulirano okruženje koje se događa u stvarnom životu ili hipotetsku situaciju u grafički bogatom i dinamičnom okruženju. Studenti su stavljeni u situaciju i uključeni su u proces učenja kroz interaktivne simulacije i igre temeljene na *game-based* aplikacijama. *Immersive* učenje podupire jednog učenika ili više malih timova da rade zajedno kako bi riješili problem, uvježbali tehnike ili poboljšali svoje vještine.

4. MOBILNO UČENJE (engl. *M-learning*)

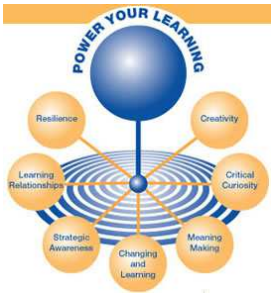


Slika 8. Mobilno učenje (engl. *M-learning*)

Mobilno učenje (engl. *M-learning*) je uporaba prijenosnih računalnih uređaja, kao što su različiti iPad-i i drugi tableti, prijenosna računala, osobni digitalni pomoć-nici (PDA) i pametni telefoni povezani s bežičnim mreža-ma. To omogućuje da se podučavanje i učenje proširi i na prostore izvan tradicionalne učionice te daje nastavnicima i polaznicima povećanu fleksibilnost i nove mogućnosti za

interakcije koje su prikladne i dostupne virtualno s bilo kojeg mjesta.

5. SURADNIČKO ILI KOLABORATIVNO UČENJE



Slika 9. Suradničko ili kolaborativno učenje

Suradničko ili kolaborativno učenje (engl. *Collaborative learning*) podupire nastavu kroz interaktivne metode ili online komunikacije koristeći društvene medijske alate kao što su chat sobe, tribine i video alate kao što su webi-nari i web-emitiranja. Ovi alati se mogu implementirani u učionici i online okruženju kako bi podržali tijek nastave. Oni daju mogućnost velikom broju polaznika da rade zajedno s nastavnikom-instruktorom.

6. KOMBINIRANO ILI HIBRIDNO UČENJE (engl. *Blended learning*)



Slika 10. Kombinirano ili hibridno učenje

U zemljama članicama NATO saveza i EU koriste se LMS-ovi: ILIAS, MOODLE i MERLIN.





Slika 11. LMS sustav *ILIAS*



Slika 12. LMS sustav *MOODLE*

IV. EUROPSKO VOJNO OBRAZOVANJE EMILYO - RAZMJENA KADETA I MLADIH ČASNIKA MEĐU VOJNIM AKADEMIJAMA EU

10. studenog 2008. godine, tijekom francuskog predsjedanja EU-om, 27 ministara obrane EU jednoglasno su podržali osnivanje „Europske inicijative za razmjenu mladih časnika“ inspirirana programom ERASMUS (engl. *European Community Action Scheme for the Mobility of University Students*). *ERASMUS militaire* je ubrzo postao poznat po kratici EMILYO (engl. *Exchange of Military Young Officers*). Od rujna 2009. do rujna 2013. godine EMILYO je omogućio razmjenu 630 kadeta i mladih časnika te znatno ojačao europeizaciju visokog vojnog obrazovanja.

U suvremenom sigurnosnom okružju biti časnik podrazumijeva iznimnu međunarodnu aktivnost, gotovo od prvog dana aktivne službe. Ovaj poziv zahtijeva ne samo razumijevanje kompleksnosti područja operacija i razumijevanje opcija međunarodnog odgovora, nego također poznavanje rada, običaja i standarda drugih nacija i partnera u misijama i operacijama. Dvoje renomiranih europskih vojnih stručnjaka, Hans Wiedner i Herald Gell, navode kako “časnici trebaju kreativnost, prilagodljivost i ostale instrumente vođenja kako bi mogli na pravi način odraditi zadaću… Ako ih moramo učiti vođenju u ekstremnim uvjetima, koja je razlika da li ih to učimo u pustinji, džungli ili na Arktiku? Ako ih moramo učiti radu u međunarodnom okruženju, moramo ih tijekom obuke poslati u inozemstvo. Budući časnici moraju gledati preko horizonta, a profesori i obučavatelji na prvoj razini vojnog obrazovanja moraju ih ohrabriti i učiti tome. Samo tada možemo opravdano reći da smo pružili najbolje moguće obrazovanje našim budući časnicima.“

Ovakve opće težnje razvoja vojne obuke, dodatno su usložene intenziviranjem zahtjeva za stvaranjem prepoznatljivog europskog vojnog identiteta i veće vidljivosti europskog odgovora na međunarodne krize. EMILYO je stoga osnovan u skladu s Europskom sigurnosnom strategijom (*ESS)* i djeluje kao jedan od modela implementacije europske Zajedničke sigurnosne i obrambene politike (*CSDP*) s primarnim ciljem promicanja europske kulture sigurnosti i obrane tijekom inicijalnog vojnog obrazovanja i obuke budućih nacionalnih vojnih elita.

V. EMILYO PROGRAMI

EMILYO je načelno podređen Europskom sigurnosnom i obrambenom fakultetu (*ESDC*) te Implementacijskoj grupi (engl. *Implementation Group - IG*) za EMILYO. Ove institucije razvile su način europeizacije visokog vojnog obrazovanja s posebnim naglaskom na poticanje zajedničke kulture sigurnosti i obrane, kroz dvije glavne smjernice razvoja:

1. Obveza obrazovanja i obuke mladih časnika o svim aspektima CSDP-a i
2. Nužnost stavljanja mladih časnika u europsko okruže-nje tijekom raznih faza inicijalne časničke obuke.

Naime, cijela inicijativa je izgrađena na dvije glavne pretpostavke: prva je da mladi časnici sve češće rade u

međunarodnom okruženju već u inicijalnim fazama svoje karijere. Ovo se vidi na terenu u mirovnim i humanitar-nim misijama i operacijama, posebno u vojskama zapad-nih zemalja. Druga pretpostavka je da će mladi časnici pu-no bolje raditi na terenu u međunarodnom okružju, ako su već tijekom školovanja bili izloženi radu, obuci i koopera-ciji s pripadnicima drugih nacija. Na taj način se već u najranijoj fazi kulturološki senzibiliziraju, prilagođavaju svoja znanja i vještine primjeni na stranom jeziku, te ubr-zano uče kako ono što je „normalno" u jednoj državi čla-nici, nije neophodno „normalno" i u drugoj.

Temeljem ovih smjernica u proteklih pet godina razvijena su četiri oblika implementacije EMILYO inicijative:

1. Zajednički i međunarodno akreditirani akademski moduli u trajanju od tjedan dana. Moduli su razvijeni kao krucijalno važni za sve mlade europske časnike, a neki od njih su:
a) „Common module on CSDP" - jednotjedni ključni zajednički modul koji upoznaje polaznike s konceptom, mehanizmima i izazovima CSDP-a i služi europeizaciji inicijalne časničke obuke. Prvi put je proveden 2009. godine u Vojnoj akademiji Portugala, a danas se provodi u 8 zemalja članica EU za sve ostale države Unije. Iz njega je proizašla i CSDP Olimpijada na kojoj se kadeti i mladi časnici natječu u poznavanju tematike Europske sigurnosne i obrambene politike. Održava se jednom godišnje na Cipru.
b) „Common module on media training" - jednotjedni, akreditirani, zajednički modul koji upoznaje polaznike kako raditi i ponašati se s medijima u području operacija. Modul je razvila Vojna akademija Belgije.

2. Akademski moduli u trajanju 3-6 mjeseci

Svaka od akademija razvija određene semestralne programe koji su otvoreni za međunarodne studente. Npr. Austrija ima razvijen cijeli zimski semestar iz vojnog vođenja *„Bachelor Program Military Leadership"*, Španjolska nudi ukrcaj stranih kadeta na svoje školske jedrilice i brodove u trajanju do 6 mjeseci s ciljem uvježbavanja procedura mornaričkog vođenja i zapovijedanja, Kopnena vojna akademija Poljske također nudi semestar za strane studente iz područja sigurnosnih strategija i međunaro-dnih odnosa, a Francuska vojna akademija u Rennesu za EU vojne kadete nudi *„International political science semester"* u trajanju od 3 mjeseca i nosi 30 ECTS bodova.
Dio zemalja je razvio i vlastite nacionalne programe razmjene, poput austrijskog *„International Training for Job ITJ"*. Ovaj program obvezuje SVE austrijske kadete da moraju 6 tjedana stručne prakse provesti u stranoj zemlji kao zapovjednici vodova.

3. Završetak cjelokupnog vojnog obrazovanja u jednoj državi je moguć u Austriji, Poljskoj, Češkoj, Slovačkoj i Francuskoj, ali se rjeđe događa zbog usklađivanja sa standardima civilnog obrazovanja i dijela vojne obuke.

4. Razmjena kadeta i mladih časnika u vojnoj obuci.
Ove se razmjene najčešće događaju tijekom ljetnih i zimskih kampova, vojnih obuka ili stažiranja. Razmjene su intenzivne s preko 100 obučnih aktivnosti godišnje.

Samo u ovom trenutku ponuđeno je više od 36 oblika vojnog obrazovanja i obuke u 10 europskih vojnih akademija u trajanju od tjedan do 6 mjeseci. Zemlja primateljica najčešće snosi sve troškove školovanja, smještaja i prehrane, dok matična zemlja snosi troškove putovanja i dnevnica.

Pet godina programa EMILYO omogućio je razmjenu 630 kadeta i mladih časnika diljem europskih vojnih akademija, te se iz godinu u godinu, broj dana razmjene intenzivno povećavao.

Iako je EMILYO primarno orijentiran na kadete i mlade časnike, on je značajno utjecao i na same vojne obrazovne institucije koje su se morale prilagoditi akreditiranim programima, sadržajima i standardima obučavanja, provođenju obučnih sadržaja na stranom jeziku (najčešće engleskom ili francuskom), apsorbirati strane polaznike i vrjednovati vlastite programe i kroz njihova iskustva, viđenja i stavove.

VI. ORUŽANE SNAGE RH I INICIJATIVA I PROGRAM EMILYO

Osnivanjem programa „Kadet" 2003. godine, RH je dobila mogućnost sudjelovanja u razmjenama i obukama kadeta i mladih časnika diljem Europe. Aktivnosti su se intenzivirale 2008. godine, a Terezijanska vojna akademija iz Bečkog Novog Mjesta formalno je započela ovu suradnju pozivajući hrvatske kadete da po prvi put sudjeluju na austrijskom časničkom balu. Hrvatima u čast, dozvoljeno je da naši kadeti i kadetkinje otvore bal uz zvukove Jelačićevog marša. Bal je označio i simbolički ulazak RH i OS RH u europske sustave vojnog obrazovanja jer od tada kadeti započinju sa sudjelovanjem u vojnim obučnim modulima (C i D) austrijske vojne akademije. S druge strane, Austrija je izabrala OS RH kao mjesto za provođenje stručne prakse svojih kandidata za časnike (programa ITJ), te se svake godine 2 kadeta upućuju u postrojbe hrvatske kopnene vojske na 6 tjedna, gdje uz nadzor našeg mentora zapovijedaju i rade s vodom OS RH.
Uz aktivnosti u Austriji, hrvatski kadeti sudjeluju i u EMILYO inicijativi Njemačke časničke škole u Dresdenu, koja organizira jednotjedne module upoznavanja s njemačkom kopnenom vojskom, postrojbama, opremom i sustavom obuke.
Od 2013. godine OS RH su se i uključile po prvi put u Konferenciju mladih vojnih znanstvenika CEFME, koja se održala u Budimpešti. Na konferenciji se prezentiraju vojno-stručni društveni i tehnički radovi. Navedena konferencija održana je u svibnju 2014. u Slovačkoj i 2015. u Poljskoj. Konferencija uvezuje i rad iMAF-a (engl. *International Military Academic Forum*) i njenog znanstvenog odbora čiji članovi ocjenjuju dostavljene i prezentirane radove.
Tijekom 2015. godiine upućeni su hrvatski kadeti na jednotjedni EU CSDP modul u Vojnu akademiju Italije u

Torinu, koji im omogućio sudjelovanje i na sljedećoj CSDP olimpijadi koja će se održati 2016. godine na Cipru.



Slika 13. Hrvatski kadeti na modulu D vojne obuke u Austriji



Slika 14. Hrvatski i njemački kadeti na međunarodnom tjednu u Dresdenu

VII. POGLED U BUDUĆNOST



Slika 15. Hrvatsko vojno učilište „Dr. Franjo Tuđman“



Slika 16. Primjena elemenata e-učenja na terenu

HVU „Dr. Franjo Tuđman“ je u procesu transformacije u visoko obrazovnu i znanstveno-istraživačku vojnu ustanovu odnosno instituciju tj. sastavnicu sveučilišne zajednice. Od 2014. godine provodi dva nova preddiplomska sveučilišna studijska programa za potrebe OS RH. Vizija je da se uvođenjem e-učenja želi u procesu obrazovanja ostvariti aktivan partnerski odnos nastavnika i polaznika, prilagođen osobinama i potrebama svakog polaznika, s jasno postavljenim obrazovnim ciljevima i jasno dodijeljenim odgovornostima za postizanje tih ciljeva. Uvođenjem i aktivnom primjenom e-učenja HVU „Dr. Franjo Tuđman“ namjerava ostvariti četiri strateška cilja: unaprijediti kvalitetu obrazovanja, omogućiti nastavnicima i studentima ostvarivanje novih uloga u obrazovnom procesu, povećati konkurentnost i dostupnost učilišta i učilišnih programa i osposobiti studente za uporabu tehnologija cjeloživotnog učenja.

I u europskom, kao i u hrvatskom, vojnom obrazovanju ponuđen je određeni broj tečajeva i zajedničkih modula u području e-učenja. Svaki od njih obuhvaća IDL i f2f fazu. Sadašnje sposobnosti su na zavidnoj razini, a njihovim daljnjim razvojem, koje su precizirane Smjernicama za daljnje unapređenje sustava e-učenja, osnažit će se novim sposobnostima, a polaznici će sistematizirati postojeća i usvojiti nova znanja, vještine i sposobnosti kojima će biti osposobljeni za još učinkovitiju njihovu primjenu u svakodnevnim zadaćama.

VIII. ZAKLJUČAK

U informacijskom društvu i vremenu *e-učenje* dobiva sve više na značenju poglavito u europskom i hrvatskom vojnom obrazovanju. Može se zaključiti kako je u europskom i hrvatskom vojnom obrazovanju napravljen značajan pomak u razvoju i uporabi sustava e-učenja, ali zbog važnosti stalnog izlaganja međunarodnom okruženju tijekom obuke i obrazovanja, većina ih ipak uvezuje ADL i f2f module u vojnom obrazovanju svojih pripadnika. Razvijen je i ponuđen veliki broj „gotovih“ *online* tečajeva za potrebe obrambenog i vojnog sustava. Stvoreni su potrebni preduvjeti za implementaciju *e-učenja* i u obrambenom i vojnom sustavu RH.

„Znanjem u budućnost“ moto je sadašnjih naraštaja i ostat će moto i naraštajima koji će doći na vojno obrazovanje u škole Hrvatskog vojnog učilišta „Dr. Franjo Tuđman“ u Zagrebu.

LITERATURA

[1] A. W. (Tony) Bates, „Upravljanje tehnološkim promjenama - Strategije za voditelje visokih učilišta“, *CARNet* - Zagreb i *Benja* - Lokve, 2004. (prijevod), Izvornik: „Managing Techological Change: Strategies for College and University Leaders“, *Jossey-Bass Publishers*, San Francisco, USA, 2000.

[2] B. Shackelford, „Project Managing E-Learning“, *American Society for Training & Development Press*, Alexandria, USA, 2002.

[3] M. M. Lynch & J. Roecker, „Project Managing E-Learning“, A handbook for successful design, delivery and management, *MPG Books Ltd*, Bodmin, Great Britain, 2007.

[4] J. Greene & A. Stellman,“Head First PMP“ A Learner's Companion to Passing the Project Management Professional Exam, A Brain-Friendly Guide, *Oreilly Media, Inc*., CA, USA, 2007.

[5] Grupa autora, „E-Lerning Concept“, NATO unclassified, *NATO, Bruxelles, Belgium*, 8.2.2012.

[6] M. Ćukušić & M. Jadrić, „E-učenje: koncept i primjena“, *Školska knjiga d.d.*, Zagreb, 2012.

[7] B. Divjak, „Ishodi učenja u visokom školstvu“, TIVA, FOI, Varaždin, 2008.

[8] P. Jandrić & D. Boras, „Kritičko e-obrazovanje Borba za moć i značenje u umreženom društvu“, *Tehničko veleučilišta u Zagrebu i Filozofski fakultet Sveučilišta u Zagrebu*, Zagreb, 2012.

[9] J.D. Fletcher, „Education and Training Technology in the Military“, *Science magazine*, vol. 323, 2 January 2009., str. 72

Prezentacijski alati za prikaz matematičkih sadržaja

M. Štefan Trubić*, I. Radošević **
* Sveučilište u Rijeci/Tehnički fakultet, Rijeka, Hrvatska
** Sveučilište u Rijeci/Odjel za matematiku, Rijeka, Hrvatska
mstefan@riteh.hr, inesr@math.uniri.hr

Sažetak – Učinkovito predstavljanje rezultata svoga rada u javnosti zahtijeva pripremu odgovarajuće prezentacije prilagođene određenoj publici. Sat moderne nastave matematike se može znatno unaprijediti i obogatiti uvođenjem prezentacija za pojedine teme čime obrazovni program dobiva novu kvalitetniju dimenziju. Pored sadržaja i načina na koji se isti prikazuje, ključni elementi prezentiranja su i odabir prikladnih prezentacijskih programa i alata koji su usklađeni i prilagođeni sadržaju. Danas postoje različiti alati koji olakšavaju stvaranje prezentacija. Posebno će nas zanimati alati pogodni za predstavljanje matematičkih sadržaja koji uključuju prikaz složenih matematičkih izraza. U radu će biti opisani poznati i često korišteni prezentacijski alati: Beamer LaTeX, PowerPoint i Prezi, pri čemu će se istaknuti njihove najvažnije prednosti i nedostaci, te napraviti međusobna usporedba. Posebno će biti istaknuta uloga Iguana TeX-a kao poveznice između Beamer LaTeX i PowerPoint alata. Uvid u osnovne elemente svakog pojedinog alata dobit će se primjenom istih na prezentaciju jedne odabrane matematičke teme, konkretno numeričke integracije.

I. UVOD

Kvalitetno izrađena prezentacija temelj je gotovo svakog uspješnog predavanja, neovisno je li cilj edukacija publike unutar nastavnog procesa ili informiranje javnosti o rezultatima rada. Kvalitetu prezentacije primarno određuju sadržaj i struktura. Poznato je da sadržaj treba predstaviti jasno, precizno, dosljedno i na što jednostavniji način. S druge strane, struktura je određena svojim sastavnim dijelovima, odnosno pojedinim temama koje bi trebale biti poredane logičnim slijedom. U svakoj temi treba istaknuti ključnu poruku ili ideju koja će doprinijeti da se cjelina uspješno zaokruži, izvede željeni zaključak i postigne unaprijed zadani cilj predavanja.

Znatan doprinos kvaliteti prezentacije daje i pravilan odabir prezentacijskih programa i alata koji su usklađeni i prilagođeni odabranom sadržaju koji se prezentira. Prikladni programi i alati za izradu prezentacija mogu pojednostavniti predstavljanje određenih ideja, znanja i zaključaka namijenjenih odgovarajućoj publici. Pri odabiru prezentacijskog alata autor bi trebao voditi računa da prikaz odabranog sadržaja bude što kvalitetniji, ali i da je istovremeno izrada prezentacije u odabranom prezentacijskom programu što jednostavnija. To se prvenstveno odnosi na odabir dizajna predložaka prezentacije, unos teksta i različitih formula, dodavanje slika, animacija i zvukova, usklađivanje fontova i boja, te na kraju jednostavno generiranje odgovarajućeg prikaza koji će se prezentirati javnosti. Pri tome bi se prezentacije namijenjene nastavnom i odgojno-obrazovnom procesu trebale voditi metodičkim načelima, posebno načelima motivacije, sistematičnosti i postupnosti, zornosti, primjerenosti i znanstvenosti.

U ovom će se radu posebna pozornost usmjeriti na one prezentacijske alate koji su prije svega pogodni za kvalitetno i pravilno predstavljanje matematičkih sadržaja koji uključuju prikaz složenih matematičkih izraza.

Bit će opisan Beamer LaTeX kao prezentacijski alat koji se najčešće koristi u znanstvenim krugovima, ali i u nastavi matematike i prirodnih predmeta, posebno zbog izuzetno lijepog prikaza složenih sadržaja. Istaknut će se važnost Microsoft PowerPoint-a kao alata za izradu prezentacija koji je zbog svoje jednostavnosti najpopularniji, te ujedno i najzastupljeniji među svim prezentacijskim alatima koji se danas koriste. Način na koji se ta dva najmoćnija prezentacijska alata mogu povezati bit će predstavljen kroz primjenu Iguana TeX-a, kao posebnog programa kojim se složeni matematički izrazi mogu uvesti u PowerPoint prezentacije iz LaTeX okruženja. Također, bit će predstavljen i Prezi kao jedan relativno novi prezentacijski alat čije su posebnosti nelinearni i izuzetno dinamični prijelazi između pojedinih stranica. Kroz rad će se istaknuti najvažnije prednosti i nedostaci svakog od navedenih alata, te napraviti međusobna usporedba na odabranoj matematičkoj temi pod naslovom „Numerička integracija“.

II. PREZENTACIJSKI ALATI

A. *Beamer LaTeX*

TeX je kompjuterski alat kojim je moguće izraditi raznovrsni tiskani materijal, od jednostavnog teksta do složenih matematičkih izraza i to najviše tiskarske razine. Njegov je idejni začetnik i tvorac Donald Ervin Knut, američki informatičar i počasni profesor na Stanford University-ju, koji je još davne 1978.g. održao prvo predavanje o Tex-u (vidi Donald E. Knuth: The TeXbook, Volume A of Computers and Typesetting. Addison-Wesley, Massachusetts, 1984.).

Matematički tekstovi, u širem smislu i tehnički tekstovi, nerijetko sadrže mnoštvo formula i simbola što ih čini zahtjevnijima za izradu. Stoga ne čudi što je programski paket LaTeX, zasnovan na TeX-u, omiljen upravo matematičarima, za pisanje članaka, knjiga i nastavnog materijala u obliku prezentacija. LaTeX je izradio Leslie Lamport 1986. godine (vidi Leslie Lamport: A Document Preparation System. Addison-Wesley, Massachusetts, second edition, 1994.). Neosporna posebnost istog je upravo lijepo oblikovanje matematičkih izraza. Dodatna prednost je što je ovaj prezentacijski alat zajedno sa svim svojim dodatnim paketima besplatan.

Prema [1] slaganje LaTeX dokumenta započinje definiranjem logičke strukture dokumenta - poglavlja, odjeljaka, odlomaka, i to samo s nekoliko naredbi, čime će se postići grafički uređen dokument. Tekst procesori, kao što je Microsoft Word, imaju drugačiji pristup izradi dokumenta. Korisnik interaktivno određuje izgled dokumenta, tijekom unošenja teksta i drugih objekata, definira i radi izmjene na dokumentu, te u svakom trenutku vidi kako on izgleda. To su takozvani WYSWYG programi, čiji naziv proizlazi iz prvih slova engleskih riječi what you see is what you get, što u prijevodu znači što vidiš to i dobiješ. Uvid u izgled LaTeX dokumenta moguć je tek nakon procesiranja tekstualne datoteke formata tex u dokument formata dvi, ps ili pdf.

Rasprave u smislu prednosti i nedostataka, s jedne strane LaTeX-a i s druge tekst procesora, mogu se promatrati s više aspekata, na primjer onog estetskog, strukturnog, onog zbog specifičnih zahtjeva ili jednostavno zbog osobnog afiniteta. O ukusima se ne raspravlja, no možemo navesti samo neke činjenice, koje se nama matematičarima ipak čine kao prednosti LaTeX-a u odnosu na ostale tekst procesore. Kao što je već rečeno, LaTeX dokumenti su dobro i uniformno strukturirani. Iako je u novijim tekst procesorima moguće, pa čak i solidno, zapisati matematički izraz, ipak prednost dajemo LaTeX-u, jer matematički tekst nerijetko sadrži podosta matematičkih izraza i specijalnih simbola koje u LaTeX-u možemo zapisati i dobiti dokument profesionalnog dizajna. Prema [2] jednostavno se generiraju i složene strukture kao što su bibliografija, popis slika i tablica, indeksi i sadržaj. Također, pri uvođenju i korištenju raznih numeracija, fusnota, referiranja i povezivanja sadržaja ne treba voditi računa o njihovom namještanju, redoslijedu i izmjenama. Uz zaista velike mogućnosti i odabir simbola, ono što još možemo spomenuti je to da nam LaTeX omogućuje definiranje nama specifičnih izraza i simbola, baš onih koje ćemo koristiti u našem dokumentu, te posebno dijakritičkih slova naše abecede.

Kad govorimo o izradi matematičkog teksta u užem smislu, odnosno onoga u prezentacijama, nameće se pitanje: koji alat koristiti? Odmah ćemo spomenuti Beamer, vidi [3], LaTeX-ov paket za izradu prezentacija kojeg je razvio Till Tantau 2003.godine (vidi Till Tantau: User Guide to the Beamer Class Version 3.01, 2004.).

Izrada prezentacije u Beamer-u je drugačija od one korištenjem WYSWYG programa, na primjer korištenjem Microsoft PowerPoint-a koji će biti kasnije opisan, a gotovo jednaka kao i izrada ostalih LaTeX dokumenata.

U nastavi matematike, za izradu prezentacija, češće se koristi LaTeX Beamer. Razlog tome je što se nastava gotovo uvijek priprema u obliku skripte, naravno u LaTeX-u, pa uz definiranje strukture novog LaTeX dokumenta koristimo već postojeći kod. Beamer LaTeX najviše koriste znanstvenici za prezentaciju rezultata istraživanja, posebno onih u prirodnim znanostima. Prezentacija je dostupna za distribuciju u pdf formatu, što je memorijski optimalno i savršeno za predavanja ili seminare.

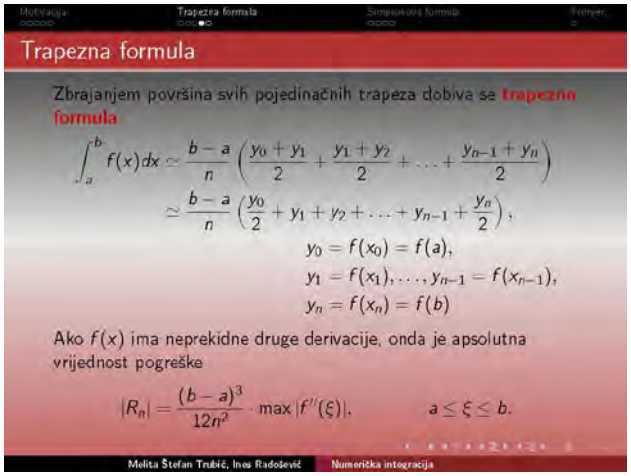
Na primjeru prezentacije iz područja matematike, konkretno jednim dijelom prezentacije s temom „Numerička integracija“, načinjen je kratak uvid u osnovne elemente prezentacije izrađene u LaTeX Beamer-u. Dvije odabrane stranice, po sadržaju, bit će načinjene i u drugim prezentacijskim alatima i time poslužiti njihovoj usporedbi. Odabrane stranice iz LaTeX Beamer prezentacije prikazane su u pdf formatu, na gornjim dijelovima slika 1 i 2, a njima pripadajući TeX kodovi, na donjim dijelovima istih slika.

Beamer LaTeX dokument iz prethodnog primjera započinje naredbom \documentclass{beamer} kojom se definira klasa dokumenta u LaTeX-u, te naredbom \usetheme{ime} kojom se određuje izgled cijele prezentacije, vidi [4]. Imena su obično vezana za veće gradove, konkretno u primjeru odabrana je tema Warsaw. Zatim se odabire boja podloge i fontova, njihova veličina i stil, te uključuju različiti paketi ovisno o sadržaju koji se




```
\documentclass[compress,red]{beamer}
\mode<presentation>

\usepackage[croatian]{babel}

\usetheme{Warsaw}
\usecolortheme{whale}
\useoutertheme[subsection=false]{smoothbars}
\setbeamertemplate{background canvas}[vertical shading]
        [bottom=palette primary.bg!80!white!,top=black!90!white!,middle=white!90!black]
\setbeamerfont{title}{size=\Large}
\setbeamerfont{date}{size=\scriptsize}


\begin{document}

\title{\bf{Numerička integracija}}
\author{\bf{Melita Štefan Trubić, Ines Radošević}}

\end{document}
```

Slika 1. Beamer Latex prezentacija “Numerička integracija” - naslovna stranica u pdf i tex formatu




```
\subsection{Trapezna formula}
\frame{
\frametitle {Trapezna formula}
\setbeamercovered{invisible}
Zbrajanjem površina svih pojedinaènih trapeza dobiva se
\textcolor[rgb]{0.804,0.00,0.00}{\textbf{trapezna formula}}
 \begin{equation*}
        \begin{split}
                \int_{a}^{b} f(x)dx &\simeq
                \frac{b-a}{n}\left( \frac{y_{0}+y_1}{2}+\frac{y_{1}+y_{2}}{2}+\ldots
    + \frac{y_{n-1}+y_n}{2}  \right)\\
                &\simeq \frac{b-a}{n}\left( \frac{y_{0}}{2}+y_1 +y_{2}+\ldots
    + y_{n-1}+\frac{y_n}{2}  \right),\\
                & \hspace{3cm} y_0=f(x_0)=f(a),\\
                 & \hspace{3cm} y_1=f(x_1),\ldots, y_{n-1}=f(x_{n-1}),\\
                & \hspace{3cm} y_n=f(x_n)=f(b)
         \end{split}
        \end{equation*}
        \pause
 Ako $f(x)$ ima neprekidne druge derivacije, onda je apsolutna vrijednost pogreške
\begin{equation*}
        |R_{n}|=\frac{(b-a)^3}{12n^2}\cdot \max|f''(\xi)|, \hspace{1.5cm} a\leq \xi \leq b.
\end{equation*}
}
```

Slika 2. Beamer Latex prezentacija "Numerička integracija" - odabrana stranica u pdf i tex formatu

želi prezentirati. Osnovna prezentacija se sastoji od okvira koji nakon procesiranja predstavljaju pojedine stranice prezentacije.

Pored unošenja teksta, moguće je kreiranje numeriranih i nenumeriranih lista, crtanje slika, umetanje animacija i slika u različitim formatima, biranje boja, te ono najvažnije - upisivanje vrlo složenih matematičkih izraza i formula. Nakon procesiranja one dobivaju matematički savršen oblik i iznimno lijepo uređen prikaz. Dinamika prezentacije se može određivati pauzama i različitim mogućnostima transparentnosti teksta, a jednostavna su i pozivanja na različite dijelove stranica i cijele stranice.

B. Microsoft PowerPoint

Za prezentacije na konferencijama gdje se ne inzistira toliko na matematičkim izrazima nego više na idejama i zaključcima ili za prezentacije popularnih predavanja češće se koriste drugi prezentacijski alati. Uz sve spomenute prednosti LaTeX dokumenta, ipak se kod prezentacija ponekad napušta LaTeX-ov Beamer i to ponajviše zbog standardiziranog, jednoobraznog izgleda prezentacija.

Poboljšanje u tom smislu daje Microsoft PowerPoint. Uz jednostavno unošenje teksta, nudi i bogat odabir predložaka, velike mogućnosti obrade raznih objekata, slika, filmova i zvukova, omogućuje animacije i razne vizualne efekte, te posebno zanimljive prijelaze između pojedinih stranica. Nedostatak toga je što će PowerPoint prezentacije zahtijevati nešto više memorijskog prostora, te što nije zanemarivo, program nije besplatan kao Beamer.

Uz neospornu važnost sadržaja prezentacije i samog izlaganja, važna je vizualna privlačnost i dinamičnost prezentacije, čime je Microsoft PowerPoint postao toliko popularan. Može se prilagoditi potrebama korisnika neovisno o području njihovog djelovanja. Čak i učenici u osnovnoj školi koji po prvi put ulaze u svijet prezentacija i prezentiranja različitih sadržaja čine to koristeći upravo Microsoft PowerPoint.

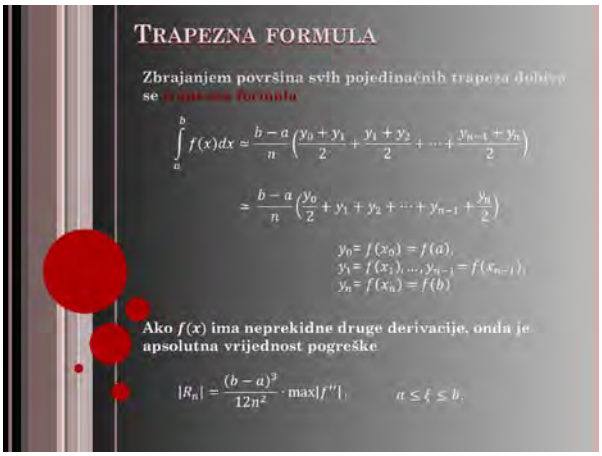
Ovaj je prezentacijski alat sastavni dio programskog paketa Microsoft Office. Zato slaganje prezentacije dodatno olakšava mogućnost unošenja teksta iz dokumenata napisanih u Microsoft Wordu i u svim Office alatima. Nastavci koji se dodijeljuju osnovnim formatima za spremanje dokumenata su pptx za PowerPoint prezentacije i ppsx za PowerPoint projekcije.

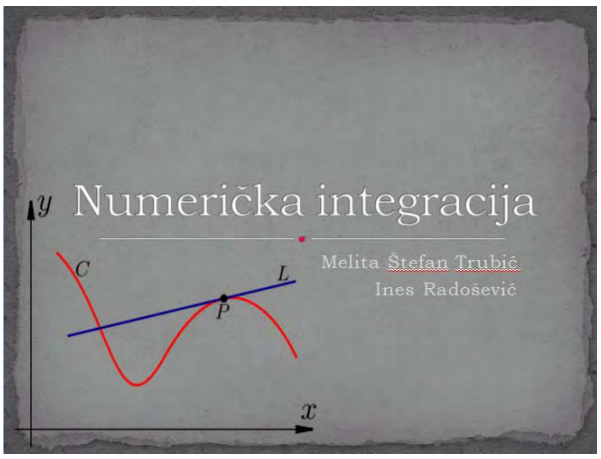
Posljednja inačica prezentacijskog alata za Windows-e je PowerPoint 2013, vidi [5]. Sa svakom novom inačicom prezentacijski alat nudi više mogućnosti. Jedan iznimno pozitivan novitet je korištenje prikaza izlagača čime se odvaja ono što vidi predavač od onoga što vidi publika, te omogućuje predavaču da tijekom predavanja koristi svoje bilješke. Prezentaciju je moguće ispisati u obliku brošura zajedno s bilješkama predavača. Prema [6] zanimljiva je i upotreba matrice stranica kojom se određuju zajedničke komponente svih stranica (isti fontovi, slike ili logotip koji može biti u obliku vodenog žiga). Bogata je ponuda niza unaprijeđenih ili novih efekata kao što su obrezivanje slika radi umetanja u oblik, dodavanje zvučnih efekata u prijelaze i animacije, animiranje različitih objekata - slika, isječaka crteža i tekstova, umetanje različitih tipova grafikona, videozapisa i glazbe, kao i mogućnost kreiranja puta kojim će se određeni objekt premjestiti po odabranoj putanji.

Veliki je pomak u PowerPoint-u 2013 učinjen i u upisivanju i u prikazivanju matematičkih izraza. Nekada su se matematičke formule pisale u Microsoft Word-ovom Equation Editor-u koji je često automatski mijenjao veličinu formule, pa čak i vrstu fonta u istom tekstu. Takve formule su se u prezentaciju unosile u obliku slika lošije kvalitete, koje nisu imale dovoljnu oštrinu, niti su se mogle lijepo oblikovati i uklopiti u ostali tekst. To je bio glavni razlog zašto su se matematičari i stručnjaci iz ostalih prirodnih znanosti radije priklonili LaTeX-ovom Beamer-u. Danas, nova inačica programa podržava relativno jednostavno umetanje složenih matematičkih izraza direktno unutar PowerPointa, pri čemu je dostignuta kvaliteta prikaza na tiskarskoj razini, slično kao u LaTeX-ovom Beamer-u. Time se Microsoft PowerPoint uspio prilično približiti Beamer LaTeX- u u prikazivanju zahtjevnih matematičkih sadržaja.

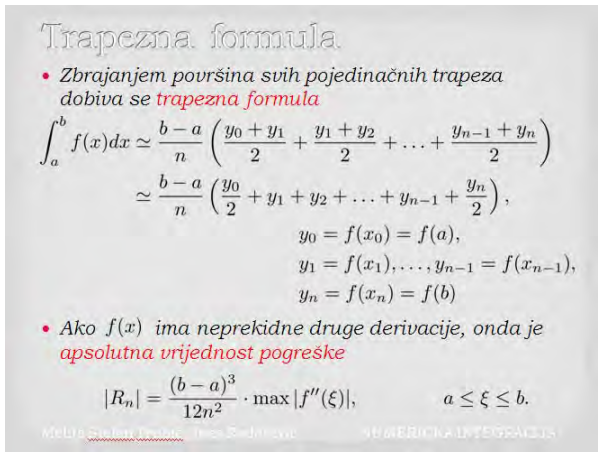
Kao što se može vidjeti na slikama 3 i 4 koje prikazuju naslovnicu i odabranu stranicu iste matematičke teme







Slika 3. Microsoft PowerPoint prezentacija "Numerička integracija" - naslovna stranica u dva predloška



Slika 4. Microsoft PowerPoint prezentacija "Numerička integracija" - odabrana stranica u dva predloška

„Numerička integracija" u dva različita predloška izrađena pomoću PowerPoint alata, razlike u kvaliteti prikaza matematičkih izraza u odnosu na Beamer LaTeX gotovo i nema.

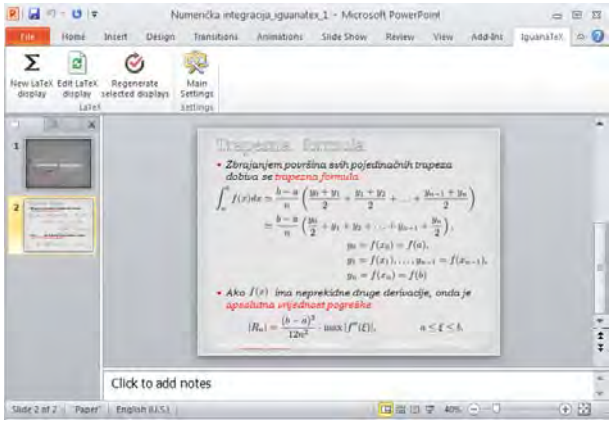
C. *Iguana Tex*

Sada kad smo iznijeli neke činjenice o Power Point-u i LaTeX-u, idemo vidjeti kako se može izraditi prezentacija kao njihov spoj. Uz pomoć dodatnog programskog alata IguanaTeX-a, vidi [7], koji se poziva iz Power Point-a, moguće je složenije matematičke izraze izraditi u LaTeX-u, slika 5. Dakle, ovim spojem bi se mogao postići maksimum u smislu dizajna prezentacije.

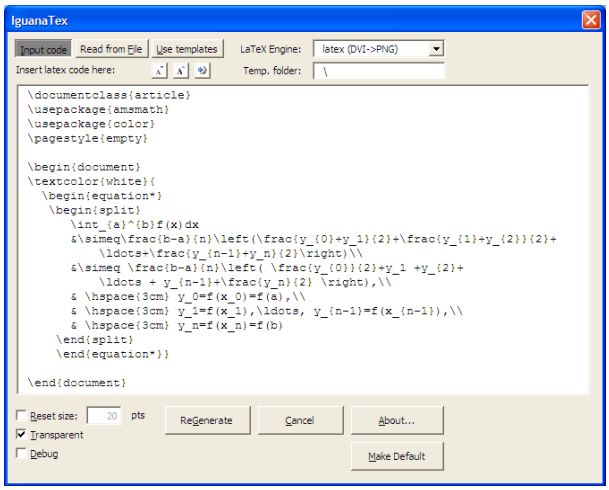
Iguana TeX je besplatan program. Prvu verziju Iguana Tex-a razvio je Zvika Ben-Haim 1978. godine, dok je za poslijednju verziju ovog dodatnog programskog alata Iguana TeX v1.47 koja je izašla krajem 2015. godine zaslužan Jonathan Le Roux, vidi http://www.jonathanleroux.org/software/iguanatex/.

Odabirom opcije New LaTeX display-a otvara se prozor, slika 6, u koji upisujemo kod formule ili ga možemo pozvati iz već postojeće TeX datoteke. Nakon procesiranja stvara se slika u PowerPoint prezentaciji. Kao i svaku drugu sliku, možemo je pomicati, rotirati, grupirati, mijenjati njezinu veličinu ili koristiti kao dio animacije. Ako trebamo učiniti dodatnu promjenu na dobivenoj slici, odnosno formuli, odaberemo istu i kliknemo na opciju Edit LaTeX display unutar IguanaTex-a, čime će se ponovno otvoriti prozor s kodom formule. Naravno da se osim formula može upisati i tekst prezentacije kao dio koda u IguaniTeX-u, ukoliko tako želimo, vidi [8].

Spremanjem prezentacije snima se i slika i njezin LaTeX kod. Snimljena prezentacija je postojana i može se pokrenuti i na računalu koje nema instaliran IguanaTeX.



Slika 5. Programski alat IguanaTex pozvan u Microsoft PowerPoint-u



Slika 6. Kod Trapezne formule u prozoru IguanaTex-a

D. *Prezi*

Linearno pomicanje sa stranice na stranicu osnova je Beamer LaTeX-a i PowerPointa. Prema [9] Prezi uvodi potpuno novi način izrade prezentacija čija je osnova zumiranje u korisničko sučelje. Prezentacijski alat Prezi razvili su Adam Somlai-Fischer, Peter Halacsy i Peter Arvai 2009. godine.

Budući je ovaj prezentacijski program dostupan preko posebnog poslužitelja govori se o usluzi u oblaku (cloud), odnosno prezentaciji na virtualnom platnu. To znači da ga mogu koristiti svi korisnici koji imaju pristup internetu i relativnu dobru vezu.

Nakon registracije na njegovoj mrežnoj stranici, vidi [10], Prezi dolazi u tri oblika, od kojih su dva osnovna dostupna na Prezi-jevoj mrežnoj stranici, a treći uz znatnu doplatu omogućava i izvanmrežno uređivanje. Besplatna verzija nudi relativno malo mrežnog prostora za pohranu prezentacije (100MB) i prezentacije su javno dostupne, tj. pristup im je slobodan. To svakako nije poželjno u akademskim krugovima, pri objavi novih znanstvenih teza i rezultata istraživanja. Uz manju godišnju doplatu, prezentacije mogu postati privatne i prostor za spremanje se može povećati, a u nešto skupljem paketu nudi se verzija na radnoj površini s dodatnim efektima i funkcijama. Dakle, ovisno o paketu, prezentacije u Prezi-ju mogu biti dostupne i preko mreže i izvan mreže.

Prezi je interaktivni prezentacijski alat čija je posebnost prezentiranje pomoću vrlo dinamičnih i nelinearnih prijelaza, vidi [11]. Prezentacija je najčešće postavljena kao jedna velika stranica ili slika na koju se unose svi elementi prezentacije, pri čemu se s jednog područja na drugo prelazi naglim uvećanjem, te se tako ističe veza između polazne veće slike i novog sitnijeg detalja te slike. Na taj se način jednostavnom nelinearnom navigacijom postiže vrlo učinkovita vizualna stimulacija. To može biti prednost pri naglašavanju određenih informacija unutar procesa poučavanja, ali i poteškoća pri koncentriranju na sadržaj predavanja. Štoviše, prebrzo prebacivanje s jednog elementa prezentacije na drugi, uz vrlo izražene vizualne efekte različitih boja, slika i popratnih zvukova kod nekih slušatelja mogu izazvati prekomjernu vizualnu stimulaciju, a time i nelagodu, vidi [12]. Zbog toga se pri izradi prezentacija treba voditi preporukama i uputama za korištenje programa propisanim od proizvođača.

Ovakvom se prezentacijom može usmjeriti predavanje tako da se ono prilagodi trenutnoj raspravi i razgovoru s publikom. Dodatno zbog svoje preglednosti, Prezi omogućava i bržu dostupnost pojedinih informacija jer je sve smješteno na jednoj stranici s koje se vrlo brzo dolazi do traženih podataka. To je svakako prednost u odnosu na linearno pretraživanje pojedinih informacija u PowerPoint-u i Beamer LaTeX-u, iako i tamo postoji prostor vidljiv samo predavaču u kojem je moguće brže doći do pojedinih informacija. Prezi ima mogućnost zajedničkog istovremenog rada više korisnika na kreiranju jedne prezentacije. Ovaj prezentacijski alat, za razliku od ostalih, ima posebno nestandardno korisničko sučelje na koje se pri izradi prezentacije treba priviknuti.

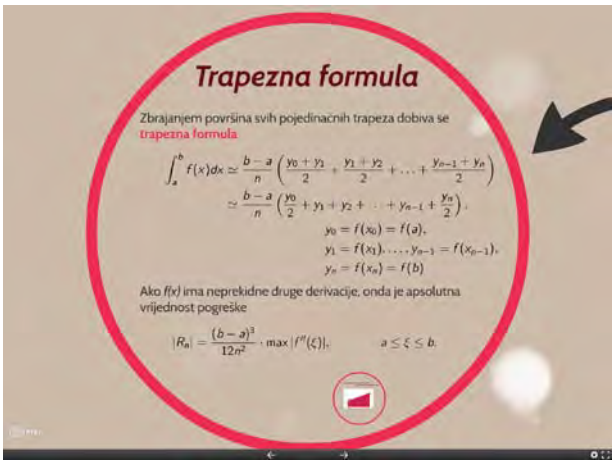
Prezi nudi mnogo predložaka različitih dizajna, uz mogućnost umetanja različitih okvira, oblika, linija i strelica. Pored teksta moguće je unošenje video i audio datoteka, pdf dokumenata i slika, ali ograničenih veličina.

Iako ovaj prezentacijski alat može bitno doprinjeti obogaćivanju, kretivnosti i kvaliteti poučavanja određene nastavne jedinice, ključni nedostatak ima upravo u prikazu matematičkih sadržaja. Prezi nije prilagođen unošenju složenih matematičkih formula, već se one moraju napisati u nekom drugom programu i u samu prezentaciju unositi kao slike ili dijelovi pdf dokumenata. Vizualno, to može biti vrlo korektno izvedeno, ali zahtijeva nešto veći anganžman oko same izrade prezentacije. Posebno ako se želi prezentirati sadržaj s mnogo složenih matematičkih izraza.

Na primjeru prikaza predavanja pod naslovom „Numerička integracija“ formule su napravljene u LaTeX-u, i unesene su u Prezi u pdf formatu, vidi slike 7 i 8. Ukoliko smo spremni zanemariti dodatni napor uložen u izradu same prezentacije, prezentiranje sadržaja pomoću Prezi-ja može biti posve novo i originalno iskustvo, i za predavača i za publiku.



Slika 7. Prezi prezentacija “Numerička integracija” - naslovna stranica



Slika 8. Prezi prezentacija "Numerička integracija" - odabrana stranica

III. ZAKLJUČAK

Predstavljanje rezultata svog rada u javnosti ili edukacija publike unutra odgojno obrazovnog procesa često zahtijeva pripremu odgovarajuće prezentacije. Pravilan odabir prezentacijskog alata pogodnog za predstavljanje određenog sadržaja neposredno ovisi o istome i bitan je za uspješno i kvalitetno predavanje. To se posebno odnosi na sadržaje iz različitih područja matematike, fizike, kemije, biologije i računarstva, koje obiluju posebnim simbolima i složenim matematičkim izrazima. Na odabranoj matematičkoj temi pokazano je da je za njihov prikaz najbolje koristiti Beamer LaTeX, što se najčešće kao prezentacijski alat koristi u znanstvenim krugovima i Microsoft PowerPoint alat koji je jako popularan kod većine predavača upravo zbog svoje jednostavnosti. Nekadašnji nedostaci i ograničenja PowerPointa pri prikazivanju formula uspješno su u novoj inačici istog programa uspješno otklonjeni. Time je PowerPoint postao prilično dobra alternativa Beamer LaTeX-u, te su matematičari dobili mogućnost gotovo ravnopravno koristiti još jedan prezentacijski alat pored Beamera. Za one koji ne bi napuštali LaTeX okruženje za prikaz formula, a opet bi željeli koristiti niz različitih mogućnosti PowerPointa, postoji Iguana TeX kao poseban program koji na jednostavan način pruža upravo takvu opciju. Primjerom je pokazano da prezentacijski alat Prezi nije osobito pogodan za prikaz matematičkih sadržaja iako daje jednu potpunu novu dimenziju pojmu prezentiranja, predstavljanjem informacija na nelinearn i vrlo dinamičan način.

LITERATURA

[1] K. Berry, "LaTeX: An unofficial reference manual," December 2015. (http://svn.gna.org/viewcvs/*checkout*/latexrefman/trunk/latex2e.pdf)

[2] https://web.math.pmf.unizg.hr/~ungar/NASTAVA/MS/msLaTeX-handouts.pdf

[3] http://bitbucket.org/rivanvy/beamer

[4] T. Tantau, J. Wright, and V. Miletić, "The BEAMER class, User Guide for version 3.36," March 2015. (http://get-software.net/macros/latex/contrib/beamer/doc/beameruserguide.pdf)

[5] J.Cox, J. Lambert, "Microsoft PowerPoint 2013," Microsoft Press, Washington, 2013.

[6] https://support.office.com/hr-hr/article/Novosti-u-programu-PowerPoint-2013-1c38822e-0284-4acb-8099-23dc6f3207c5

[7] http://www.jonathanleroux.org/software/iguanatex/

[8] http://www.technion.ac.it/~zvikabh/software/iguanatex/

[9] Prezi, "Prezi Desktop Guide," 2009-2010. (http://assets.cdn01.prezi.com/assets/desktop/Prezi_Desktop_3_Guide.pdf)

[10] https://prezi.com/

[11] https://prezi.com/4dm-fjhakau3/prezi-a-users-guide/

[12] http://eucenje.efst.hr/alat-prezi/

Siguran put do škole

Davor Šokac*, i Ivan Biuklija**
*Osnovna škola Matije Petra Katančića, Valpovo, Hrvatska
davor.sokac@skole.hr
**Osnovna škola Matije Petra Katančića, Valpovo, Hrvatska
ivan.biuklija@skole.hr

Besplatne komercijalne geografske karte na Internetu ne pokrivaju sve dijelove svijeta jednako detaljno, ne omogućuju jednostavno ažuriranje i ograničene su po tome kako se smiju koristiti, zbog čega smo učenicima predstavili OpenStreetMap (http://osm.org), otvorenu kartu svijeta, odnosno „Wikipediju" kartografije koju svatko može jednostavno dopuniti i ažurirati. Učenike smo upoznali prednostima ovog projekta, pravima korištenja i načinima ispravnog prikupljanja podataka za unos u kartu. Kod prikupljanja podataka učenici su koristili svoje mobilne uređaje s GPS prijemnikom i odgovarajuće besplatne aplikacije, dok je za unos podataka bilo dovoljno osobno računalo s pristupom internetu. Učenici su snimili GPS tragove svoje ulice, kućne brojeve i točke od interesa, te su prikupljene podatke unijeli u OpenStreetMap.

Projekt nije ograničen na ucrtavanje ulica i kuća učenika koji su trenutno uključeni u projekt, već će se kontinuirano provoditi kako bi podaci za naš grad u svakome trenutku bili ažurni i potpuni jer predstavljaju osnovni preduvjet za ostvarivanje planiranog povezanog programa obrazovanja za sigurno sudjelovanje u prometu u sklopu UN-ova Desetljeća sigurnosti cestovnog prometa.

Ovakvim projektima nastavu informatike u osnovnim školama učinit ćemo atraktivnijom, a sami učenici ponijet će više korisnih znanja i vještina koje će moći primijeniti tijekom nastavka školovanja i kasnije u životu.

I. UVOD

Svjedoci smo osjetnog pada zainteresiranosti učenika osnovne škole za izbornu nastavu općenito pa tako i za izbornu nastavu informatike. Do toga je dovelo nekoliko činjenica. Svjesni smo da današnji osnovnoškolci, i to velika većina, posjeduju vlastita računala te sami svladavaju dijelove osnova rada s računalima, upotrebe interneta i internetskih servisa, ali je to znanje kod velikog broja učenika na niskoj razini. Nezainteresiranost za svladavanjem veće i svrhovitije količine znanja i vještina predstavlja jedan od problema manje zainteresiranosti za izbor informatike kao izbornog predmeta. Učenici koji ipak odabiru izborni predmet informatike u osnovnoj školi, tijekom srednjoškolskog obrazovanja susreću se s drugim problemom: uče većinu znanja i vještina koje su već svladali za vrijeme osnovnog školovanja u izbornoj nastavi informatike – te im stoga informatika postaje dosadna.

Potaknuti takvim saznanjima odlučili smo učenike kroz uporabu besplatnih internetskih alata i izvora znanja i informacija motivirati za rad s besplatnim resursima digitalnog internetskog doba.

Osim izrade specijalizirane digitalne geografske karte područja između mjesta stanovanja i škole, učenici će usvojiti znanja o pravima korištenja i primjene informacija dostupnih na internetu. Znamo da internet pruža mnoštvo besplatnih sadržaja, ali dobar dio tih sadržaja, iako besplatan, nije slobodan, jer su načini uporabe ograničeni. Kroz ažuriranje karte na dionici između svoje kuće i škole, učenik će praktično uvidjeti važnost autorskih prava i intelektualnog vlasništva te će kod buduće uporabe internetskih sadržaja znati vrednovati tuđe, ali i informacije koje je sam sakupio, obradio i objavio.

Ovakvim pristupom prema drugim intelektualnim radovima promičemo ispravan i pozitivan stav prema objavljenim internetskim i drugim digitalnim sadržajima.

U drugom dijelu vezanom uz prometna pravila i sigurnost pješaka u prometu, učenici će usvojiti znanja o ispravnom kretanju pješaka, kako po pješačkoj stazi tako i po cesti ako na nekim dijelovima ne postoji pješačka staza, važnosti uporabe obilježenih pješačkih prijelaza i uporabe svjetlosnih prometnih znakova prilikom prelaska ceste (semafori), između ostalog. Ova će im znanja omogućiti i uočavanje opasnih mjesta na putu do škole i natrag te obilježavanje takvih mjesta na mapi i pješačkoj dionici od kuće do škole.

Ovdje smo naveli samo neka znanja i vještine koje će učenici usvojiti radeći na ovome projektu, kojim će učenici uvelike obogatiti svoje informatičko znanje i vještine te se cjelovitije pripremiti za nastavak školovanje ili svojeg budućeg životnog poziva.

II. PROVEDBA PROJEKTA

Projekt je zamišljen kroz nekoliko faza. U uvodnoj fazi učenike smo upoznali s digitalnim sadržajima, važnošću poštivanja intelektualnog vlasništva na internetu te razjasnili razliku između „slobodnih" i „besplatnih" sadržaja. Upoznali smo i razjasnili činjenicu da svi besplatni sadržaji nisu istovremeno i slobodni. Upoznali smo sve prednosti objavljivanja radova pod otvorenim licencama Creative Commons/Open Data Commons koji drugima omogućuju slobodno korištenje autorskih radova uz obvezu navođenja autora i daljnje distribucije pod istim uvjetima.

U ovome dijelu projekta dotaknuli smo se i pitanja ažurnosti i potpunosti dostupnih podataka. Na internetu, besplatno je dostupno mnoštvo različitih geografskih karata i digitalnih ortofoto snimaka bio kojeg dijela zemljine kugle, ali ostaje pitanje jesu li baš svi dijelovi svijeta jednako detaljno obrađeni i prikazani. Kod komercijalnih karata, iz razumljivih razloga, u većim urbanim područjima takve karte i prikazi znatno su

detaljniji te sadrže podatke o svim ulicama i mnogim točkama od interesa na području. U manjim središtima, besplatni izvori najčešće imaju podatke o ulicama i glavnim točkama interesa, ali rijetko imaju ucrtane kuće i sve objekte. U još manjim središtima zemljovidi su „siromašni“ takvim podacima, nisu ucrtane kuće, često nedostaju kućni bojevi, oznake pješačkih prijelaza ili se takvi podaci namjerno ne prikazuju na zemljovidima, ne nudeći pritom mogućnost da se do tih podataka ikako dođe. Ovo smo učenicima zorno prikazali pomoću alata za usporedbu zemljovida gdje se na najlakši način može vidjeti razlika.



Slika 1. Usporedba karata: OSM - Google

Zbog svega navedenog, upoznali smo učenike s otvorenim servisom za uporabu i izradu geografskih karata – OpenStreetMap, http://www.openstreetmap.org, pravima i uvjetima korištenja servisa te alatima i uređajima koji omogućuju njegovo uređivanje i korištenje.

Nakon predstavljanja potrebnog hardvera i softvera, sudionike projekta poučili smo kako da svoje mobilne telefone s odgovarajućim aplikacijama mogu koristiti na terenu za pravilno prikupljanje geografskih podataka, snimanje dionica i bilježenje točaka od interesa na putu od kuće do škole. Podatke prikupljene na taj način učenici su sa svojih mobilnih uređaja prebacili na računala, učitali u servis OSM i potom te podatke iskoristili za ažuriranje zemljovida našega grada, ucrtavajući i zanimljive lokacije, upisujući kućne brojeve, pješačke prijelaze, semafore i slično.



Slika 2. Upotreba OSM servisa – uređivanje podataka

Učenik je imao zadatak ucrtati svoju ulicu i kuću s adresom, kako bi u nastavku projekta pomoću specijaliziranih servisa za mjerenje i dijeljenje dionica mogao iscrtati svoj put od kuće do škole, očitati udaljenosti i prepoznati te označiti opasna mjesta u prometu.

III. HARDVER I SOFTVER U PROJEKTU

Kod prikupljanja podataka za ucrtavanje, učenici su koristili nekoliko specijaliziranih aplikacija za tu svrhu (OSM Tracker i Keypad-Mapper za bilježenje točaka i kućnih brojeva na sustavu Android, Go Map!! na Apple iOSu) i geotagirane fotografije, dok su unos podataka u OpenStreetMap obavljali korištenjem uređivača iD (web-aplikacija) i Java OpenStreetMap Editor (aplikacija za računala).

IV. REZULTATI

Rezultati ovakve edukacije su višestruki. Učenici svladavaju znanja iz osnova uporabe informacijsko-komunikacijskih sredstava i mobilnih uređaja. Upoznaju se s modernim digitalnim načinom stvaranja zemljovidnih karata te uče prometna pravila pješaka kao sudionika pometa. Na kraju, upoznaju se i s autorskim pravima te mogućnostima uporabe sadržaja s interneta.

Znanja stečena prikupljanjem, obradom i uređivanjem podataka omogućit će učenicima bolje planiranje, određivanje udaljenosti od točke A do točke B te kretanje gradom na siguran način.

V. ZAKLJUČAK

Ovaj projekt ne završava ucrtavanjem samo ulica i kuća učenika trenutno uključenih u projekt. Projekt će se provoditi i s narednim generacijama učenika, sve dok se ne urede sve ulice mjesta stanovanja, ali rad na uređivanju zemljovida ni tada ne prestaje. Tijekom godina dolazit će do izgradnje novih prometnica, prečaca, ili staza koje neće biti promptno ugrađene u karte komercijalnih davatelja mapa, koji najčešće ne unose i ne objavljuju promjene kontinuirano. Učenici i korisnici besplatnog servisa OSM mogu to urediti u najkraćem vremenskom odmaku od nastanka promjena. Na taj će način podaci uvijek biti ažurirani i točni.

Podatci koji nastaju na besplatnom web servisu OSM dostupni su svima za slobodnu uporabu, a točnost i pouzdanost, kao i kod Wikipedije, ovise o aktivnosti i broju suradnika, volontera.

Ovakvim projektima nastavu informatike u osnovnim školama učinit ćemo atraktivnijom, a sami učenici ponijet će više korisnih znanja i vještina koje će im vjerojatno zatrebati u nekom dijelu budućeg školovanja, ali i života.

LITERATURA

[1] OpenStreetMap, <https://en.wikipedia.org/w/index.php?title=OpenStreetMap&oldid=705371380> (Pristupljeno: 10. siječnja 2016.)

[2] Main Page. (2014, July 10). OpenStreetMap Wiki. <http://wiki.openstreetmap.org/> (Pristupljeno: 12. siječnja 2016.)

[3] Geofabrik tools. (2015). Geofabrik GmbH. <http://tools.geofabrik.de/> (Pristupljeno: 12. siječnja 2016.)

[4] Wikipedia contributors, 'Comparison of web map services', Wikipedia, The Free Encyclopedia, <https://en.wikipedia.org/w/index.php?title=Comparison_of_web_

map_services&oldid=703979556> (Pristupljeno: 15. siječnja 2016)

[5] JOSM. (2015). <https://josm.openstreetmap.de/> (Pristupljeno: 10. siječnja 2016.)

Primjena obrazovne društvene mreže Edmodo u nastavi III. osnovne škole Čakovec

Nataša Boj*
* III. osnovna škola Čakovec, Čakovec, Republika Hrvatska
natasha.boj@gmail.com

Sažetak - Obrazovna društvena mreža Edmodo stvorena je 2008. godine kako bi obrazovanje dovela u okruženje 21. stoljeća. Danas je sa svojih preko 60 milijuna članova broj 1 u svijetu obrazovnih društvenih mreža i predstavlja mjesto gdje se susreću edukacija i inovacija.
Riječ je o sigurnom online okruženju za učenike jer su korisnici Edmoda učitelji, učenici i roditelji.
Edmodo je u primjeni u III. OŠ Čakovec drugu školsku godinu za redom i koristi se za potrebe nastave matematike, informatike, hrvatskog, engleskog i njemačkog jezika, povijesti, geografije, likovne i glazbene kulture te fizike. Edmodo se koristi i za rad Vijeća učenika. Najviše se koristi za komunikaciju između učitelja i učenika u smislu slanja različitih poruka i obavijesti te za objavu domaćih zadaća.
Nadalje, učitelji koriste kvizove znanja za ponavljanje gradiva i za provjeru znanja, postavljaju ankete, daju zadatke za rad u školi ili od kuće. Učenici zadatke rješavaju na računalu i predaju na Edmodo.
Edmodo nudi i naprednije opcije od kojih je na nastavi informatike korišten rad s aplikacijama.
Provedena anketa među učenicima je pokazala kako je većina učenika zadovoljna primjenom Edmoda.

I. UVOD

III. OŠ Čakovec informatički je dobro opremljena. Škola raspolaže informatičkom učionicom s 25 osobnih računala koja su povezana lokalno i na Internet. Nadalje, u informatičkoj se učionici za potrebe nastave koristi pametna ploča, LCD projektor, printer, skener i kamera.
Informatička učionica se koristi za potrebe nastave informatike ali i za povremene potrebe nastave ostalih predmeta, za organiziranje županijskih stručnih vijeća, različite radionice u suradnji s udrugom MIS, za ogledna predavanja koja drže studenti Učiteljskog fakulteta učenicima razredne nastave.

U svim ostalim učionicama postoji jedno prijenosno računalo s vezom na Internet, pisač i LCD projektor.

Dobra tehnička opremljenost škole i veliki postotak učenika koji imaju računalo i pristup internetu od kuće znatno olakšavaju primjenu suvremenih metoda poučavanja i učenicima daju mogućnost učenja na nov, suvremen način, način koji im više odgovara i na koji lakše postižu obrazovne ishode.

Korisničko sučelje Edmoda je vrlo slično sučelju društvene mreže Facebook. Razlika je ta što je Edmodo znatno sigurnije okruženje za učenike. Naime, korisnici Edmoda su učitelji, učenici i roditelji. Sigurnosti doprinosi i činjenica da unutar grupe učenika ne postoji komunikaciju u koju nije uključen i učitelj.

Informacijsko komunikacijsku tehnologiju u školi treba maksimalno iskoristiti u nastavi i to ne samo u nastavi informatike već u nastavi ostalih nastavnih predmeta. Njezina će adekvatna primjena potaknuti aktivno učenje zasnovano na istraživanjima, razmjeni informacija, razumijevanju i primjeni istih. Tu je u središtu nastavnog procesa učenik. Učitelj ima ulogu voditelja i savjetnika, on treba motivirati učenike za rad. Takva suvremena nastava potiče kritičko mišljenje i timski rad što u konačnici rezultira boljim ishodima učenja.

Obrazovna društvena mreža Edmodo je jedan od mogućih načina primjene tehnologije za potrebe nastave. To je besplatan suradnički alat koji je nastao s ciljem povezivanja i suradnje učenika i učitelja. U početku je to bila platforma za mikrobloging, da bi se do danas pretvorio u cjelovitu platformu s brojnim aplikacijama korisnim za učenike i učitelje.

Mogućnost otvaranja korisničkog računa imaju učitelji, učenici i roditelji. Postoji mogućnost formiranja grupa, dijeljenja dokumenata, praćenja rada učenika ili određene grupe, komuniciranja s drugim učiteljima, roditeljima ali i učenicima individualno.

II. GLAVNI DIO

Edmodo je u primjeni u III. OŠ Čakovec drugu školsku godinu i danas se koristi za potrebe nastave matematike, informatike, hrvatskog jezika, engleskog jezika, njemačkog jezika, povijesti, geografije, likovne kulture, glazbene kulture i fizike.
Edmodo se u školi koristi i za rad Vijeća učenika. Članovi grupe Vijeće učenika su uz predstavnike učenika i zainteresirani učitelji, psihologinja i ravnateljica škole.

Edmodo se najviše koristi za komunikaciju između učitelja i učenika u smislu slanja različitih poruka i obavijesti od strane učitelja ili obrnuto – učenici postavljaju upite učiteljima ili članovima svoje grupe tj. članovima razreda. To su sve pitanja koja imaju veze isključivo s nastavom i aktivnostima koje se odnose na

školu. Naime, učenicima je na početku korištenja mreže naglašeno da se ista koristi samo za potrebe škole i to zaista funkcionira dobro.

Nadalje, učitelji koriste kvizove znanja za ponavljanje gradiva i za provjeru znanja, postavljaju ankete, daju zadatke za rad u školi ili od kuće. Učenici zadatke rješavaju na računalu i predaju ovdje na Edmodo. Dakle, učitelj ima sve na jednom mjestu i to u digitalnom obliku.

Učenici su dodatno motivirani za rad nagrađivanjem -dodjelom znački za što su korištene Moje učeničke značke (npr. najviše bodova u kvizu, najbolji učenički rad) i značke Edmoda.

Kako bi učenicima obogatio obrazovno iskustvo Edmodo nudi aplikacije iz široke zajednice izdavača aplikacija i pruža pristup aplikacijama treće strane koje pokrivaju razna područja. Uporaba ovih aplikacija je izvrstan dodatak nastavi, korisne su za učenje novih sadržaja kao i za pregledavanje i praćenje učeničkih postignuća. Privlačne i jednostavne za korištenje ove aplikacije odgovaraju digitalnoj generaciji učenika, a mogu se povezati s temeljnim mogućnostima Edmoda kao što su dodjeljivanje znački za nagrađivanje učenika, primjena zadataka. Učenici trebaju samo svoje Edmodo korisničke podatke za prijavu kako bi pristupili aplikaciji koju je učitelj dodao grupi. Edmodo Store je središnje mjesto za obrazovne aplikacije gdje trenutno nude razne aplikacije koje pokrivaju širok spektar obrazovnih materijala, besplatne i premium verzije.

A. Primjeri uporabe Edmoda u III. OŠ Čakovec

U nastavku su navedeni neki konkretni primjeri uporabe Edmoda.
Primjer 1: Biologija – popis tema za referat i upute
Primjer 2: Biologija – popis učenika s odabranom temom, termini za predaju radova, upute za predaju referata i prezentacije
Primjer 3: Hrvatski jezik – domaća zadaća, priprema za pisanu provjeru znanja
Primjer 4: Hrvatski jezik – analiza testa, zadaci s rješenjima
Primjer 4: Njemački jezik – podsjetnik za DOP, domaću zadaću i skoru pisanu provjeru znanja
Primjer 5: Njemački jezik – radni listići za vježbanje, planovi ploče
Primjer 6: Engleski jezik – komunikacija učenica-učiteljica, zamolba učenice za odgodom usmenog odgovaranja uz navedeni razlog, zamolba uvažena od strane učiteljice
Primjer 7: Informatika – poruka učenicima – izraz nezadovoljstva prezentacijama koje su napravili učenici, objašnjenje što se očekuje ubuduće, primjer dobre prezentacije koja je ocjenjena s odličnim
Primjer 8: Informatika – obavijest učenicima polaznicima dodatne nastave o školskom natjecanju iz informatike
Primjer 9: Informatika – objava edukativnog videa o sigurnosti na internetu
Primjer 10: Informatika – pregled zadaća koje su predali učenici s povratnom informacijom o ocjeni
Primjer 11: više predmeta, prikaz znački koje je dobio učenik od određenih učitelja za svoja postignuća
Primjer 12: Vijeće učenika – objava zapisnika s održanog sastanka Vijeća učenika
Primjer 13: Informatika - objava ankete o primjeni Edmoda – za učenike
Primjer 14: Informatika – objava rezultata ankete o primjeni Edmoda

B. Rezultati istraživanja o primjeni Edmoda vezanog za učenike

U školi je provedena anketa među učenicima o zadovoljstvu primjene obrazovne društvene mreže Edmodo.
Uzorak: 141 učenik (5. – 8. razred)
Razdoblje provođenja ankete: 11.12. – 15.12.2015.
Anketa je provedena putem Edmoda, na zid je stavljena poveznica do ankete izrađene u Google okruženju.

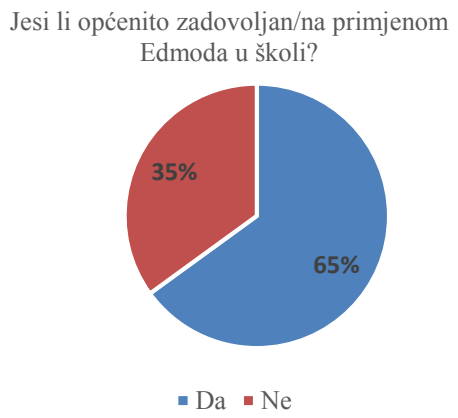
Po završetku ankete, rezultati su također objavljeni na zidu Edmoda i bili su dostupni svim učenicima, učiteljima i roditeljima, korisnicima Edmoda. Rezultati ankete pokazuju kako je 65% učenika zadovoljno primjenom Edmoda. 35% učenika nije zadovoljno iz razloga što sada trebaju više raditi na računalu pa im manje vremena ostane za zabavu.

Osim odgovora većine učenika koji su pozitivni, na to upućuju i poruke koje su upisivali na kraju ankete. Iz njihovih poruka se saznaje da im nedostaje Edmodo na nekim nastavnim predmetima, jer nažalost, nisu svi učitelji korisnici Edmoda.

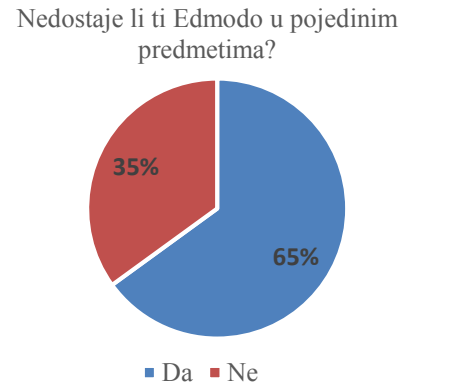
Poruke učenika:

- „Edmodo je super“
- „Edmodo je stvarno koristan“
- „Edmodo je zakon“
- „Edmodo je jako koristan za one koji ne znaju što je za dz
- „Neka učitelji koji nemaju Edmodo, naprave grupu“
- „Sviđa mi se da mogu predati zadaću“
- „Edmodo je korisna društvena mreža za sve potrebe u vezi škole“
- „Sviđa mi se što Edmodo mogu koristiti i kod kuće“
- „Svi moji digitalni radovi su na jednome mjestu – Biblioteci Edmoda“
- „Ako nešto ne stignem napraviti u školi, stavim na Edmodo i dovršim kod kuće“
- „Edood je jednostavan za korištenje“

Slika 1 prikazuje da je primjenom Edmoda u školi zadovoljno 65% učenika a nezadovoljno 35% učenika. Slika 2 prikazuje da isti postotak učenika koji je zadovoljan primjenom Edmoda isto tako nije zadovoljan time što Edmodo nije zastupljen i u nekim drugim premetima. To isto su pokazale i poruke učenika iz kojih se vidi da mole učitelje tih predmeta da počnu koristiti Edmodo. Pojedini kolege ne koriste Edmodo za potrebe nastave iako bi većina anketiranih učenika to željela.



Slika 1. (Ne)zadovoljstvo učenika primjenom Edmoda



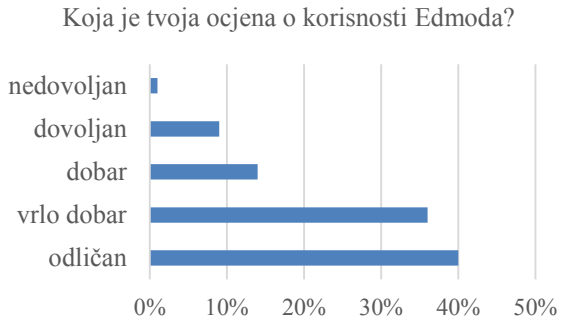
Slika 2. Nezadovoljstvo većine učenika neprimjenom Edmoda u nekim predmetima

Slika 3 prikazuje da 40% anketiranih učenika ocjenjuje korisnost Edmoda za potrebe njihove nastave ocjenom odličan, preko 30% učenika je dalo ocjenu vrlo dobar. Manji postotak učenika smatra da korištenje Edmoda u nastavi ne zaslužuje visoke ocjene.

Slika 4 pokazuje da je učenicima najbitnije da je na Edmodu objavljeno što je za domaću zadaću kao i to što tu istu zadaću mogu vratiti u digitalnom obliku na Edmodo. Ovo je posebno važno za učenike koji nisu bili u školi jer tako na jednostavan način mogu saznati što se radilo kao i što je za domaću zadaću.

Zatim je učenicima bitno što mogu biti u komunikaciji s učiteljem kao i to što su im dostupne različite obavijesti. Manje im je bitna komunikacija s članovima grupe kao i nagrađivanje značkama. Pojedini učenici vole nagrađivanje značkama jer ih to dodatno motivira na rad. To su većinom učenici petih razreda.

Iz grafičkog prikaza na slici 4 je vidljivo da je učenicima manje bitno to da roditelji prate što oni rade u školi. Istraživanje je pokazalo da manji postotak roditelja ima otvoreni korisnički račun na Edmodu, njih 20%..



Slika 3. Ocjena učenika o korisnosti Edmoda

III. ZAKLJUČAK

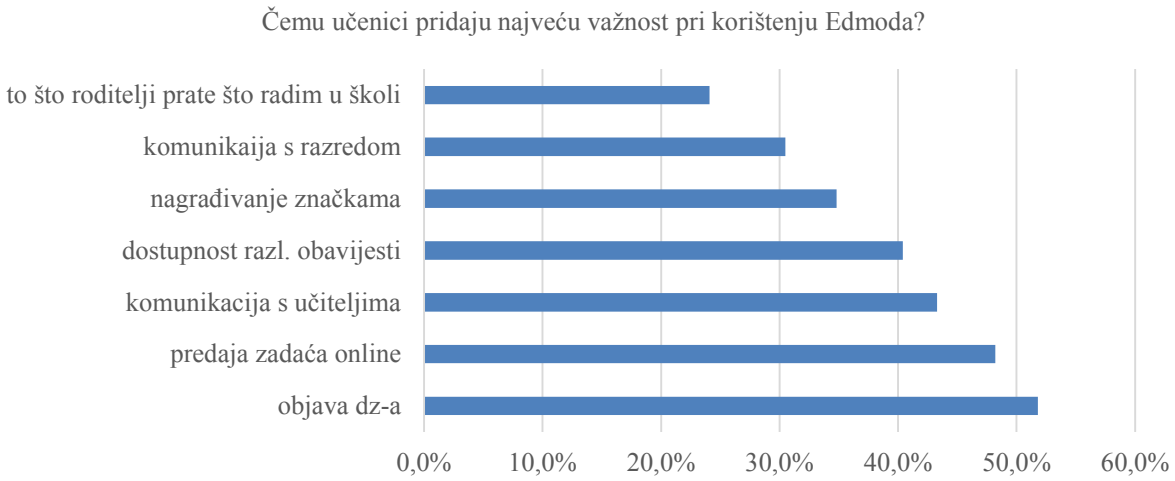
Može se zaključiti kako je način na koji učenici uče jednako važan kao i ono što uče. S obzirom da živimo u digitalnom dobu, a naši učenici su korisnici različitih tehnologija koje, ako im se ne ponudi ništa bolje, uglavnom koriste za razonodu, zadatak učitelja je taj da učenicima omogući da jedan veći dio provedenog vremena za računalima, mobitelima i sl. uređajima provedu upravo koristeći ih za edukaciju.

Edmodo je dobar primjer iskoristivosti informacijsko komunikacijske tehnologije za potrebe nastave. Jednostavan je za korištenje, dostupan je i na hrvatskom jeziku. Edmodo nudi brojne mogućnosti: razmjenu ideja, dijeljenje nastavnih sadržaja, povezivanje s nastavnicima diljem svijeta, dijeljenje poruka unutar grupa, individualni razgovori, dijeljenje slika i video zapisa, bilježenje aktivnosti u kalendaru, formiranje osobne biblioteka s raznim adresama web stranica i dokumentima, praćenje rada učenika i objedinjavanje svih učeničkih radova na jednome mjestu, kreiranje kvizova i anketa čiji se odgovori automatizmom statistički obrađuju.

Edmodo je prikladan za korištenje u svrhu suradnje nastavnika unutar škole, unutar Županijskih stručnih vijeća, nastavnika sudionika u projektima i sl.

Istraživanje je pokazalo da je većina učenika, njih 65%, zadovoljno primjenom Edmoda u nastavi. Kao najveću prednost Edmoda učenici vide u tome što su im nastavni materijali dostupni i nakon izlaska iz učionice kao i to što zahvaljujući komunikaciji s učiteljima i članovima grupe (razreda) mogu doći do svih potrebnih informacija u vrlo kratkom vremenu. Učenicima se sviđa mogućnost predaje radova u digitalnom obliku i objedinjenost svih radova ne jednome mjestu.

Jasno je da tehnologija ne može zamijeniti učitelje, ali je samo pitanje vremena kada će učitelji koji koriste tehnologiju zamijeniti one koji ju ne koriste.



Slika 4. Što je učenicima bitno prilikom korištenja Edmoda

LITERATURA

[1]. Grupa autora: „Suradničko učenje i Edmodo“ – priručnik, CARNet, AZOO, MZOS, siječanj 2015.
[2]. Lidija Kralj – „Edmodo – američko rješenje za društvenu obrazovnu mrežu“, (lidija-kralj.forum.hr), veljača 2015.

Prilozi:

1. Prezentacija – primjeri uporabe Edmoda u III. OŠ Čakovec

Digitalni scenariji učenja

Marina Mirković, dipl. inženjer elektrotehnike
Tehnička škola, Požega, Republika Hrvatska
marina.mirkovic@po.t-com.hr

Sažetak

Digitalni scenariji učenja pružaju učenicima mogućnost da primjenjuju znanje i praktične vještine bitne za situaciju koju nalaze pri rješavanju određenih zadataka. Pri tome analiziraju, prepoznaju i rješavaju probleme uz vještine kritičkog razmišljanja. Uče uporabom strukturiranih aktivnosti učenja koje nalikuju na izazove slične onima s kojima će se suočiti u stvarnom svijetu.

Nastava uz primjenu scenarija učenja pomaže učenicima da postanu svjesni propusta u učenju, lakše prebrode teškoće u učenju te promijene svoje ponašanje u određenim situacijama.

Digitalni scenarij učenja pretvara podatke u praktičnu primjenu, predstavlja zadatke u vezi s ishodima učenja. Kako bi nastavnici izgradili učinkovit i zanimljiv scenarij, potrebno je integrirati sadržaj, metodiku i razne digitalne alate. Oni omogućuju stvaranje privlačnog interaktivnog scenarija koji će biti zanimljiv učenicima u njihovu radu na nastavi uz suradnju sa svojim vršnjacima, te ujedno omogućiti razvoj njihovih digitalnih kompetencija.

Hrvatska akademska i istraživačka mreža – CARNet u sklopu projekta Inspiring Science Education, kojemu je cilj popularizacija znanosti uz korištenje tehnologije i dostupnih e-resursa, pokrenula je ISE MOOC. U ovom masovnom otvorenom online tečaju polaznici osmišljavaju nastavni sat uz primjenu informacijsko-komunikacijske tehnologije. ISE digitalni scenariji učenja su strukturirani prema metodi istraživačkog učenja. Primijenjivi su u različitim kontekstima te pogodni za razmjenu među nastavnicima na školskoj, državnoj i međunarodnoj razini.

I. UVOD

Nastava putem digitalnog scenarija učenja može biti učinkovit način kako učenici mogu primjeniti ono što su naučili u realnim situacijama. Prikladno je kada želimo da učenici primjenjuju svoje znanje i vještine kritičkog razmišljanja za donošenje odluka, rješavanje ili istraživanje problema. Ova tehnika može se koristiti tijekom online tečaja kao i u direktnoj nastavi licem u lice.

Digitalni scenariji su odličan način edukacije kojim nastava i primjena online tečaja postaju privlačniji. Pomažu učenicima u učenju na dubljoj razini. Pri tome čitaju s namjerom da razumiju i postave se u određenoj situaciji, promišljajući o postupcima i donošenju svojih odluka. Scenariji time pretvaraju podatke u praktiču primjenu [1].

II. ULOGA SCENARIJA UČENJA

Scenarij predstavlja proces učenja u kojem su pedagoški precizno definirani uvjeti i ishodi učenja. Nastavnik u kreiranju scenarija učenja dizajnira svoje pedagoške i metodičke ideje kroz ulogu organizatora odnosno režisera koji vodi brigu o svojoj i motivaciji učenika te međusobnom partnerstvu u pedagoškoj interakciji. Pri tome je važan timski rad više nastavnika koji mogu zajedno razraditi ideju i scenarij.

Pri stvaranju učinkovitih digitalnih scenarija učenja potrebno je analizirati elemente među kojima su:

- obrada i povezivanje sadržaja putem odgovaranja na pitanja ili rješavanja zadatka
- uporaba modela iz stvarnog života s problemima odnosno zadacima koje će učenik moći izvršiti
- uporaba zadataka vezanih za ishode učenja (učenik posjeduje kognitivne i praktične vještine u svom području te može koristiti to znanje i vještine u rješavanju posebnih problema iz svoga područja, ocjenjuje vlastite kompetencije te poboljšava vlastiti rad ili učenje)
- događaji odnosno aktivnosti čine logički redoslijed
- sadržaj je točan, relevantan i prikladan.

Scenarij je odraz stvarne situacije i sadrži sve složene elemente potrebne za pristup određenom problemu pri čemu učenici uče kroz rad, svojim aktivnostima razvijaju vještine i znanja koji su im potrebna. Može imati oblik priče, u kojoj postoji zaplet, likovi i problem koji će učenici prepoznati, analizirati, i/ili riješiti. Ponekad može imati oblik igre. Uz visoku interaktivnost, poželjno je da učenici surađuju sa svojim vršnjacima.

Nastava uz primjenu scenarija učenja pomaže učenicima da postanu svjesni praznina u učenju, lakše prebrode teškoće u učenju, promijene svoje ponašanje u određenim situacijama. Isto tako scenariji mogu pomoći učiteljima i mentorima u opisivanju teških koncepata.

Da bi izgradili učinkovit i zanimljiv scenarij potrebno je integrirati sadržaj, metodiku i razne digitalne alate kojima stvaramo privlačan interaktivan scenarij zaimljiv učenicima u njihovu radu na nastavi.

Digitalni scenariji učenja nastavniku daju ideju za provedbu nastavnog sata uz pomoć tehnologije. Nastavnicima su strukturirana priprema za izvođenje nastavnog sata a učenicima plan i uputa za provođenje nastavnih aktivnosti. Pogodni su za razmjenu među nastavnicima različitih razreda, predmeta, škola i država.

Učenici su motivirani za učenje zahvaljujući činjenici što znaju da posjeduju vještine potrebne za rješavanje zadatka. Izazvani su, ali svjesni da imaju alate za susret s izazovom i nastoje ga uspješno svladati. Umjesto nastavnih materijala iz knjige ili putem predavanja, učenici sudjeluju u scenariju u kojem mogu izravno vidjeti kako će vještine i znanje koristi dugoročno. Dobar scenarij se temelji na aktivnosti ili zadataku u kojem će se učenici suočiti sa stvarnim preprekama i problemima kako bi proširili svoje vještine i bazu znanja. Uz to je potrebno zadržati njihovu motivaciju za sudjelovanje u procesu učenja [2].

Pri kreiranju digitalnog scenarija učenja potrebno je uključiti što više interaktivnih elemenata. Učenicima treba omogućiti da izaberu put koji ih vodi kroz pitanja i odgovore. Važno je da su scenariji realni i informativni, jednostavni i dovoljno jasni kako bi se izbjeglo preopterećenje informacijama.

Djeca uče iz primjera. Inspirirani su pričama borbe, teškoća, hrabrosti uz pobjede stvarnih heroja koji ih motiviraju da ih oponašaju. Virtualni scenariji pružaju učenicima stručnost i iskustvo kroz treninge koji im omogućuju učenje kroz proces pokušaja i pogreški, bez suočavanja s posljedicama ili troškovima pogrešnih odluka.

Postoji nekoliko vrsta scenarija. Svaki tip je pogodan za podučavanje određene vrste vještina. U scenariju koji se temelji na vještinama, od učenika se očekuje da će pokazati vještine i znanje koje je već stekao. Problemski scenariji idealni su za situacije u kojima učenici moraju integrirati svoje teorijsko i praktično znanje u istraživanju problema. Odlučivanje, logično zaključivanje i kritičke analize sastavni su dio tih scenarija. Kod scenarija procjenjivanja učenici moraju predvidjeti ishod nekog događaja u budućnosti na temelju njihovog znanja. Scenariji igre uključuju uporabu igre kao alata za učenje [3].

Likovi i aktivnosti u digitalnom scenariju učenja trebaju potaknuti učenike na pozitivne promjene u stjecanju znanja, vještina ali i u ponašanju. Nužne su povratne informacije putem učinkovitih nastavnih alata. Učenicima treba objasniti pogreške u radu te predložiti alternativne načine kako doći do rješenja problema ili zadatka.

III. ELEMENTI SCENARIJA UČENJA

Scenariji učenja sadrže elemente koji opisuju način, mjesto odvijanja i primjenjena nastavna sredstva. Među podacima nalazimo:

Osnovne informacije:

- kratki opis scenarija
- područje
- dobna skupina i razred
- koliko je vremena potrebno za realizaciju scenarija učenja
- prikaz digitalnih materijala koji će se koristiti
- lokacija gdje se scenarij odvija (npr. školska učionica, lokalna knjižnica, na otvorenom prostoru, online)
- vrijeme događanja scenarija (npr. tijekom školskih sati, izvan nastave)

Pedagoške informacije:

- ishodi učenja
- nastavne metode
- vještine koje se razvijaju scenarijem
- sudionici - tko je uključen u scenariju (npr. nastavnici, učenici, studenti, školski knjižničari, članovi zajednice, vanjski stručnjaci, koje su njihove uloge)
- razina interaktivnosti (koje vrste interakcije postoje između sudionika u scenariju)
- pristup poučavanju.

Tehničke informacije:

- web adresa digitalnog/ih materijala ili aplikacije koja se koristi
- oprema potrebna za izvedbu nastavnog sata
- tehnologija (što se i kako od tehnologije koristi u scenariju).

Informacije o sadržaju:

- popis aktivnosti učenja.

IV. IZRADA DIGITALNOG SCENARIJA UČENJA

Postoji više različitih načina kako stvoriti scenarije učenja, od tekstualnih studija slučaja do interaktivnih sadržaja odnosno simulacija. Bez obzira na resurse koje imamo na raspolaganju, na učinkovite načine možemo aktivirati učenike u scenarijima kako bi što više pridonijeli njihovom učenju.

Potrebno je odlučiti o obliku scenarija, da li se koristi za nastavu s pristupom lice u lice ili za online okruženje. Nužno je analizirati medije (fotografije, audio, video) i druge resurse koji će biti potrebni. Online scenariji učenja omogućuju dodatne aktivnosti kao što su forumi i wiki.

Početna točka scenarija je prepoznavanje događaja ili situacije, a potom kjučnih područja za povratne

informacije i učeničke refleksije. Izrada plana scenarija je učinkovit način pripreme. Zamolite kolege za pomoć u analizi i praćenju aktivnosti kroz scenarij kako bi osigurali da se odvija prema vašim očekivanjima i postiže rezultate koje ste namjeravali.

Jednostavan scenarij u kojem se koristi priča stvara situaciju u kojoj učenici mogu koristiti raspravu kako bi podijeliti svoja tumačenja i odluke. Uporaba slika može biti učinkovit način za poboljšanje scenarija. Prateći sadržaje slika učenici mogu identificirati određene probleme koje će poslije analizirati i rješavati.

Sofisticiraniji scenariji imaju ugrađene određene interakcije i mogućnost višestrukog izbora. Oblik grananja koristi se često u kvizovima. Kada učenici odaberu pogrešan odgovor, dobivaju poruku koja objašnjava zašto je odgovor pogrešan i usmjerava ih pregledati određene materijale u scenariju.

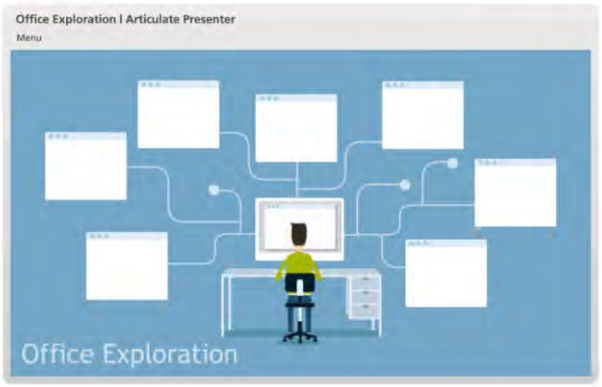
Model "3C" (challenge, choices, consequences) karakterizira izazov, izbor, a potom posljedice izbora. Učenik donosi odluku, a zatim odabire odgovarajući izbor koji stvara posljedice. U tom trenutku nastavnik može dati povratnu informaciju i pustiti učenika u novi izazov u nastavku učenja. Nastavnik može dodati novu 3C strukturu na kraju posljedica. Umjesto dobivanja povratnih informacija, učenik može direktno prijeći u novi izazov/aktivnost. Posljedica izbora učenika stvara novi put [4]. Kod kreiranja složenih scenarija koji će pružiti zanimljivo okruženje za učenje treba dobro razumjeti predmet odnosno temu problemskog učenja. Razgranati i interaktivni scenariji moraju biti relevantni i smisleni kako kod učenika ne bi izazvali nezainteresiranost i gubljenje vremena.

V. ALATI ZA IZRADU SCENARIJA UČENJA

Način izrade scenarija učenja ovisi o raspoloživim resursima i sadržaju. Temelj je dijagram tijeka razvoja problema i aktivnosti, gdje je učenik usmjeren prema pitanjima i odlukama definiranim u scenariju. Put analize i rješavanja koje će učenik primjeniti ovisi o nastavniku. Dijagram tijeka i sadržaja scenarija nastavnik može pripremiti pomoću umnih mapa. Na brz način organizira svoje misli prilikom izgradnje tijeka svog zamišljenog scenarija. Ideje pikazane umnom mapom lako se mogu prenijeti u slajdove PowerPoint prezentacije [5].

Scenarij treba biti izazovan, kako učenici ne bi lako riješili zadatke odnosno probleme. Važno je odrediti okruženje za scenu. Hoće li to biti učionica, ured, radionica? Potrebno je prikupiti materijale koji će se koristiti u scenariju, kao što su slike, grafikoni i drugi multimedijski sadržaji. U PowerPoint prezentaciji jednostavan je pregled sadržaja i uređivanje. Klikom na strelice omogućeno je kretanje kroz slajdove scenarija. Učenici donose odluke klikom na gumbe unutar scenarija, a na temelju tih odluka prezentiraju se odgovarajuće povratne informacije. Dostupni su besplatni PowerPoint predlošci kao osnova za kreiranje scenarija učenja [6].

Na slici 1. prikazan je predložak za scenarij koji korisniku omogućuje istraživanje niz interaktivnih ekrana. Klikom na određeni zaslon u glavnom izborniku učenik prelazi na određeni modul scenarija.

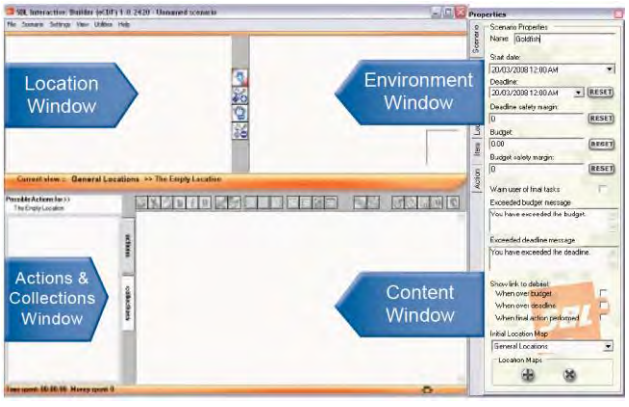


Slika 1. PowerPoint predložak scenarija učenja

SBLi (Scenario Based Learning Interactive) je besplatni eLearning softver za stvaranje i isporuku multimedijskih scenarija problemskog učenja putem interneta. Oni omogućuju učenicima sudjelovanje u novom obrazovnom iskustvu.

Novorazvijeni paket alata dizajniran je kako bi se omogućilo nastavnicima, predavačima i ostalima koji rade u obrazovanju, stručnom osposobljavanju ili usavršavanju, stvaranje i isporuku interaktivnih scenarija temeljenih na problemskom učenju kao pomoć nastavnim metodama. Softver uključuje:

- Modul kreiranja - za stvaranje scenarija (slika 2). Osmišljen je kako bi proces stvaranja scenarija bio jednostavan i lak.
- Modul izvođenja - za predstavljanje stvorenih scenarija učenicima, omogućujući im interakciju kroz scenarij (slika 3). Postoje dvije vrste izvođenja: offline prikaz putem CD-a (ne zahtijeva pristup internetu) i online prikaz scenarija preko interneta putem web preglednika.



Slika 2. Sučelje modula kreiranja SBLi scenarija učenja



Slika 3. Sučelje modula izvođenja SBLi scenarija učenja

SBLi scenariji omogućuju okruženje za učenje usmjereno na određeni problem ili područje istraživanja. Scenariji se mogu koristiti kako bi korisnik postao svjestan manjih problema koje treba uzeti u obzir u rješavanju određene vrste složenih problema. Pri tome testira svoje dijagnostičke vještine u određivanju uzroka određenog problema ili sudjeluje u studijama slučaja koje zahtijevaju od korisnika da se uključe u niz zadataka [7].

Mogu ih koristiti učenici i korisnici svih uzrasta. SBLi pruža mogućnost za stvaranje interaktivnog scenarija koji omogućuje učenicima da istražuju različite lokacije i okruženja te poduzimaju razne aktivnosti u pokušaju da se riješi problem ili zadatak. Iskustvo učenja stečeno kroz interakciju u scenarijima može pomoći učenicima shvaćanje principa, stjecanje znanja potrebnih za uspješno razumijevanje i dijagnosticiranje. Kritički promišljaju i analiziraju probleme, uključujući iterativni ciklus istraživanja, promišljanja i stvaranje hipoteza.

Ako su scenariji povezani s izvornim materijalima (kao što su web poveznice, Word dokumenti, Excel proračunske tablice, knjižnice resursa i slično), učenici će moći istražiti brojne izvore informacija i znanja.

SBLi scenariji razvijeni su u sljedećim disciplinama: poljoprivreda, arhitektura i izgradnja, zdravlje-njega, veterina, trgovina, ugostiteljstvo, geografija, informacijska tehnologija, obrazovanje, socijalni rad, inženjering, pravo, upravljanje okolišem, poslovanje i ostalo.

Edukata je suradnički proces projektiranja kojim nastavnici mogu povezati vizionarske scenarije i nastavnu praksu. Dizajnerski proces Edukata je izvor inspiracije za nastavnike koji žele ojačati svoje sposobnosti i samopouzdanje kao dizajneri budućih aktivnosti učenja u učionici te proširiti i poboljšati obrazovne dizajnerske vještine. Inspirativni scenariji su vodič nastavnicima za promišljanje i oblikovanje budućih aktivnosti učenja koje odražavaju ključne aspekte scenarija, osobne interese i potrebe učenika [8]. Primjeri aktivnosti kod stvaranja određenog proizvoda ili projekta (slika 4) mogu biti: predstavljanje problema, razumijevanje kroz ispitivanje, promatranje i prikupljanje podataka, stvaranje mentalnih mapa kako bi razumjeli odnose između prikupljenih informacija, promišljanje kroz snimanje audio-vizualne refleksije, stvaranje i dizajniranje rješenja, izvođenje radionice prikupljanja povratnih informacija o svom radu od stručnjaka, osoba koje mogu predstavljati buduće korisnike, objavljivanje i prezentiranje rješanja, suradnja s učenicima drugih škola.



Slika 4. Razvoj aktivnosti učenja u procesu Edukata

Tijekom radoblja od 2011. do 2013. godine, učenje kroz aktivnosti osmišljene procesom Edukata bio je pilot projekt u više od 17 europskih zemalja [8]. Analizom je utvrđeno da su aktivnosti pozitivno utjecale na učenje te postizanje i poticanje vještina 21. stoljeća (komunikacija na materinjem jeziku; komunikacija na stranim jezicima; matematička kompetencija i osnovne kompetencije iz prirodoslovlja i tehnologije; digitalna kompetencija; sposobnost učenja; socijalne i građanske kompetencije; smisao za inicijativu i poduzetništvo; kulturološka osviještenost i izražavanje).

VI. ISE SCENARIJI UČENJA

Hrvatska akademska i istraživačka mreža – CARNet u sklopu projekta Inspiring Science Education (ISE), kojemu je cilj popularizacija znanosti uz korištenje tehnologije i dostupnih e-resursa [9], pokrenula je ISE MOOC (masovni otvoreni online tečaj).

Sastoji se od dva dijela:

- Prvi dio - Osnove korištenja obrazovnih trendova i digitalnih alata u nastavi
- Drugi dio - Napredna primjena obrazovnih trendova i digitalnih alata u nastavi.

U tečaju polaznici osmišljavaju vlastiti nastavni sat uz informacijsko komunikacijsku tehnologiju i primjenu stečenog znanja. Upoznju se s otvorenim obrazovnim sadržajima i izradom digitalnih scenarija istraživačke nastave [10].

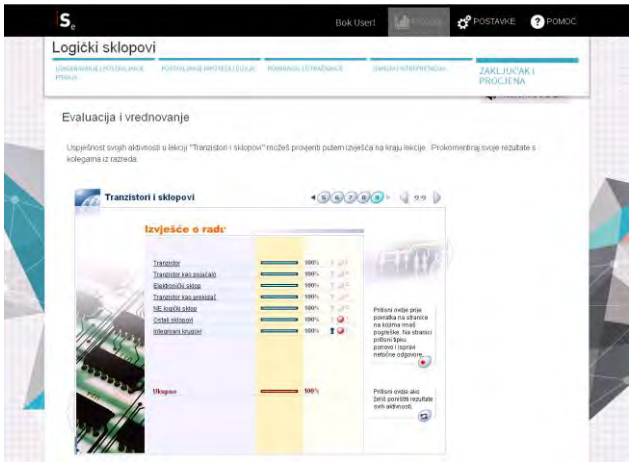
ISE alat za izradu digitalnih obrazovnih sadržaja, nastavnih materijala i scenarija učenja konstruiran je na temelju faza istraživačkog učenja. Sadržaj dijeli na pet stranica odnosno pet faza istraživačkog učenja: orijentacija i postavljanje pitanja, postavljanje hipoteza, planiranje i istraživanje, analiza i interpretacija te zaključci i evaluacija (slika 5). ISE scenarij učenja koji se njime izrađuje je prilagođen istraživačkom učenju. Alat omogućuje nastavniku upisivanje bilješki koje će samo on vidjeti prilikom primjene scenarija učenja u nastavi.

ISE digitalni scenariji učenja objavljuju se na ISE portalu [11] te ovisno o postavkama privatnosti, postaju dostupni svim nastavnicima koji žele svoju nastavu osuvremeniti upotrebom digitalnih alata posebno za prirodoslovna područja.

Nastavnici mogu pregledavati scenarije učenja drugih autora, a uz njihovo dopuštenje mogu ih kopirati odnosno klonirati te potom izmijeniti prema svojim potrebama. Važan dio scenarija učenja su detaljno razrađeni metapodaci scenarija koji omogućuju lakše pretraživanje na ISE portalu.

Alat za izradu scenarija istraživačkog učenja je otvoreni obrazovni sadržaj jer omogućuje korisniku ponovnu uporabu sadržaja i prilagodbu prema kulturnim, pedagoškim i kurikularnim specifičnostima. Scenariji učenja doprinose mješovitom učenju kombiniranjem online i licem-u-lice pristupa.

Jedan od važnijih dijelova ISE scenarija su i pitanja za provjeru znanja pomoću kojih učenici mjere svoje kompetencije. Nakon ispunjavanja učenik dobiva povratnu informaciju o točnosti svojih odgovora. Ova pitanja se dakako mogu kreirati i kao pitanja za samoprocjenu znanja. Alat među ostalim pruža nastavnicima detaljan statistički pregled ispunjavanja testova za svakog pojedinog učenika te prikaz vremenskog trajanja posjeta pojedinim stranicama scenarija. Sučelje alata je na hrvatskom jeziku.



Slika 5. Evaluacija i vrednovanje u ISE scenariju učenja - izvješće o radu učenika u lekciji o Tranzistorima i sklopovima

Nastavnik kod izrade scenarija učenja započinje njegovim osmišljavanjem. Prvo odlučuje o temi scenarija a zatim prikuplja sadržaje u obliku tekstova, fotografija, video isječaka ili poveznica na primjerice simulacije procesa koje želi prikazati. Uz to razmišlja o digitalnom alatu kojeg želi koristiti u podučavanju. Sadržaj scenarija potrebno je razraditi prema pet faza istraživačkog učenja:

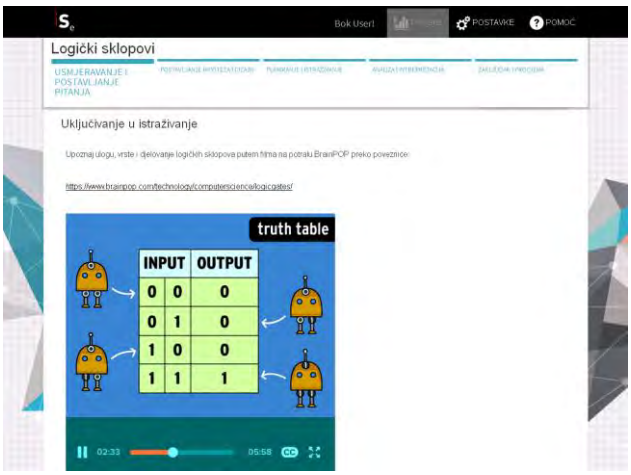
1. uključivanje u istraživanje postavljanjem znanstvenih pitanja
2. istraživanje
3. oblikovanje objašnjenja temeljenog na znanstvenim dokazima
4. prikaz i potvrđivanje znanstvenih spoznaja
5. evaluacija/ vrednovanje.

Osim pripreme sadržaja scenarija i osmišljavanja istraživačkih aktivnosti potrebno je definirati problemska pitanja za učenike.

Edukativne materijale koji se koriste u scenariju moguće je besplatno preuzeti s raznih portala s digitalnim obrazovnim sadržajima. Nacionalni portal za udaljeno učenje Nikola Tesla je repozitorij materijala za samostalno učenje ili za upotrebu u nastavi koji su strukturirani u tečajeve u raznim kategorijama.

BrainPOP je zabavni portal namijenjen djeci osnovnoškolskog i srednjoškolskog uzrasta koji sadrži razne animacije, videa i interaktivne igre (slika 6) koje se baziraju na STEM predmetima kao i engleskom jeziku, glazbi, umjetnosti i društvenim temama [12].

PhET interaktivne simulacije koje su nastale na Sveučilištu u Coloradu pojašnjavanju procese iz fizike, biologije, kemije, kao i nastavne sadržaje iz matematike i geografije [13].



Slika 6. Sadržaji s portala BrainPOP u ISE scenariju učenja

VII. ZAKLJUČAK

U nastavi koja primjenjuje scenarij, učenici razvijaju strategiju učenja i samonadgledanja učinkovitosti naučenog te time stječu vještine samostalnog upravljanja svojim učenjem. Tijekom aktivnosti i zadataka istražuju, promišljaju, procjenjuju svoje znanje i vlastiti angažman, uspoređuju rezultate rada s ostalim učenicima.

Nastavnik prati rad učenika i pomaže kod rješavanja problema. On je facilitator, moderator, mentor, promatrač i davatelj kritičkog osvrta. Prati proces rada, po potrebi ga usmjerava i potiče. Aktivno uključuje sve učenike u proces učenja, postavlja jasna i dogovorena pravila za rad te ih dosljedno provodi. Potiče i razvija kod učenika više razine mišljenja i kritički odnos prema vlastitu znanju i znanju drugih. Nastavnik osigurava razredno ozračje u kojem će se uvažavati ideje i stavovi svih učenika

Kroz digitalni scenarij učenja nastavnici razvijaju tehničke, organizacijske, pedagoške i digitalne kompetencije. Pažljivo razmatraju interese i potrebe učenika u oblikovanju nastavnih aktivnosti. Učenici rezimiraju, komuniciraju, planiraju svoj rad, promišljaju o njemu, pružaju i primaju kritike. Komentiraju što su naučili, što bi dalje htjeli istražiti. Učenički rad može se koristiti za otvorene povratne informacije i razmišljanja.

Ključ učinkovitog scenarija je usredotočiti se na ishode učenja, a ne na tehnologiju. Potrebno je razmisliti što želimo da učenici nauče, a zatim pronaći resurse koji će nam to omogućiti.

Nastavniku se pruža mogućnost kreiranja raznih metodičkih scenarija kojima se postiže nova kvaliteta nastave, povećan interes i motivacija. Učenici se od pasivnih promatrača pretvaraju u aktivne i zainteresirane sudionike koji sudjeluju u pripremi, ostvarivanju i vrednovanju nastavnih situacija.

Učenici su aktivni sudionici nastavnog procesa koji uz vodstvo nastavnika imaju priliku samostalno istraživati i surađivati s drugim učenicima. Metode rada primjenjene u scenariju učenja, primjena programa/alata koji učenicima omogućuju razne aktivnosti na nastavi pružaju ležerniju i zanimljiviju nastavu, opće poboljšanje, veće sudjelovanje i motiviranost učenika, bolju komunikaciju učenika i radnu atmosferu.

LITERATURA

[1] Scenario-Based Learning in the Online Classroom, dostupno na http://www.facultyfocus.com/articles/online-education/scenario-based-learning-in-the-online-classroom/

[2] Scenario-Based e-Learning to Improve Learners' Engagement, dostupno na http://elearningindustry.com/the-basics-of-scenario-based-e-learning

[3] How to Build an Effective an Engaging Scenario, dostupno na http://info.shiftelearning.com/blog/a-5-step-plan-to-create-your-own-scenario-based-elearning-course

[4] Build Branched E-Learning Scenarios in Three Simple Steps, dostupno na http://blogs.articulate.com/rapid-elearning/build-branched-e-learning-scenarios-in-three-simple-steps/

[5] How I Built that E-Learning Scenario, dostupno na http://blogs.articulate.com/rapid-elearning/how-i-built-that-e-learning-scenario/?_ga=1.191153578.1600776400.1452951216

[6] E-Learning Heroes, dostupno na https://community.articulate.com/downloads/templates/powerpoint?utm_source=REL&utm_medium=Email&utm_campaign=RELFEED

[7] SBL Interactive, dostupno na http://www.sblinteractive.org/

[8] Edukata, The Innovative Learning Activity Design Toolkit, dostupno na http://itec.aalto.fi/2013/09/edukata-the-innovative-learning-activity-design-toolkit/

[9] Inspiring Science Education, dostupno na http://www.carnet.hr/ise

[10] ISE MOOC, dostupno na http://my.proficiens.eu/

[11] ISE portal, dostupno na http://portal.opendiscoveryspace.eu/ise

[12] BrainPOP, dostupno na https://www.brainpop.com/

[13] PhET Interactive simulations, dostupno na https://phet.colorado.edu/

Detekcija najčešćih sintaktičkih i logičkih grešaka učenika kod stvaranja programa u početnim godinama učenja programiranja

K. Blažeka, Tehnička škola Čakovec, Čakovec, Hrvatska
kristinka.blazeka@skole.hr

Sažetak - Ovaj se članak referira na istraživanje najčešćih grešaka koje prilikom usvajanja vještine programiranja rješavajući jednostavne zadatke manifestiraju učenici početnici u programiranju. U tu se svrhu analiziraju podaci dobiveni prevođenjem (*compilation log data*) programskih uradaka programera početnika istražujući koje su najveće poteškoće učenika početnika u programiranju kod usvajanja uvodnih, osnovnih koncepata. Zato se istražuju prikazane poruke o greškama različitih vrsta i zaključci izvode temeljem uočenih sličnosti i razlika u poteškoćama na koje učenici početnici nailaze, kao i učinkovitost učenika u postupku otklanjanja grešaka. Uočava se koje programske greške početnici rade najčešće. Pokazuje se kako početnici nailaze na slične probleme koje rješavaju sukladno razvijenosti vlastitih kognitivnih sposobnosti viših razina (sposobnost analize, prosuđivanja, stvaranja) što izravno utječe na brzinu i kvalitetu usvajanja same vještine programiranja. Također, rad sadrži i osvrt na pojavnost čestih logičkih grešaka učenika početnika analizom njihovih rješenja za najjednostavnije zadatake poslanih na online evaluator u okviru Hrvatskog otvorenog natjecanja iz informatike (HONI) sa preporukama kako učenicima odgovarajućom promjenom načina poučavanja pomagati da takve greške uspješnije izbjegavaju. Rezultati vode do korisnih preporuka za poboljšanje poučavanja i učenja programiranja, od koncepcije nastavnih sadržaja, odgovarajućeg priručnika do sugestije za stvaranja alata koji bi bili usmjereni pomaganju u izbjegavanju, odnosno otklanjanju najčešćih grešaka.

I. UVOD

Informacijsko doba obilježeno je nesmanjenom potražnjom za informatičkim stručnjacima. Među brojnim vještinama za koje se očekuje da tijekom školovanja u ovom području usvoje suvremeni stručnjaci u području informatike i računalstva, vještina programiranja zauzima značajno mjesto. Kako potražnja za dobrim programerima i dalje raste, pokazuje se kako je sposobnost programiranja računala vrlo korisna i zahvalna za stvaranje uspješne poslovne karijere.

No, učenje i usvajanje ovih vještina nije jednostavno. Kako je poznato da poučavanje vještine programiranja obilježava niz izazova, za učenike i studente, ali i za njihove nastavnik, razumijevanje i ovladavanje najvećim izazovima s kojima se suočavaju početnici važno je za kvalitetu postignutih ishoda. U nizu razloga za stvarne poteškoće na koje nailaze početnici kod učenja pisanja programa za računala mogu se naglasiti:

- često slabo strukturirani sadržaji koje početnici moraju savladati, koji su im zato teški, a postaju i povodom za odustajanje,
- općenito nerazumijevanje procesa rješavanja problema,
- nerazumijevanje samih programskih koncepata (oblikovanje i tijek izvođenja programa, oblikovanje i uloga funkcija i njihovih argumenata i dr.),
- nerazumijevanje semantike i sintakse programskih jezika koje uče,
- nerazumijevanje procesa implementacije rješenja problema u računalni program.

Ni vremenska zahtjevnost procesa usvajanja vještine uspješnog programiranja računala nije zanemariva. Utvrđeno je kako je učenje programiranja, osim što je teško, vremenski dugotrajan proces. U svom radu Winslow [E] spominje kako je početniku potrebno oko deset godina da postane stručnjak u programiranju. Istraživanjem je također utvrđeno da čak i nakon dvije godine učenja programiranja, većina programera početnika još uvijek ima primjetne poteškoće.

Naime, kako bi uspješno pisali programe za računala, od učenika se zahtjeva stručnost u svim fazama postupka rješavanja problema, u upotrebi integrirane razvojne okoline, te u upotrebi programskog jezika i prevoditelja. Zato učenici, osim što trebaju dobro ovladati detaljima sintakse i semantike korištenog programskog jezika, trebaju usvojiti niz mentalnih modela koji upućuju na to kako riješiti određeni problem te steći sposobnosti rješavanja problema kako bi došli do specifikacije samog rješenja. To su samo neki od zahtjeva koji čine učenje programiranja teškim i kognitivno vrlo izazovnim zadatkom.

Iako im sama integrirana razvojna okolina (IDE) u kojoj razvijaju svoja programska rješenja pomoću odgovarajuće programske potpore pomaže u uočavanju, prezentiranju i otklanjanju pogrešaka, učenici početnici često nisu u mogućnosti ispravno intrepretirati poruku o pogrešci koju im ispisuje prevoditelj kad pri provjeri sintaksne i semantičke ispravnosti njihovog izvornog koda ispiše poruke o pogreškama zbog nepoštivanja sintakse i semantike određenog programskog jezika.

Važnost koju briga za ispravnost programskog rješenja ima za računalne znanstvenike izražena je u opisu sadržanom u izvješću [12]: ‘‘Računalni znanstvenik brine se za robusnost, intuitivnost, održivost i iznad svega ispravnost računalnih rješenja za poslovne, znanstvene i inženjerski probleme.“ Zato je vrlo važno učenike izložiti ispravnim postupcima provjere, traženja, uočavanja i ispravljanja pogrešaka. Ispravan program (ili algoritam) definira se kao program koji stvara očekivani izlaz (*output*) – u značenju: sav očekivani izlaz i ništa osim tog očekivanog izlaza – za svaki valjani ulaz (*input*) [14].

Analizom pogrešaka početnika u programiranju s obzirom na vrstu i učestalost njihovih pogrešaka, moguće je identificirati najproblematičnija područja na koja treba obratiti pažnju pri njihovom učenju i poučavanju te odgovarajućim pristupima pomoći pri unapređivanju poučavanja, ali i utjecati na povećanje kvalitete kreiranih i usvojenih strategija rješavanja problema koje početnici koriste. Također se može više vremena i pažnje posvetiti temama i područjima sa kojima učenici imaju najviše problema. Ovo malo istraživanje bi trebalo pomoći identificirati uobičajene pogreške koje rade početnici, te njihovom analizom prepoznati problematična područja na koja bi se trebao fokusirati kurikulum. Dakle, ovaj rad predstavlja pokušaj traženja odgovora na pitanje: “Koje su i koliko česte uobičajene vrste programskih pogrešaka (sintaksnih i logičkih) koje rade učenici u prvim godinama učenja programiranja.“

Teoretski okvir na kojem istraživanje počiva je konstruktivizam prema kojem je (a) učenje proces konstrukcije znanja, (b) učenje ovisno o znanju, odnosno ljudi koriste postojeće znanje kako bi konstruirali novo, i (c) učenje je visoko usklađeno sa situacijom u kojoj se događa. Ovdje minimalistički pristup [4] podupire prirodne strategije učenika [1], a jedna od odrednica koje ga obilježavaju je snažna potpora prepoznavanju i popravljanju pogrešaka.

II. PRIJAŠNJA ISTRAŽIVANJA

Neki od najutjecajnijih izazova na učenje programiranja su oblikovanje programa jezičnim konstruktima programskog jezika (sintaksno i semantički), te oblikovanje problema kroz identificiranje i strukturiranje potrebnih programskih rješenja za određeni problem.

Prijašnja istraživanja pokazala su da početnicima očekivano nedostaju znanja programskih strategija i da imaju poteškoća u upotrebi struktura izbora, zatim da imaju poteškoće u izražavanju rješenja oblikovanih govornim jezicima kompjutorskim programskim jezicima, kao i da im izazov predstavlja analiza i dizajniranje matematičkih izraza, imenovanje varijabli i pridruživanje odgovarajućih tipova podataka i struktura tim varijablama, te ispravna evaluacija izlaznih naredbi, aritmetičkih i relacijskih izraza.

Učinkovitije okoline učenja moguće je stvarati boljim shvaćanjem procesa učenja prvog programskog jezika, odnosno načina učenja učenika početnika,. Analizom uočenih poteškoća i pogrešaka moguće je unaprijediti pristupe poučavanja i vremensku alokaciju tema [13].

Nadalje, istraživanje temeljeno na međunarodnom ispitivanju više od 500 učenika i nastavnika [9], evaluiralo je najveće poteškoće pri učenju programiranja: (a) razumijevanje kako dizajnirati program da bi rješavao određeni zadatak; (b) razumijevanje kako dijeliti punu funkcionalnost u procedure; (c) razumijevanje kako nalaziti pogreške (bugove) u vlastitim programima.

U prijašnjim je istraživanjima zabilježeno također kako početnici imaju poteškoće u uočavanju i ispravljanju pogrešaka (*debugging/fixing*) kod uvjeta u petljama, kod logičkog funkcioniranja uvjetnih naredbi. Česte su i aritmetičke greške, te greške kod ažuriranja i inicijalizacije podataka. Takva situacija ima za posljedicu da se umjesto fokusiranja na rješavanje problema, ti učenici uglavnom bave često dugotrajnim i iscrpljujućim pokušajima ispravljanja pogrešaka koje im je pokazao prevoditelj. Postoje i rješenja koja pri tom pomažu učenicima, primjer je upotreba programa koji preprocesira izvorni kod ('Gauntlet') [7] i na jednostavan način detaljno učenicima objašnjava svaku identificiranu sintaksnu pogrešku sa sugestijama o tipičnim uzrocima takve pogreške i pomaže im dajući prijedloge kako pogrešku ispraviti. Na taj se način smanjuje i količina potrebnih intervencija od strane nastavnika. Pokazuje se da učenici često imaju dobro razumijevanje ciljeva učenja, ali da su prvenstveno nedostatak iskustva u programiranju i jednostavne sintaksne pogreške odgovorni za neuspjehe učenika u programiranju. [10]

I u istraživanju [11] praćene su, bilježene i kategorizirane (prema određenim programskim konceptima) pogreške (sintaktičke i semantičke) koje su učenici radili prilikom rješavanja praktičnih zadataka iz programiranja. Kategorizirane su prema: varijable i tipovi, aritmetičke operacije, logičke operacije, ulazno/izlazne operacije, uvjetne upravljačke strukture, ponavljajuće upravljačke strukture, te modularizacija. Utvrđeno je da su glavni uzroci pogrešaka učenika slaba sposobnost rješavanja problema i nedostatak znanja i razumijevanja programskog jezika.

Važan alat za lociranje izvora pogreške u izvornom kodu programa, a koji uz to pomaže i razumijevanju programa, je *debugger*. Rezultati istraživanja [5] upotrebe *debuggera* pokazali su da: (a) formalno osposobljavanje u traženju pogrešaka pomaže studentima razviti vještine dijagnosticiranja i uklanjanja neispravnosti iz računalnih programa i (b) učenici koji su obavili dobrovoljne vježbe u traženju pogrešaka provodili su značajno manje vremena na traženje pogrešaka u svojim programima.

III. PISANJE ISPRAVNIH PROGRAMA

Odgovarajućim učinkovitim poučavanjem moglo bi se pomoći učenicima početnicima da smanje, odnosno čak da ni ne rade greške prilikom programiranja. Time bi se smanjilo vrijeme koje oni troše na ispravljanje pogrešaka, a ojačalo njihovu uspješnost učenja učinkovitih programskih koncepata. Razvijena svijest o važnosti sposobnosti pisanja ispravnih programa od esencijalnog je značaja za uspješnog programera.

Prilikom učenja učenik početnik često dobija niz pogrešaka koje nije sposoban otkloniti. Postupak otklanjanja pogrešaka dodatno je otežan učenicima kojima engleski jezik nije materinji. Naime, programski konstrukti, semantički koncepti i sintaktički elementi bazirani su na engleskom, zbog čega neki korisnici pokazuju ozbiljne

kognitivne, semantičke i sintaktičke poteškoće u razumijevanju i prevođenju svojih ideja u sintaksu i semantiku u na engleskom baziranu paradigmu programskog jezika.

Uz to, iako se povratna veza smatra jednim od najvažnijih utjecaja na učenje i motivaciju učenika, uobičajena povratna veza prevoditelja dizajnirana je za stručnjake što početnicima otežava interpretaciju i razumijevanje prezentiranih upozorenja i poruka.

U postupku stvaranja ispravnih programa važno je učenike naučiti i kako verificirati ispravnost programa postupkom testiranja, ispitivanja, odnosno provođenjem postupka koji se sastoji od identifikacije ulaznih primjera koji obuhvaćaju čitav prostor valjanih ulaznih podataka i ispitivanjem tih primjera, odnosno verifikacijom da li izvođenje programa proizvodi očekivani izlaz za sve te ulazne slučajeve.

Ovaj postupak od učenika traži da analiziraju ponašanje danog algoritma i/ili problema koji je u pozadini algoritma, odnosno da nađu pogrešku u nizu naredbi (*debugging*) ili da pripreme iscrpan skup ulaznih scenarija (*testing*) i tako razviju odgovarajuće vještine. Svrha ovakvih zadataka analize je potpora razvoju vještina važnih za dizajn algoritama s obzirom da programeri trebaju biti sposobni analizirati ispravnost, učinkovitost i robustnost njihovog vlastitog koda [2].

Besprijekorne stvari rijetko se izvedu odmah, pa je potrebno uočiti da li je stvoreni algoritam ispravan, odnosno prepoznati i ispraviti pogreške ako ih ima. Programska okolina učinit će ovu aktivnost učinkovitijom, no ovo ispitivanje ipak zahtjeva dosta razmišljanja, a nakon otkrivanja pogreške, potrebne su i analitičke sposobnosti za njihovo uspješno otklanjanje.

IV. PRIMIJENJENA METODOLOGIJA

Ovaj rad obuhvaća rezultate i razmatranja vezana uz dva ispitivanja manjeg opsega.

Prvo ispitivanje (A) vođeno je za vrijeme izvođenja praktične nastave (u računalnom laboratoriju) nastavnih predmeta iz područja računalstva u srednjoj strukovnoj školi (Algoritmi i programiranje, Programiranje i programski jezici i Računalstvo). Ključni cilj bio je utvrđivanje učestalosti pojavnosti najčešćih karakterističnih pogrešaka (sintaksnog ili semantičkog tipa) koje manifestiraju početnici u programiranju.

Drugo ispitivanje (B) odnosi se na analizu neispravnih rješenja najjednostavnijih zadataka koja su na online evaluator poslana u nekom od prvih pet kola ovosezonskog natjecanja u programiranju pod nazivom Hrvatsko otvoreno natjecanje u informatici (HONI 2015./2016.). Ovdje je ključni cilj bio utvrđivanje učestalosti pojavnosti najčešćih karakterističnih pogrešaka (logičkog tipa) koje manifestiraju početnici u programiranju. To su pogreške u sintaksno ispravnim programima, no koje treba smatrati neispravnima jer ne proizvode ispravan izlaz za određeni podskup ulaznih podataka.

Utvrđene činjenice i informacije prezentirane su sa ciljem primjene radi ostvarenja veće učinkovitosti i boljih i kvalitetnijih rezultata u procesu učenja programiranja.

A. Sudionici

Sudionici prvog istraživanja (A) bili su učenici jedne srednje četverogodišnje tehničke škole tijekom prvih godina učenja programiranja u programskom jeziku C, odnosno C++. Analizirana su programska rješenja 24 učenika prvih, drugih i četvrtih razreda u čijim su tekstovima detektirane pogreške. Za pisanje programa učenici su koristili neku od posljednjih inačica programa Microsoft Visual Studio Express for Desktop, odnosno njegovu integriranu razvojnu okolinu ili Dev C++ razvojno sučelje.

Programiranje u programskom jeziku C učili su od jedne do tri godine, odnosno u prosjeku 1,3 godine.

Sudionici drugog istraživanja (B) su učenici srednjih škola u Republici Hrvatskoj (od prvog do četvrtog razreda) – natjecatelji, odnosno članovi prijavljenih timova, nekog od prvih pet kola Hrvatskog otvorenog natjecanja u informatici. Analizirana su rješenja 39 učenika. Učenici su svoje programe pisali najvećim dijelom u programskom jeziku C, odnosno C++, a nekoliko je rješenja bilo u drugim programskim jezicima (Python i Pascal).

B. Prikupljanje podataka

Prikupljanje podataka istraživanja (A) provedeno je prilikom rješavanja nekih najjednostavnijih zadataka iz arhive zadataka natjecanja (HONI [15]). Mala složenost zadataka omogućavala je tek osnovnu provjeru znanja i vještina učenika temeljnih koncepata programiranja (deklaracija, inicijalizacija i pridruživanje vrijednosti, strukture grananja i ponavljanja) u trajanju od jednog sata. Izvorni kod svakog učenika analiziran je u razvojnoj okolini kako bi se identificirale programske pogreške (kršenje pravila sintakse i semantike) te njihova pojavnost i učestalost.

Određena sintaksna pogreška (*syntax error*) ukazuje na nemogućnost prevođenja pripadajućeg izvornog programskog koda. Prevoditelj (dio IDE, softver zadužen za provjeru pogrešaka u kodu i njegove usklađenosti sa semantikom programskog jezika) prikazuje pogrešku i njenu približnu lokaciju. Zapisane su pronađena pogreška i njena kategorija.

Prikupljanje podataka istraživanja (B) provedeno je analizom rješenja nekih najjednostavnijih zadataka iz arhive zadataka natjecanja (HONI 2015./2016.). Svoja su rješenja učenici dostavljali na *online* evaluator [6]. Izvorni kod svakog učenika vrednovan je s obzirom na pripremljene test podatke i analiziran kako bi se identificirale najčešće logičke programske pogreške.

Logičke pogreške ukazuju na krivu ili neispravnu programsku logiku koja vodi do neočekivanih izlaznih rezultata nakon što je program preveden i izveden. Primjećuje se tada kada je izvorni kod izveden, ali stvarni izlazni rezultati različiti su od željenih i očekivanih. Primjećene logičke pogreške također su zapisane.

Prijašnja istraživanja pokazala su da je logičke pogreške teže otkrivati i popravljati od sintaksnih. Logičke pogreške uglavnom ukazuju na upotrebu neispravnog algoritma u programu [3] i zahtjevaju mentalnu usmjerenost, logičko razmišljanje i strategiju pronalaženja i ispravljanja jer prevoditelj obično ne ukazuje na to da se logička pogreška

pojavila, samo je poznato da se izlazni rezultati razlikuju od očekivanih. Logičke pogreške zahtjevaju bolje strategije poučavanja i učenja, priručnici tu obično nisu dovoljni.

V. REZULTATI I RASPRAVA

A. Ispitivanje A

Analizirane su poruke prevoditelja i detektirane pogreške podijeljene su u sedam (7) kategorija (Tablica 1.). Ukupni broj zabilježenih pogrešaka je 33.

TABLICA I. RASPODJELA UOČENIH SINTAKTIČKIH I SEMANTIČKIH POGREŠAKA U ISPITIVANJU (A)

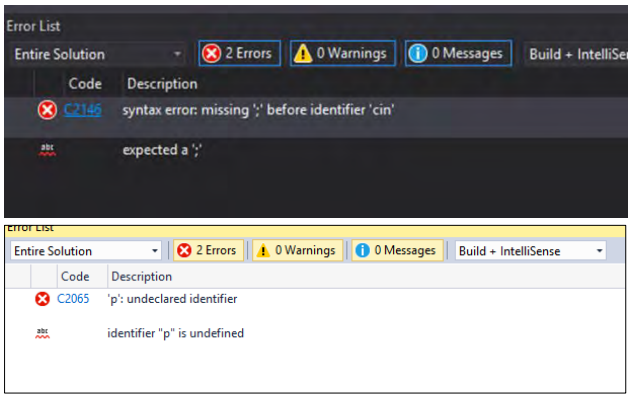
R	C	EI	ET	ER	EUI	EIE	EUN	ETS
1	0	1						
1	0		1					
1	0		1					
1	0		1					
1	0	1						
1	0	1						
1	0	1						
1	0			1				
1	0	1	1					
2	1				1	1		
2	1						1	
2	2		1					1
2	2		1			1		
2	1		1					
2	1		1					
2	1		1					
4	3		1					1
4	3	1	1					
4	3		1					
4	3		1					
4	3						1	
4	3		1		1			1
4	3		1		1			
4	2		1					
	1,3	6	16	1	3	2	2	3

Objašnjenje oznaka stupaca:
R – razred sudionika
C – godine učenja programskog jezika C/C++
EI – pogreška u pisanju naziva identifikatora
ET – pogreška u pisanju izraza (operatora) ili izostanak graničnog znaka naredbe (;) ili nekog drugog simbola (token)
ER – pogreška redefiniranja identifikatora varijable
EUI – pogreška nedeklariranog identifikatora
EIE – pogreška u pisanju uvjetne if-else() naredbe
EUN – pogreška izostanka inicijalizacije
ETS – pogreška krivog specifikatora tipa

Najčešći uočeni tip pogreške bio je izostanak nekog od *punktuatora* (! % ^ & * ( ) – + = { } | ~ [ ] \ ; ' : " < > ? , . / #). Ova se pogreška u tekstovima programa pojavila čak 16 puta što čini gotovo polovicu (48,5%) svih zabilježenih pogrešaka.

Slijedeća pogreška po učestalosti bilo je krivo pisanje identifikatora, a u programima učenika pojavila se 6 puta.

Učestalost svake od ostalih pet vrsta pogrešaka manja je od 10%.



Slika 1. Primjeri obavijesti o sintaktičkim pogreškama u programima

Tako je npr. pogreška u pisanju if-else() naredbe prijavljena dva puta, a tri puta su učenici zaboravili prvo deklarirati upotrebljeni identifikator. Izostanak inicijalizacije korištene varijable također je prijavljen dva puta.

Ovi rezultati navode na zaključak da nije toliko nepoznavanje naredbi i načina njihovog pisanja uzrokom pogrešaka, koliko to predstavlja pridržavanje strogoj sintaksi i semantici pisanja programa, jer potrebne simbole učenici najčešće izostavljaju zbog nepažnje.

B. Ispitivanje B

Ukupni broj zabilježenih pogrešaka je 39.

U tablici je navedena raspodjela uočenih logičkih pogrešaka u promatranim rješenjima.

TABLICA II. RASPODJELA UOČENIH LOGIČKIH POGREŠAKA U ISPITIVANJU (B)

Naziv zadatka	Neispravan izraz ili formula	Neispravna provjera granica	Previd u rubnom uvjetu petlje ili grananja	Neispravna veličina polja	Gramatička pogreška	Nekorektno logičko zaključivanje (If-else)	Manjkajući raniji prekid petlje
IZLET	1		6				
NOMIS		3	1				
TNT	1						
KARTE	1			3			
KOCKICE	7		2		5	2	
NOGOMET	1		1				2
DOMACICA	1				1		1
Ukupno: 39	12	3	10	3	6	2	3

Ovi rezultati pokazuju kako su u tekstovima izvornih kodova promatranih rješenja učenici najčešće griješili pri oblikovanju izraza, bilo da su izveli i/ili primjenjivali krivi izraz, odnosno formulu, bilo da su upotrebljavali krivu varijablu. Zbog takvih pogrešaka program im je bio neispravan u 12 uzoraka, odnosno gotovo trećini slučajeva (31%). Razlog ovoj pogrešci često je i nepažnja kod korištenja *copy-paste* postupka.

Druga pogreška po učestalosti koja je bila razlogom neispravnosti programa bili su rubni uvjeti naredbe ponavljanja ili grananja. Ta se pogreška pojavila u otprilike četvrtini rješenja (25,6%).

Prilično visoko, na trećem mjestu učestalosti, javila se i gramatička pogreška, tipično vezana uz potrebu da se kao izlazni rezultat ispiše neki tekst. Njena je pojavnost u obliku neispravnog ispisa traženog teksta zabilježena u 6 slučajeva (oko 15%).

Ostale pogreške pojavile su se u manje od 10% slučajeva. U toj su skupini zabilježeni: neispravna provjera parnosti, neispravna (premala) veličina polja, nekorektno logičko zaključivanje, odnosno njegova pretvorba u if-else() naredbu grananja, te izostavljeni raniji prekid petlje.

```
int main () {
    scanf ("%d %d", &m, &s);

    if (s == 1) s = 6;
    else if (s == 2) s = 5;
    else if (s == 3) s = 4;

    if (m == 1) m = 6;
    else if (m == 2) m = 5;
    else if (m == 3) m = 5;

    if (m > s) printf ("mirko");
    else if (m < s) printf ("slavko");
    else printf ("nerijeseno");

    return 0;
}

void func(){
    if (7 - a > a)
        a = 7 - a;
    if (7 - a > a)
        a = 7 - a;
    if (a > b)printf("mirko");
    if (a == b)printf("nerijeseno");
    if (a < b)printf("slavko");
}
```

Slika 2. Primjeri logičkih pogrešaka u programima

ZAKLJUČAK

Jedan od ciljeva ovog istraživanja bio je uvid u najčešće pogreške koje učenici početnici u programiranju rade tijekom prvih godina učenja programskog jezika, njihova detekcija i kategorizacija.

Učinkovite tehnike koje bi pomogle u jednostavnijoj identifikaciji, ispravljanju i smanjivanju najčešćih, odnosno uobičajenih pogrešaka pri pisanju računalnih programa omogućile bi učenicima početnicima da se više usmjere na rješavanje problema. Ti bi sadržaji mogli biti obuhvaćeni priručnikom koji bi opisivao uobičajene poteškoće i pogreške prilikom pisanja programa i tako pomogao u obrazovanju učenika početnika da smanje, pa čak i spriječe pojavu različitih programskih pogrešaka [A]. Takav bi referentni materijal sadržavao popis pogrešaka sa detaljnim primjerima i uputama kako ih izbjeći čime bi se ostvario pozitivan utjecaj na smanjivanje broja pogrešaka u programiranju i time poboljšale vještine programiranja učenika.

Poželjno bi bilo u budućnosti provesti i više istraživanja usmjerenih razvoju nastavnih strategija i strategija učenja kojima bi se pomoglo učenicima i studentima početnicima da prevladaju poteškoće u učenju koje za posljedicu imaju pojavu pogreški logičkog tipa.

Također bi se u *plug-in* obliku moglo napraviti rješenje koje bi služilo za detektiranje pogrešaka i predlaganje mogućih rješenja za određeni, česti tip pogreške.

Vrlo je važno i da nastavnici računalstva mlade generacije educiraju kako je svaka pogreška, ma kako mala ona bila, pokazatelj da je program neispravan i da zbog toga može izazvati ozbiljne troškove, pa i posljedice za ljudske živote [8].

LITERATURA

[1] Boyle11, T., **Design for Multimedia Learning**, Prentice Hall, 1997.

[2] Burton A., B. **Encouraging Algorithmic Thinking Without a Computer** Olympiads in Informatics, 2010, Vol. 4, 3–14 3 © 2010 Institute of Mathematics and Informatics, Vilnius

[3] Özmen Büşra, Altun Arif: **Undergraduate Students' Experiences in Programming: Difficulties and Obstacles**, Turkish Online Journal of Qualitative Inquiry, July 2014, 5(3)

[4] Carroll, J. M., **The Nuremberg funnel: designing Minimalist Instruction for practical computer skill**, MIT Press, 1990, in Sneiderman, B. **Designing the User Interface: Strategies for Effective Human-Computer Interaction,** Addison Wesley Publishing Company, 1992.

[5] R. Chmiel & M.C.Loui: "**Debugging: from Novice to Expert**", Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education, SIGCSE 2004, Norfolk, Virginia, USA, March, 2004.

[6] http://evaluator.hsin.hr.

[7] T.C.A Flowers, C. Carver, and J. Jackson. 2004. "**Empowering Students and Building Confidence in Novice Programmers through Gauntlet.**" Savannah, GA 34Th ASEE/IEEE Frontiers in Education Conference T3H-10 20 – 23.

[8] Y. Ben-David Kolikant* and M. Mussai: **''So my program doesn't run!'' Definition, origins, and practical expressions of students' (mis)conceptions of correctness**, Computer Science Education, Vol. 18, No. 2, June 2008, 135–151

[9] E. Lahtinen, K. Ala-Mutka, H.M. Jarvinen: **"A study of the difficulties of novice programmers "**, ACM SIGCSE Bulletin, New York, NY, USA, ACM Press, 3 : Vol.37. pp. 14-18, 2005.

[10] Murtala Mohammed, Abdulwahab Lawan, Bashir S. Galadanci, Mijinyawa M.K: **Detection and Categorization of Errors by Novice Programmers in a First Year Java Programming Class: A Comparative Analysis,** Proceedings of the The Third International Conference on Digital Enterprise and Information Systems, Shenzhen, China, 2015

[11] Nelishia Pillay, Vikash R. Jugoo: **An Analysis of the Errors Made by Novice Programmers in a First Course in Procedural Programming in Java**, Dec 28, 2013

[12] Tucker, A.: **A model curriculum for K–12 computer science**. Final report of the ACM K–12 task force curriculum committee, 2003. Dohvaćeno u siječnju, 2016., sa http://www1.acm.org/education/k12/k12final1022.pdf, p. 12.

[13] Ioana Tuugalei Chan Mow: **Analyses of Student Programming Errors In Java Programming Courses**, Journal of Emerging Trends in Computing and Information Sciences, Vol. 3, No. 5, May 2012

[14] GinatJ, D. (1998).**Foundations of computer science** I. Rehovot, Israel: The Weizmann Institute of Science.

[15] www.hsin.hr/honi (dohvaćeno veljača, 2016.

Nastava matematike na SageMathCloud platformi

Željka Tutek
Geodetski fakultet, Zagreb
zeljkat@geof.hr

Sažetak - Jedan od temelja inženjerskog znanja je matematika. Matematički softver omogućuje prijelaz od malog broja zadataka koji se mogu riješiti analitičkim metodama na daleko veći broj složenijih problema čije rješavanje uključuje i numeričke metode. Za razliku od MATLAB, Mathematica i Maple, SageMath je slobodan softver otvorenog koda koji koristi poznate programske biblioteke: NumPy, SciPy, matplotlib, SymPy, Maxima, GAP, FLINT, R,… Iza mojih studenata i mene je iskustvo rada sa SageMath na odnedavno dostupnoj SageMathCloud platformi kao alatu za nastavu; i iz te perspektive daje nekoliko odgovora na pitanje: koliko i kako uvođenje matematičkog softvera u nastavi matematike prve godine za inženjere uistinu olakšava učenje i poučavanje?

I. Uvod

Kad je 2015 godine trebalo predložiti izmjene nastavnog programa, nisam imala ni malo sumnje da se matematičke predmete prve godine studija lako može osnažiti korištenjem matematičkog softvera SageMath. Uz sve njegove poznate prednosti, jednostavnost pristupa na novoj SageMathCloud platformi je SageMath učinila izvrsnim izborom za nastavu.

II. SageMath

SageMath je slobodan matematički softver otvorenog koda licenciran kao GNU GPL. Početak njegovog razvoja 2004 godine je bio vođen idejom matematičara Williama Steina da se omogući jedinstven pristup za korištenje velikog broja postojećih slobodnih specijaliziranih programskih biblioteka i alata napisanih u raznim programskim jezicima. S misijom „creating a viable free open source alternative to Magma, Maple, Mathematica and Matlab“ [1], u njegov razvoj uključeno je preko pet stotina istraživača, studenata i inženjera.

Kroz sučelje napisano u programskom jeziku Python moguć je kombiniran pristup ka stotinjak slobodnih programskim biblioteka: NumPy, SciPy, matplotlib, SymPy, Maxima, GAP, FLINT, PARI, Singular, R,… za razna područja matematike: diferencijalni i integralni račun, linearnu algebru, diskretnu matematiku, algebru, logiku, geometriju i topologiju, teoriju brojeva, algebarsku geometriju, vjerojatnost i statistiku,… [2].

Kako bi se postigla brzina i efikasnost, dijelovi SageMath koda se mogu implementirati u programskim jezicima: Cyton, C/C++, Fortran 95, GAP, Common Lisp, Singular i PARI/GP [2]. Za implementaciju se mogu koristiti i Octave, Magma, Mathematica i Maple ako su zasebno instalirani. SageMath je trenutno u inačici 7.0. Donedavno se zvao samo Sage ili SAGE što je akronim od **S**oftware for **A**lgebra and **G**eometry **E**xperimentation.



Slika 1. Logo matematičkog softvera Sage

III. SageMathCloud

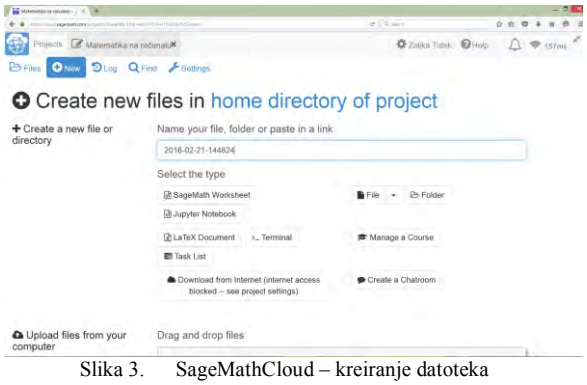
SageMathCloud (kraće: SMC) je zajedničko web sučelje za SageMath, IPython, LaTeX i Terminal kojem se slobodno može pristupiti s korisničkim računom na web adresi https://cloud.sagemath.org/. Dio je SageMath projekta koji su potpomogli Google [4] i US National Science Foundation. Još uvijek je u fazi testiranja, ali pokazuje stabilnost i sve veću funkcionalnost kroz dvije godine postojanja. Iako usluge poslužitelja od 2015 godine vodi kompanija "SageMath, Inc." [6], cjelokupni SageMathCloud kod je otvoreno dostupan u GitHub repozitoriju.

Glavne značajke SageMathCloud platforme [3] su

- alat za nastavu (lako kreiranje projekata i uključivanje studenata, zadavanje zadataka, praćenje rada studenata te preuzimanje i ocjenjivanje rješenja),
- računalna matematika (mogućnost korištenja: SageMath, IPython, Scientific Python Stack, R, Julia, GAP, Octave,…),
- jednostavna suradnja među korisnicima (više članova tima može u realnom vremenu uređivati isti dokument i koristiti video chat ili chat uz pomoć HTML/Markdown/LaTeX),
- jedinstveno programsko razvojno okruženje (pisanje, prevođenje i izvođene programa u raznim programskim jezicima),
- ugrađen LaTeX editor (jednostavna i kvalitetna izrada i najsloženijih dokumenata).



Slika 2. Značajke SageMathCloud platforme



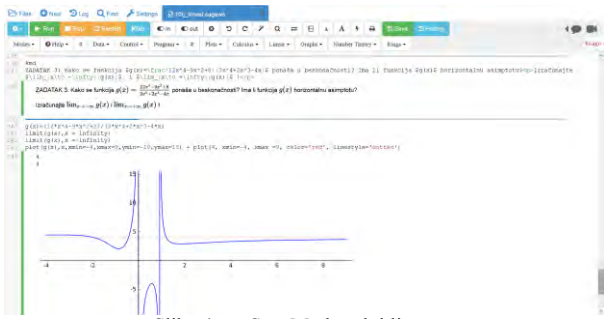
Slika 3. SageMathCloud – kreiranje datoteka

SageMathCloud omogućava rad s interaktivnim SageMath radnim listovima (engl. worksheets). To su datoteke tipa sagews koje se sastoje od niza ćelija što mogu sadržavati kod (uz SageMath i R, Python, Julia, Javascript, Ruby,…) ili komentare (HTML, Markdown, LaTeX) te izlazne rezultate dobivene samostalnim izvođenjem tih ćelija. Kao i za Jupyter bilježnice (engl. notebook) [5], za rad treba samo jedan prozor preglednika. SageMath radni listovi se mogu kreirati, otvoriti, spremiti, kopirati, preimenovati, premještati i dijeliti preko pridružene URL web poveznice. Mogu pripadati raznim projektima i direktorijima korisnika.

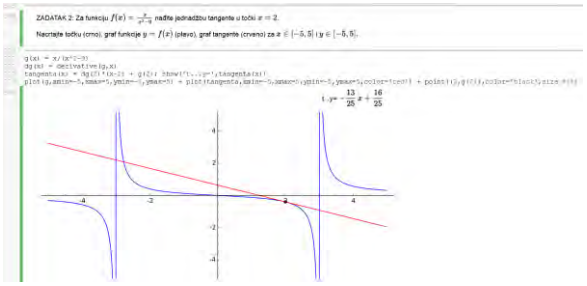
Već i nakon kratkog iskustva rada sa SageMath radnim listovima lako se prepozna što William Stein u svom blogu od 10.2.2016 [6] i piše „Open source is now ready to compete with Mathematica for use in the classroom“.

IV. NASTAVA MATEMATIKE

Najčešće se studenti prve godine tehničkih studija susreću s apstraktnim pojmovima u predmetima koji uključuju sadržaje linearne algebre i matematičke analize. I često imaju probleme u razumijevanju brojnih novih matematičkih pojmova. No bez obzira što u ovom 21. stoljeću svi prihvaćamo da je nužno uključivanje matematičkog softvera i internetskih izvora u nastavu jer je jasno da vizualizacija i numeričke metode mogu znatno olakšati učenje, realizacija često i nije moguća ili nije uspješna. Neki od problema su npr. visoke cijene programskih licenci, ionako prepune računaonice za najbrojniju prvu godinu studija, veliko vrijeme potrebno za stjecanja vještine samostalnog korištenja kvalitetnog matematičkog softvera i njegova složenost što može učiniti problem još i složenijim, mijenjanje načina vrednovanja i ocjenjivanja,…



Slika 4. SageMath radni list



Slika 5. SageMath – primjer zadatka

U ak. godini 2015/16 po prvi put sam izvela nastavu izbornog predmeta koji je imao za ciljeve

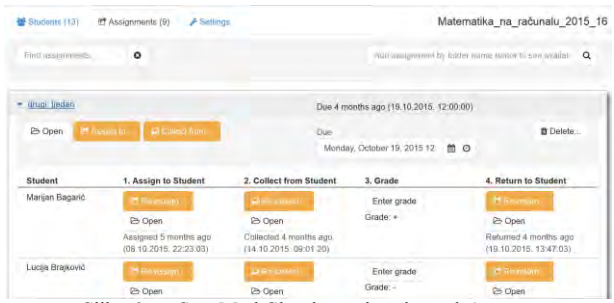
- steći vještinu samostalnog korištenja matematičkog softvera za simboličko i numeričko računanje i
- podržati nastavu matematičkih predmeta (Analitička geometrija i linearna algebra i Matematička analiza).

Nastava se sastojala od petnaest jednosatnih laboratorijskih vježbi na kojima su studenti koristili svoja prijenosna računala i Eduroam za bežično spajanje na Internet.

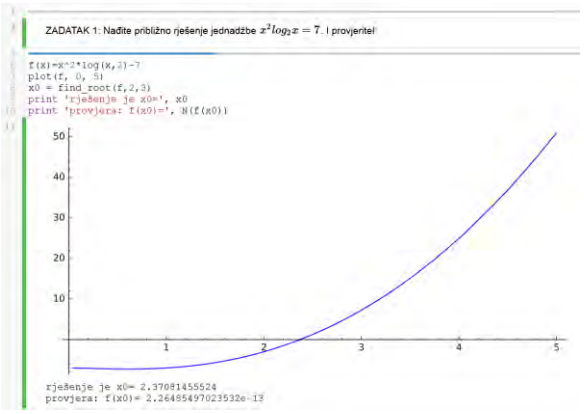
Predmet je imao svoju web stranicu u LMS Moodle-u (sustavu E-učenje Geodetskog fakulteta), ali sam koristila i alat za nastavu („Manage a Course“) u SageMathCloud. Studenti se pozivaju na tečaj (engl. course) preko e-mail adrese kojom su se prijavili. Zadavanje zadaća, praćenje aktivnosti, ocjenjivanje rješenja i vraćanje povratne informacije se pokazalo više primjereno programskom okruženju i praktičnije od onog što nudi Moodle.

SageMathCloud platforma nudi radno okruženje koje je intuitivno i ne zahtijeva instalaciju programa ili dodataka pa se već nakon prvog sata početnog upoznavanja počelo s rješavanjem matematičkih problema pomoću SageMath. Slobodan pristup platformi dopušta maksimalni diskovni prostor od 7,5 GB po projektu i maksimalnu ukupnu radnu memoriju od 1 GB za sve projekte. To se za izvođenje nastave pokazalo dovoljno.

Pripremila sam petnaest SageMath radnih listova koje su studenti samostalno prolazili i rješavali zadaće (rad s editorom i ćelijama; jednadžbe i nejednadžbe; vektori; točke, pravci i ravnine; limesi; deriviranje; integriranje; analiza svojstava funkcija; crtanje grafova funkcija u 2D i 3D; matrice; linearni sustavi; linearni operatori; svojstvene vrijednosti i svojstveni vektori). Zadaće su često tražile i grafičko i numeričko rješavanje.



Slika 6. SageMathCloud - studentske zadaće



Slika 7. Primjer rješavanja nelinearne algebarske jednažbe

Numeričke metode su najčešće tema posebnih kolegija viših godina tehničkih studija pa se u uvodnim kolegijima još uvijek koriste samo analitičke metode. Jednostavnost numeričkog rješavanja npr. nelinearnih algebarskih jednadžbi (Slika 7.) ili numeričkog integriranja u SageMath čini prihvatljivim njihovo rano uključivanje u nastavu matematike. Time se omogućuje rješavanje daleko šireg skupa problema. A i studenti dobivaju realniji uvid na doseg analitičkih metoda.

Kako sam i očekivala pokazalo se da predznanje Pythona nije bilo potrebno. Korišteno je ukupno oko sedamdeset SageMath naredbi. Tipka Tab daje popis svih naredbi koji započinju unesenim slovima. A ako se iza imena naredbe upiše znak upitnika, ispisat će se za nju ugrađena detaljna dokumentacija s mnogo primjera. Dio studenata je na kraju nastave ipak izrazio želju da samostalno nauči Python.

Svi studenti koji su pohađali nastavu su uspješno savladali definirane ishode učenja. Za očekivati je da i nadalje nastave koristiti moćni slobodni matematički softver.

V. ZAKLJUČAK

Izvođenje nastave uz pomoć SageMath na SageMathCloud platformi me uvjerilo da je SageMath nakon deset godina svog razvoja trenutno najbolji izbor matematičkog softvera koji jednostavno i uspješno može i treba biti uključen u nastavu matematičkih kolegija prve godine tehničkih studija. Vodeći komercijalnih programi slične namjene koji su svoj razvoj počeli mnogo prije (Maple 1980, Matlab 1984, Mathematica 1988), imaju zahtjevnija radna sučelja koja traže više vremena za upoznavanje i iskusnijeg korisnika.

LITERATURA

[1] SageMath – web stranica projekta, dostupna na poveznici http://www.sagemath.org/ , 2016

[2] SageMath dokumentacija, dostupna na poveznici http://doc.sagemath.org/ , 2016

[3] SageMath Cloud – web stranica za prijavu , dostupna na poveznici https://cloud.sagemath.org/ , 2016

[4] W. Evans, "The Struggle for Open Mathematics Software", Online Searcher, vol. 39 (2), str. 22-26., 2015

[5] Jupyter – web stranica projekta, dostupna na poveznici http://jupyter.org/ , 2016

[6] S. William, "Sage: Open Source Mathematics software", blog dostupan na poveznici http://sagemath.blogspot.com/, 2016

[7] SciPy – web stranica projekta, dostupna na poveznici http://www.scipy.org/about.html , 2016

[8] S. William, J. David, "SAGE: System for Algebra and Geometry Experimentation", ACM SIGSAM Bulletin, vol. 39(2), str. 61-64., 2005

[9] M. A. Gray, "Sage: A New Mathematics Software System", Computing in Science & Engineering, vol. 10(6), str. 77-75., 2008

[10] G. V. Bard, "Sage for Undergraduates", AMS, 2015

[11] A. Casamayou, N. Cohen, G. Connan, T. Dumont, L. Fousse, F. Maltey, M. Meulien, M. Mezzarobba, C. Pernet, N. M. Thiéry, P. Zimmermann, "Calcul mathématique avec Sage", CreateSpace, 2013

[12] G. A. Anastassiou, R. A. Mezei, „Numerical Analysis Using Sage", Springer, 2015

Uvod u robotiku - Arduino platforma i web aplikacija

Andrej Lacković, Bojan Fulanović
Visoko učilište Algebra, Zagreb, Hrvatska
andrej.lackovic@racunarstvo.hr, bojan.fulanovic@racunarstvo.hr

Sažetak - Arduino je jednostavna razvojna platforma zasnovana na univerzalnom mikrokontroleru. Idealna je za razvoj upravljačke elektronike i robotike kako kod učenika tako i kod studenta. Od više raznih mogućih načina upravljanja Arduino platformom jedan od mogućih je i putem mreže, odnosno putem web aplikacije. Web aplikacije su programska rješenja kojima se pristupa putem Internet preglednika na računalu ili mobilnom uređaju i danas su neminovnost koja se ne može izbjeći. Rad opisuje izradu web aplikacije za upravljanje robotskog autića zasnovanog na Arduino platformi.

I. UVOD

Mikrokontroler je digitalna elektronička naprava u obliku integriranog kruga. Mikrokontroler normalno radi u kontroliranoj petlji, dakle očitava ulaze i zatim podešava izlaze sukladno sa svojim programom. Petlja se stalno ponavlja dok traje kontrola procesa. Namjena mikrokontrolera je upravljanje uređajima i procesima, pa u sebi ima integrirani mikroprocesor, memoriju, digitalne i analogne ulaze i izlaze, komunikacijske sklopove i druge dodatke. Programi za mikrokontroler se pišu na računalu te pomoću programatora prebacuju u njihovu memoriju. Najpoznatiji proizvođači mikrokontrolera, uglavnom za hobiste, su Atmel i Microchip.

II. ARDUINO

Arduino je platforma otvorenog koda za kreiranje elektroničkih prototipova baziran na sklopovlju i programskom paketu koji je fleksibilan i jednostavan za korištenje. Namijenjen je umjetnicima, dizajnerima, hobistima, elektroničarima i svima koji su zainteresirani za kreiranje interaktivnih objekata i okruženja.



Slika 1. Arduino Uno i Genuino mikrokotroler

Arduino je skup elektroničkih i softverskih komponenti koje se mogu jednostavno povezivati u složenije cjeline s ciljem izrade zabavnih i poučnih elektroničkih sklopova. Srce Arduina jesu mikrokontroleri. Najčešće koristi 8 bitne mikrokotrolere koje proizvodi tvrtka Atmel.

A. *Verzije Arduina*

Arduino dolazi u nekoliko različitih verzija. Uglavnom se razlikuju po mikroprocesoru, količini memorije, te broju analognih i digitalnih ulaza/izlaza.

Pa tako razlikujemo:

- Arduino Uno (slika 1.),
- Arduino Mega,
- Arduino Pro,
- Arduino Zero.

B. *Programska podrška i programiranje*

Programiranje Arduino platforme radi se kroz posebno dizajnirano sučelje nazvano Arduino IDE (slika 2.) Inače sam programski jezik baziran je na AVR C programskom jeziku.

Arduino IDE ujedno služi za pisanje programa i za prenošenje samog programa na mikrokontroler. Dok kod većine drugih mikrokontrolera je potreban fizički programator, kod Arduina to nije potrebno. Na samom Arduinu nalazi se Arduino „bootloader“. Mali program koji služi za prenošenje programa na sam mikrokontroler bez nekakvog dodatnog hardvera.

C. *Moguće nadogradenje*

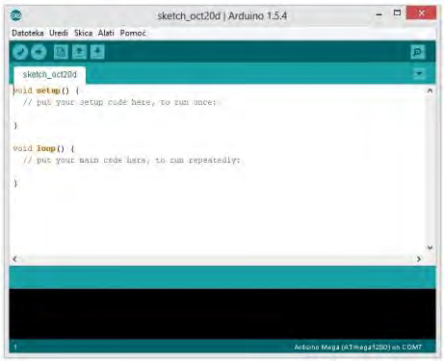
Arduino je moguće nadograditi sa njemu kompatibilnim modulima za različite potrebe. Moguće nadogradnje možemo svesti na module i dodatke.

Neki od mogućih modula:

- Ethernet Shield (mrežni modul),
- WiFi Shield (WiFi modul),
- GSM Shield (GSM modul),
- Motor Shield (kontrola motora).

Neki od mogućih dodataka:

- TFT LCD screen (LCD monitor),
- SD card adapter (dodatna memorija),
- Mini USB/Serial adapter,
- RF prijamnik i predajnik,
- Razni senzori
  - IR senzor,
  - Ultrazvučni senzor.



Slika 2. Arduino IDE

D. *Arduino biblioteke*

Programski jezici C/C++ koriste biblioteke kao osnovni način proširivanja svoje funkcionalnosti. Biblioteke imaju zaglavlje (header file) u kojem se nalaze prototipovi funkcija koje sadrže (koje može koristiti program) kao i deklaracija specijalnih tipova podataka i makroa koje koriste ove funkcije.

Arduino biblioteke omogućuje nam da kompleksne i dugačke linije koda zamijenimo vrlo jednostavnim funkcijama. Osim što će nam uvelike pojednostaviti i skratiti rad s određenim modulom ili senzorom, nemamo potrebe ulaziti u dubinu rada i problematike modula/senzora niti biti napredni poznavatelji C/C++ programskih jezika.

Gotove biblioteke moguće je preuzeti putem interneta koje korisnici pišu i dijele ga kao open-softver. Drugi način je da napišemo vlastite biblioteke. Da bi program koristio biblioteku, potrebno je učitati je u Arduino okruženje i uključiti je. Biblioteke se uključuju pomoću ključne riječi #include <nazivbiblioteke.h>

III. WEB APLIKACIJE

Web aplikacije su programska rješenja kojima se pristupa pute Internet preglednika koristeći Internet ili intranet. One su dostupne u bilo kojem trenutku s bilo kojeg mjesta pomoću računala ili mobilnih telefona. Osim toga, web aplikacije nije potrebno postavljati na svako računalo ili mobilni telefon jer im se pristupa identično kao i ostalim web stranicama s bilo kojeg web preglednika.

Web aplikacije pisane su upotrebom standarda definiranih od strane međunarodnih organizacija, što osigurava da su iste upotrebljive na svim vrstama platformi. Za samu upotrebu web aplikacija nije potrebno posebno znanje, dovoljno je poznavanje korištenja internetom. Neke od tehnologija koje se koriste pri izradi web-aplikacija:

- Apache web server,
- Ruby on Rails framework,
- MySQL i PostgreSQL baze podataka,
- HTML i CSS,
- Java Script biblioteke - jQuery, Node.js i druge.

A. *Hyper text markup language*

HTML je kratica za HyperText Markup Language, što znači prezentacijski jezik za izradu web stranica. Hipertekst dokument stvara se pomoću HTML jezika. HTML je jednostavan za uporabu i lako se uči, što je jedan od razloga njegove opće prihvaćenosti i popularnosti. Svoju raširenost zahvaljuje jednostavnosti i tome što je od početka bio zamišljen kao besplatan i tako dostupan svima. Temeljna zadaća HTML jezika jest uputiti web preglednik kako prikazati sadržaj. Pri tome se nastoji da taj dokument izgleda jednako bez obzira o kojemu je web pregledniku, računalu i operacijskom sustavu riječ. Html datoteke su zapravo obične tekstualne datoteke, ekstenzija im je .html ili .htm.

B. *Cascading style sheets*

CSS je akronim za Cascading Style Sheets i predstavlja jednostavan mehanizam za dodavanje izgleda informacijama u HTML stranici. CSS je kao tehnologija nastala iz nekoliko različitih rješenja kako bi se u osnovi odvojio izgled opisa informacije od samog opisa i same informacije. Uporaba CSS-a u stvaranju današnjih Web sadržaja je više nego poželjna, ne samo zbog svoje jednostavnosti već i zbog činjenice da se Web sadržaji danas dohvaćaju s različitih uređaja koji imaju i različita fizička ograničenja po pitanju prikazivanja informacija i njihovih opisa pa nam CSS omogućava da Web sadržaj odmah prilagodimo i tim uređajima bez dodatnog izrade novih i specifičnih HTML stranica.

CSS sintaksa se sastoji od opisa izgleda elemenata u dokumentu. Opis može da definira izgled više elemenata, i više opisa može da definira jedan element. Na taj način se opisi slažu jedan preko drugog da bi definirali konačni izgled određenog elementa (od tuda naziv Cascading (eng. Cascade – crijep) da bi se dočaralo slaganje jednog stila preko drugog u definiranju konačnog izgleda elementa)

Svaki opis se sastoji od tri elementa:

- definicija ciljnih elemenata,
- svojstva,
- vrijednosti.

C. *Java Script*

Java Script je jednostavan, interpretiran programski jezik namijenjen ponajprije razvoju interaktivnih HTML stranica. Jezgra Java Scripta uključena je u većinu današnjih preglednika (Internet Explorer, Google Chrome, Mozilla Firefox, Opera, Safari i drugi). Zbog toga su programi kraći i čitljiviji. Java Script omogućuje izvršavanje određenih radnji u inače statičnim HTML dokumentima, npr. interakciju s korisnikom, promjenu svojstava preglednikova prozora ili dinamičko stvaranje HTML-sadržaja. Java Script nije pojednostavljena inačica programskog jezika Java. Povezuje ih jedino slična sintaksa i to što se koriste za izvršavanje određenih radnji unutar preglednika.

Java Script se može pisati u bilo kojem uređivaču teksta (editor) koji podržava standard ASCII. Blok za pisanje (Notepad) prisutan je u svim inačicama operacijskog sustava Windows pa se nameće kao pogodan i za programiranje u Java Scriptu.

IV. PROJEKT ARDUINO AUTIĆ – WEB APLIKACIJA

Ideja projekta je kao uvod u robotiku napravit upravljanje Arduino autića pomoću web aplikacije. Budući je ideja upravljati autićem bežičnom putem u projektu će se koristiti dvije arduino platforme koje će međusobno komunicirati putem RF prijamnika i predajnika.

Komponente korištene u projektu:

- Dvije Arduino Uno platforme (ATmega 328),
- RF prijamnik i predajnik,
- Motor Shield,
- Računalo (Windows 8).

Shema spajanja komponenti ili tijek upravljanja je sljedeći: jedna arduino platforma je spojena na računalo i njome se upravlja putem web aplikacije kroz internet preglednik. Druga arduino platforma nalazi se na samom autiću, te putem RF prijamnika prima naredbe od prve arduino platforme. Samim time je postignuto upravljanje motorima na autiću putem web aplikacije.

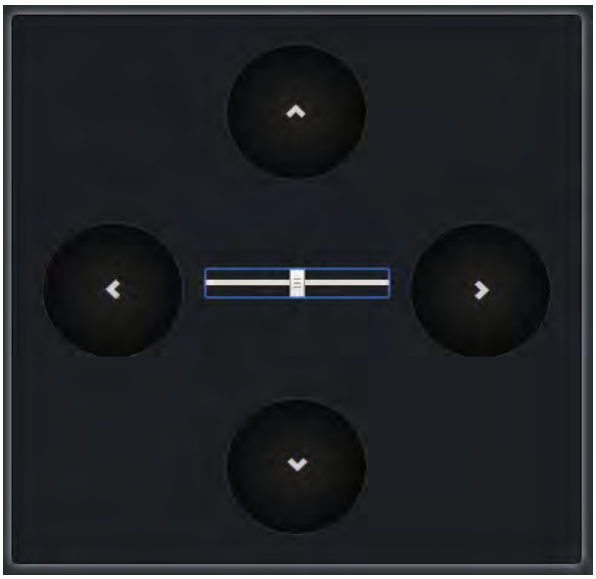
A. Web aplikacija

Za izradu same web aplikacije korišteni su HTML, CSS, te JavaScript. Ideja je da aplikacija sadrži 4 gumba koje određuju smjer gibanja autića, te klizač pomoću kojeg bi se određivala brzina gibanja autića.

Komponente koje su korištene:

- Bootstrep framework,
- jQuery Java Script Library v1.10.2,
- ardunio.js,
- ArdunoKontrole.css.

Prve dvije komponente su preuzete s interneta kao gotove open-source komponente, dok su druge dvije vlasitito napisani javascrip, te css. Izgled gotove aplikacije vidljiv je na slici 3.



Slika 3. Izgled gotove web aplikacije za upravljanje

B. HTML programski kod

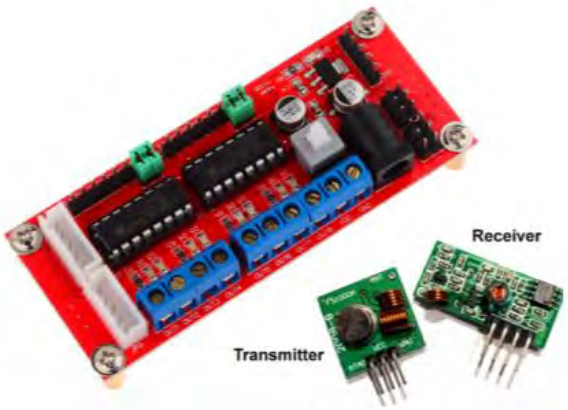
```
<!DOCTYPE html>
<html>
<head>
   <meta name="viewport" content="width=device-width"
/>
   <title>Arduino kontrole</title>
   <scriptsrc="/arduino/Scripts/jquery-1.10.2.js"></script>
   <script src="/arduino/Scripts/bootstrap.js"></script>
   <script src="/arduino/Scripts/arduino.js"></script>
   <link href="/arduino/Content/bootstrap.css"
rel="stylesheet" />
   <link href="/arduino/Content/ArduinoKontrole.css"
rel="stylesheet" />
</head>

<body>
   <div id="main">
    <div id="btnForward"><span class="glyphicon
glyphicon-chevron-up"></span></div>
<div id="btnBack"><span class="glyphicon glyphicon-
chevron-down"></span></div>
<div id="btnLeft"><span class="glyphicon glyphicon-
chevron-left"></span></div>
<div id="btnRight"><span class="glyphicon glyphicon-
chevron-right"></span></div>

<div id="duration">
   <div style="width:160px;">
      <input type="range"
           value="5"
           min="0"
           max="10" /  </div>
</div>
<p>1</p>
   </div>
</body>
</html>
```

C. Arduino.js programski kod

```
/// <reference path="_references.js" />
$(function () {
   $("[type=range]")
      .change(function () {
         //alert($(this).val());
      })
      .mousedown(function () {
         $("div#main p").css("display", "inline-block");
      })
      .mouseup(function () {
         $("div#main p").css("display", "none");
      });

   $(document).mousemove(Vrijednost);
})

function Vrijednost() {
   $("div#main p").html($("[type=range]").val());
}
```



Slika 4. Arduino autić sa 4 motora



Slika 5. 4DC motor driver modul i RF prijamnik i predajnik

D. *Arduino autić*

Arduino autić sastoji se od sljedećih komponenti:

- Kućište sa 4 kotača (slika 4.),
- 4 DC motora (slika 4.),
- Kućište za baterije,
- 4 DC motor driver modul (slika 5.),
- RF prijamnik (slika 5.),
- Arduino Uno

Za potrebe ovog projekta napisana je biblioteka za arduino platformu arduino.h. Sama biblioteka olakšava korištenje i upravljanje motorima na arduino autiću.

E. *Izgled koda bilbioteke arduino.h*

```
class Motor {
private:
byte MOTOR_LEFT[3];
byte MOTOR_RIGHT[3];
void start_driving_backward_right_motor(short speed);
void start_driving_backward_left_motor(short speed);
        void stop_left_motor();
        void stop_right_motor();
public:
        const short MAX_SPEED = 255;
        const float EASY = 0.9;
        const float MEDIUM = 0.75;
        const float HARD = 0;

Motor (byte motor_left_pwr, byte motor_left_gnd, byte motor_left_pwm, byte motor_right_pwr, byte motor_right_gnd, byte motor_right_pwm);

void start_turning_left(short speed, float amount);

void start_turning_right(short speed, float amount);

void start_turning_left_backward(short speed, float amount);

void start_turning_right_backward(short speed, float amount);

void start_rotating_left(short speed);

void start_rotating_right(short speed);

void stop_both_motors();

void start_driving_forward(short speed);

void start_driving_backward(short speed);

void start_driving_forward_right_motor(short speed);

void start_driving_forward_left_motor(short speed);

}; #endif
```

V. ZAKLJUČAK

Ovaj projekt zamišljen je kao dobar uvod u robotiku za učenike i studente. Web aplikacije i kontrola motora pomoću njih već su visoko zastupljene u svijetu oko nas, a sigurno su i budućnost robotike. Dovršenje jednog projekta je u biti gotovo uvijek početak nekog sljedećeg u smislu nadogradnje postojećeg novim idejama i mogućnostima. Opremanje ovakvog autića raznim senzorima, te prikupljanje i prikazivanje tih podatka unutar web aplikacije vjerojatno su sljedeći korak. Rad sa web aplikacijama i mikrokontrolerima je uvijek zanimljiv i dinamičan, te je samo mašta granica.

LITERATURA

[1] P. Zenzerović, Arduino kroz jednostavne primjere, Hrvatska zajednica tehničke kulture, Zagreb, 2014.

[2] M. Benzi, Getting Started with Arduino, 2nd Edition, O' Reily Media, 2011.

[3] Brock Craft, Arduino projects for dummies, Wiley, 2013.

[4] http://www.arduino.cc/ (pristupljeno, 1.2.2016.),

[5] http://www.w3schools.com/ (pristupljeno, 1.2 2016.),

[6] http://www.stackoverflow.com (pristupljeno, 2.2.2016.).

Informacijski sustav visokih učilišta - analiza slučaja za Veleučilište u Šibeniku

Sandra Krajačić*, Lana Topolčić* i Frane Urem*
* Veleučilište u Šibeniku/Odjel menadžmenta, Šibenik, Hrvatska
skrajaci@vus.hr, ltopolcic@vus.hr, frane.urem@vus.hr

Sažetak - Projekt uvođenja Informacijskog sustava visokih učilišta je započet sredinom 2000. godine od strane Ministarstva znanosti, obrazovanja i sporta u sklopu programa informatizacije visokih učilišta u Republici Hrvatskoj. Kao potpuno integriran informacijski sustav znatno je utjecao na poslovne procese svih dionika u visokom obrazovanju. U ovom radu su posebno opisane osnovne značajke sustava, ali i iskustva Veleučilišta u Šibeniku koje je uvođenje sustava završilo tijekom akademske godine 2014/2015.

I. UVOD

Informacijski sustav visokih učilišta (ISVU) predstavlja jedan od najstarijih i najvećih projekata informatizacije visokog školstva u Republici Hrvatskoj (RH). Vrlo brzo nakon uspješnog uvođenja ISVU-a, slijedio je niz drugih sličnih projekata kao što su Informacijski sustav studentskih prava (ISSP), Informacijski sustav akademskih kartica (ISAK) i Nacionalni informacijski sustav prijava na visoka učilišta (NISPVU). Iako je tijekom godina dolazilo do određenih promjena na infrastrukturi sustava (virtualizacija poslužitelja, promjene mrežne infrastrukture i sigurnosna unaprjeđenja), početna arhitektura ISVU-a zasnovana na tehnologijama otvorenog računarstva i danas bez većih problema odgovara rastućim potrebama nadogradnje i unaprjeđenja sustava. Veleučilište u Šibeniku (VUŠ) je jedno od mlađih visokih učilišta u RH pa je uvođenju ISVU-a pristupilo tek 2013. godine, a u ovom radu su opisana iskustva ustanove na takvom projektu.

II. POVIJESNI RAZVOJ SUSTAVA

Sredinom 2000. godine tadašnje Ministarstvo znanosti i tehnologije (MZT), a danas Ministarstvo znanosti obrazovanja i sporta (MZOS), započinje program informatizacije visokih učilišta. Tada je uočeno kako većina visokih učilišta ne koristi informacijske tehnologije u poslovima studentskih službi (prijava ispita, izdavanje potvrda studentima) što je dovodilo do velikog administrativnog opterećenja, ali i loše kvalitete dokumentacije i popratnih izvješća. Zato je u sklopu planirane informatizacije odobren početak projekta razvoja integriranog informacijskog sustava koji bi olakšao administrativne poslove na visokim učilištima, ali i osigurao centralnu evidenciju svih podataka bitnih za vođenje visokog školstva.

Razvoj programske podrške je početno povjeren Zavodu za primjenjeno računarstvo (ZPR) na Fakultetu elektrotehnike i računarstva (FER) u Zagrebu, a naknadno od 2004. godine održavanje sustava i podršku korisnicima preuzima Sveučilišni računski centar (Srce) u Zagrebu. Od kolovoza 2011. godine Srce preuzima i razvoj programske podrške. Prvi upisni list na produkcijskom sustavu je evidentiran na Sveučilištu u Zagrebu (Fakultet strojarstva i brodogradnje) 12. srpnja 2001. Krajem 2003. godine je isporučena verzija sustava slična današnjoj i od tada broj korisnika neprekidno raste (prema dostupnim podacima iz [1], krajem 2015. godine je ISVU sustav koristilo 111 visokih učilišta u RH).

III. ARHITEKTURA I ORGANIZACIJA SUSTAVA

Arhitektura ISVU-a je u potpunosti distribuirana preko središnjeg i lokalnih čvorova [1]. Ovakva arhitektura osigurava visoku pouzdanost (zajednički podaci su smješteni na središnjem čvoru, a lokalne baze podataka sadrže sve korisničke podatke i sinkroniziraju se preko središnje baze podatka) i propusnost sustava (korisnički upiti i unosi podataka se raspodjeljuju po lokalnim čvorovima). Svaki čvor sadrži poslužitelje s bazom podataka, poslužitelj za autentifikaciju (od rujna 2015. godine autentifikacija ISVU korisnika se obavlja preko AAI@EduHr autentifikacijske i autorizacijske infrastrukture) i web poslužitelj. Sustav je povezan s ISSP-om pa se izbjegava višestruki unos matičnih podataka o studentima. Prema dostupnim podacima iz [2] ISVU sustav je krajem 2015. godine instaliran na približno 30 virtualnih poslužitelja, a sustav čini približno 100 različitih servisa i aplikacija.

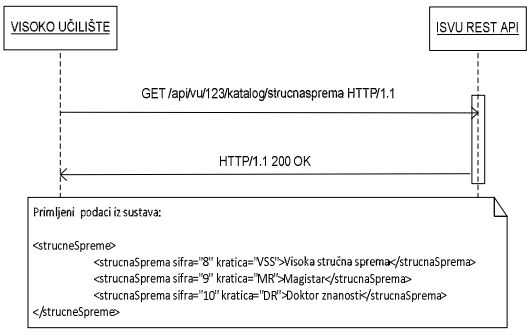
Sustav je vrlo siguran jer se unos većine podataka na visokim učilištima obavlja s posebnim klijentskim aplikacijama (modulima) koje su instalirane na računala čija je IP adresa (engl. Internet Protocol address) zabilježena na centralnom poslužitelju, u posebnoj bazi podataka. Komunikacija između klijenata i poslužitelja je kriptirana. Redovno se vrše arhiviranja baza podataka uz inkrementalno arhiviranje dnevničkih zapisa. Za korisnike su definirana prava pristupa unutar pojedinih modula, a sve operacije s podacima se bilježe za korisnike koji ih obavljaju.

Kao kod sličnih integriranih sustava (npr. ERP sustavi) i kod ISVU-a su određeni dijelovi poslovanja povezani s posebnim korisničkim modulima. Iz prethodno navedenih sigurnosnih razloga posebno su izvedeni klijentski moduli u JAVA programskom jeziku: *Studiji i studenti* (evidencija osobnih podataka o studentima i svi podaci o tijeku studiranja), *Ispiti* (evidencija ispita), *Isvu koordinator* (upravljanje dijelovima sustava od strane ISVU koordinatora), *Kadrovi* (evidencija osoba

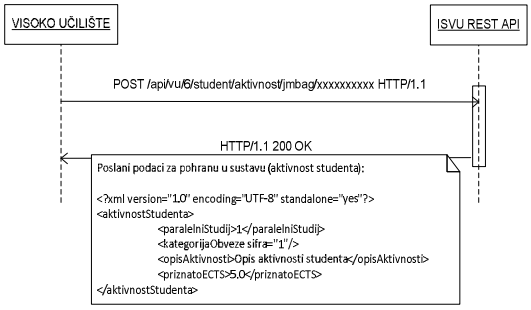
zaposlenih na visokom učilištu i organizacijskih jedinica unutar visokog učilišta), *Raspored sati* (izrada i objava rasporeda sati te rezervacija učionica), *Disis* (izrada diploma i dopunskih isprava o studiranju) i *Print server* (poseban servis za kontrolu ispisa potvrda). Nastavnici i studenti koriste web module putem uobičajenih internet preglednika s bilo kojeg računala koje ima internet pristup (prema iskustvima takvih korisnika na VUŠ-u do sada još nije zabilježen ili prijavljen slučaj nemogućnosti rada pojedinog web modula na određenom internet pregledniku, što je npr. čest slučaj kod web aplikacija u financijskom ili javnom sektoru). Takvi moduli su: *Studomat* (prijava ispita i izbornih kolegija, ispis različitih izvješća i potvrda za studente, upis studijske godine), *Skladište podataka* (razni izvještaji i statistički podaci, pohranjeni u posebnom spremištu podataka kako bi se rasteretio produkcijski sustav) i *Nastavnički portal* (unos nastavnih planova, ocjena, tema završnih i diplomskih radova).

Iako je visoko integriran sustav, ISVU omogućava korisnicima izradu vlastitih programskih sučelja za dohvat i izmjenu resursa (podataka) visokog učilišta putem tzv. REST (engl. REpresentational State Transfer) API (engl. Application Programming Interface) sučelja. Prva verzija ISVU REST API sustava je razvijena u prosincu 2011. godine, a u srpnju 2012. godine sustav izlazi iz pilot faze u punu produkciju. Pristup resursima omogućen je putem URI-ja (engl. Uniform Resource Identifier), a za korištenje REST API-ja potrebno je obaviti HTTP (engl. HyperText Transfer Protocol) zahtjev i obraditi odgovor kojeg sustav dostavlja u obliku XML (engl. EXtensible Markup Language) datoteke prema unaprijed definiranoj XML shemi.

Kod poziva se koriste standardne HTTP metode: GET (dohvat podataka s poslužitelja), POST (dostava podataka na poslužitelj) , PUT (zahtjev poslužitelju za pohranu podataka na zadanom URI-u) i DELETE (brisanje podataka). Pojednostavljeni prikaz korištenja ISVU REST API-a za dohvat podataka je vidljiv na slici 1, a primjer dostave podataka je prikazan na slici 2.



Slika 1. Zahtjev i odgovor kod dohvata podataka iz kataloga stručne spreme



Slika 2. Zahtjev i odgovor kod dostave podataka o aktivnostima studenta na paralelnom studiju

Trenutno su dostupne dvije verzije ISVU REST API-a. Više o verziji 1 je moguće pronaći na [3], a o verziji 2 na [4]. Detaljnije o REST arhitekturi je moguće pronaći na [5].

IV. POKAZATELJI KORIŠTENJA SUSTAVA

Sveučilišni računski centar redovito mjesečno obavještava korisnike ISVU-a o promjenama i unaprjeđivanju sustava, ali i o statistici korištenja sustava na [6]. Iz takvih izvora je vidljiva stalna aktivnost na edukaciji korisnika (početkom 2015. godine je održana 100. edukacija korisnika i drugo okupljanje korisnika ISVU REST API-a). Također su vidljiva vršna opterećenja sustava tijekom redovitih ispitnih rokova na visokim učilištima (npr. u zimskom ispitnom roku tijekom siječnja i veljače 2015. godine je obrađeno 450.000 ispita). Također je zanimljiv pokazatelj korištenja ISVU REST API-a prema kojema se bilježi peterostruko povećanje broja zahtjeva u 2015. godini u odnosu na 2014. godinu tako da broj zahtjeva kontinuirano raste (samo tijekom veljače 2015. obrađeno je 12.912.624 zahtjeva). U 2015. godini je zabilježeno vršno opterećenje 17/09/2015 kada je obrađeno 1.821.583 zahtjeva na ISVU REST API sustavu. Ukupno je 136.622 upisanih studenata u akademskoj godini 2015/2016. Također se bilježi stalni rast korištenje *Dsis* modula za generiranje svjedožbi, diploma i dopunskih isprava pa ih je npr. tijekom studenog 2015. godine generirano 20.206. Većina prijava ispita se odvija preko modula *Studomat* (prosječno 95%) što doprinosi znatnom radnom rasterećenju studentskih službi. Na sustavu je ispunjeno približno 100.000 anketnih obrazaca do kraja 2015. godine.

V. UVOĐENJE ISVU SUSTAVA NA VELEUČILIŠTU U ŠIBENIKU

Odluka o uvođenju ISVU-a je za VUŠ predstavljala dugoročnu i stratešku odluku. Ustanova je prilikom svog osnutka bila upoznata sa svim prednostima uvođenja ISVU-a, ali i svjesna da mora prije konačne odluke o pokretanju takvog projekta ispuniti niz preduvjeta u vezi poslovne organizacije i kvalitete nastave. Budući je VUŠ osnovano 2006. godine, u razvoju ustanove je početno definirano petogodišnje razdoblje u kojem je trebalo ispuniti sve uvjete potrebne za dobivanje trajnih dopusnica u izvođenju tri preddiplomska studija i jednog specijalističkog stručnog diplomskog studija. Ustanova je postigla zadane ciljeve 2011. godine i uspješno primjenila

normu kvalitete ISO 9001:2008. Usvojen je akcijski plan o pripremi za neovisnu periodičnu reakreditacijsku prosudbu u 2012. godini, a tijekom 2013. godine je uspješno provedena vanjska neovisna periodična prosudba unutarnjeg sustava osiguranja kvalitete VUŠ-a prema ESG standardima i smjernicama kvalitete u visokom obrazovanju.

Početkom svibnja 2013. godine je dostavljen prijedlog MZOS-a o uključenju VUŠ-a u ISVU. Projekt uvođenja službeno je pokrenut 21. svibnja 2013. godine potpisivanjem ugovora o suradnji na uspostavi, korištenju i održavanju ISVU-a između VUŠ-a, MZOS-a i Srca.

Prema dostavljenom hodogramu administrativnih poslova od strane MZOS-a i Srca, imenovana je radna grupa na VUŠ-u za uvođenje sustava u listopadu 2013. godine, a kao voditelj projekta i ključna osoba u provedbi je imenovana ISVU koordinatorica. Početno nisu bili jasno utvrđeni postupci kojima bi koordinatorica informirala upravu visokog učilišta o postignutom napretku kao i rješavanju uočenih problema. Ovaj nedostatak je uklonjen uvođenjem dodatne dokumentacije u postojećem ISO sustavu, ali su i dogovoreni češći redovni sastanci unutar radne grupe i po potrebi u proširenom sastavu s upravom VUŠ-a. U prosincu 2013. godine radna grupa je u dogovoru s upravom VUŠ-a utvrdila slijedeći vremenski okvir izvedbe projekta do početka akademske godine 2015/2016 (listopad 2015.):

- organizacija i provedba tečajeva za korisnike sustava, s rokom provedbe do lipnja 2014.
- učenje iz iskustava drugih ustanova (savjetovanja i radionice s drugim visokim učilištima koja su uspješno uvela sustav, a veličinom i organizacijom su slične Veleučilištu), s rokom provedbe do lipnja 2014.
- motiviranje zaposlenika za uvođenje novog sustava (ukazivanje na dobrobiti i nove funkcionalnosti na radionicama s predstavnicima drugih ustanova, stimulativne mjere za voditelja projekta i osobe angažirane na unosu povijesnih podataka), kontinuirano do lipnja 2015.
- upis nove generacije studenata na prvoj godini preddiplomskih studija i specijalističkog stručnog diplomskog studija u ljetnom i jesenskom upisnom roku 2014. godine
- prijepis preostalih podataka iz starog sustava, kontinuirano od siječnja do lipnja 2015.
- puna produkcija sustava na poslovima studentske službe i upisima za sve generacije studenata u ljetnom i jesenskom upisnom roku 2015. godine

Prethodno navedeni provedbeni okvir je u potpunosti ispunjen u zadanim rokovima. Vidljivo je kako se VUŠ (temeljem iskustava drugih ustanova koje su uvodile ISVU u prethodnim razdobljima) odlučio na kombinaciju sporog modela uvođenja sustava tijekom akademske godine 2014/2015 (upis samo nove generacije studenata na prvoj godini studija), a nakon utvrđene uspješne provedbe odobreno je brzo uvođenje sustava za preostale studente do početka slijedeće akademske godine. Na takav način je bitno smanjeno početno opterećenje zaposlenika službe studentskih poslova na unosu podataka (podaci o novoupisanim studentima su bili automatski dostupni u ISVU-u preko NISPVU-a) koji su se onda mogli kvalitetnije posvetiti obuci i planiranju budućeg unosa podataka iz starog sustava.

Velik doprinos uspješnom uvođenju ISVU-a na Veleučilištu u Šibeniku je dala ISVU koordinatorica jer je kao voditeljica informatičke službe posjedovala izvrsna tehnička znanja o novom i starom informacijskom sustavu. U početku su zabilježeni manji problemi u komunikaciji i izvještavanju svih dionika na projektu uvođenja ISVU-a, ali su pravovremeno rješavani dopunama izvedbene dokumentacije i češćim sastancima radne grupe. Nisu zabilježeni otpori uvođenju ISVU-a među zaposlenicima i studentima. ISVU Centar potpore u Srcu je izvrsno i na vrijeme obavljao sve zadatke na prijenosu podataka i rješavanju prijavljenih problema.

Može se reći kako je uvođenje ISVU-a na VUŠ-u bitno olakšalo sve administrativne poslove službe studentskih poslova, ali i unaprijedilo sve povezane poslove izvještavanja i statističkih obrada. U slijedećem razdoblju se očekuje proširenje vrijednosti uvedenog sustava korištenjem ISVU REST API sučelja na pomoćnim sustavima Veleučilišta u Šibeniku. Također se očekuje od svih zaposlenika da u budućnosti povećaju fond svog znanja o implementiranom sustavu, da počnu samostalno rješavati probleme konfiguriranja i postavljajnja specifikacija za nove funkcionalnosti ili optimiranje postojećih.

VI. ZAKLJUČAK

Prema svim pokazateljima uvođenje ISVU-a u prostor hrvatskog visokog obrazovanja je vrlo uspješno. Uvođenjem ISVU REST API-a otvoren je prostor visokim učilištima za razvoj vlastitih programskih rješenja koja sa sustavom mogu razmjenjivati podatke. Veleučilište u Šibeniku je uspješno uvelo sustav slijedeći zadanu projektnu metodologiju, ali i iskustva drugih sličnih visokih učilišta.

LITERATURA

[1] Srce Zagreb, Informacijski sustav visokih učilišta, 2015., dostupno na: www.isvu.hr/javno/hr/index.shtml [1. ožujka 2016]

[2] E. Prendivoj, Srce Zagreb, „“Kako je implementiran ISVU REST API?“, 2014., dostupno na: http://www.srce.unizg.hr/fileadmin/Srce/docs/isvu/ISVU_REST_API_DAN2014/ISVU-REST-API-Kako-je-implementiran.pdf [1. ožujka 2016]

[3] Srce Zagreb, “ISVU REST API v1”, 2016, dostupno na: https://www.isvu.hr/api/dokumentacija/v1/index.html [1. ožujka 2016]

[4] Srce Zagreb, “ISVU REST API v2 - HAL”, 2016, dostupno na: https://www.isvu.hr/api/dokumentacija/v2-hal/index.html [1. ožujka 2016]

[5] Martin Fowler, "Richardson Maturity Model - steps toward the glory of REST", 2010., dostupno na: http://martinfowler.com/articles/richardsonMaturityModel.html [1. ožujka 2016]

Srce Zagreb, “Periodički izvještaji usluge: Informacijski sustavi visokih učilišta (ISVU)”, 2015., dostupno na: http://www.srce.unizg.hr/dokumenti/periodicki/isvu [1. ožujka 2016]

Mobilne aplikacije u visokom obrazovanju

Maja Blašković*, Marija Fumić* i Frane Urem*
* Veleučilište u Šibeniku/Odjel menadžmenta, Šibenik, Hrvatska
mblaskov@vus.hr, m.fumic@vus.hr, frane.urem@vus.hr

Sažetak - Moderno visoko obrazovanje danas podrazumijeva potpunu informatizaciju, a mobilne aplikacije u tome postaju bitan čimbenik. Pri tome svakodnevno nastaju mobilne aplikacije koje postaju učinkoviti obrazovni alati te zadovoljavaju različitih obrazovnih i akademskih potreba studenata. U ovom radu su prikazane određene smjernice mogućih primjena mobilnih aplikacija kao i stanje primjene takvih aplikacija u visokom obrazovanju Republike Hrvatske.

I. UVOD

Mobilne aplikacije danas bilježe stalan rast preuzimanja u različitim područjima zabave, informiranja i svakodnevnog poslovanja. Glavni razlozi su dostupnost (jednostavna instalacija, vrlo često su besplatne), stalno jačanje informacijske infrastrukture (bežične mreže, računalstvo u oblaku) i procesne snage uređaja kojima su namijenjeni (dlanovnici i pametni telefoni) te jednostavnost korisničkog sučelja (ekrani osjetljivi na dodir). Uslijed naglog razvoja mobilne infrastrukture i uređaja, mobilne aplikacije postaju sve bitniji dio tržišta softvera. Pri tome se uočavaju određene svjetske smjernice u područjima njihove primjene u visokom obrazovanju.

U ovom radu je prikazano i nekoliko mobilnih aplikacija koje imaju različita područja primjene u hrvatskom visokom školstvu. Za potrebe testiranja instalirane su na odgovarajuće mobilne platforme pri čemu su provjerene osnovne funkcionalnosti i područja primjene.

II. POVIJESNI RAZVOJ MOBILNIH APLIKACIJA

Mobilna aplikacija se može definirati kao program namijenjen pametnim telefonima (engl. *smart phone*) i dlanovnicima (engl. *tablet computer*) [1]. Starija mobilna programska rješenja su objavljena za tada dostupne tzv. PDA (engl. *Personal Digital Assistant*) mobilne uređaje koje je proizvodila tvrtka Psion u '90 godinama prošlog stoljeća. Uređaji su koristili EPOC operacijski sustav sa instaliranim aplikacijama kao što su dnevnik, baze podataka, dijagrami i sl.. Na kasnijim modelima se moglo nadograđivati operacijski sustav i instalirati aplikacije drugih proizvođača korištenjem softverskih paketa koji su bili dostupni na internetu. Naknadno je razvoj EPOC operacijskog sustava nastavljen pod poznatijim nazivom Symbian, kao zajednički poduhvat tvrtki Ericsson, Nokia, Motorola i Psion 1998. godine (Nokia je u potpunosti preuzela razvoj 2010. godine).

Međutim, 2007. godine tvrtka Apple predstavlja svoj prvi pametni telefon (poznatiji pod nazivom IPhone), a 2008. godine uvodi digitalni servis AppStore koji omogućava preuzimanje aplikacija drugih proizvođača. Gotovo istovremeno (u kolovozu 2008. godine) je od strane Googlea predstavljen digitalni servis Android Market (danas poznat pod nazivom Google Play), namijenjen mobilnim uređajima koji su koristili Android operacijski sustav. U studenom 2008. godine Microsoft predstavlja svoj servis Windows Marketplace (danas poznat pod imenom Windows Store) za mobilne uređaje koji koriste Windows operacijski sustav. Koncept digitalnog servisa koji (nakon prethodne provjere) nudi potencijalnim korisnicima preuzimanje vrlo jednostavnih (u odnosu na aplikacije za stolna računala) i dobro oblikovanih aplikacija je u potpunosti i nepovratno promijenio tržište softvera.

III. BITNI POKAZATELJI PRIMJENE MOBILNIH APLIKACIJA

Prema istraživanju analitičke tvrtke IDC [2], danas je svjetsko tržište mobilnih aplikacija po raspoloživim operacijskim sustavima zastupljeno kao u Tablici I. Vidljiv je stalni rast udjela mobilnih aplikacija namijenjenih Android mobilnoj platformi.

Posebno su zanimljivi podaci prikazani u Tablici II, iz svjetskog istraživanja analitičke tvrtke Gartner [3], koje prikazuje broj preuzimanja mobilnih aplikacija od 2011. godine, s predviđanjima do 2017. godine. Vidljiv je brzi rast preuzimanja besplatnih aplikacija dok je takav rast kod aplikacija čija se uporaba naplaćuje bitno manji.

U Tablici III su navedeni podaci iz svjetskog istraživanja tvrtke Ericsson [4] koje ukazuje na buduća povećanja udjela pametnih telefona u ukupnom broju mobilnih uređaja. Prema očekivanjima iz [4], do 2020. godine većina svjetske populacije će posjedovati pametni telefon, a razlog je naglo širenje širokopojasne bežične infrastrukture kao i stalni trend pada cijena takvih uređaja.

Iako ne postoje pouzdani globalni podaci o stanju preuzimanja mobilnih aplikacija iz područja visokog obrazovanja, zanimljivi su pokazatelji mogućih područja primjena takvih aplikacija iz istraživanja [5]. U navedenom istraživanju su anketirani voditelji centara informacijske potpore na studentskim kampusima u Sjedinjenim Američkim Državama (SAD) koji su naveli najznačajnija područja buduće primjene mobilnih aplikacija u visokom obrazovanju prema Tablici IV.

Također, na području Republike Hrvatske još nisu provedena istraživanja koja bi odgovorila na bitna pitanja kao u [6] i čiji odgovori bi utvrdili glavne smjernice u

strategiji budućeg uvođenja mobilnih aplikacija za potrebe visokog obrazovanja:

- Kakve mobilne uređaji studenti već posjeduju za pristup digitalnim sadržajima i koji čimbenici utječu na kupovinu takvih uređaja (demografski, ekonomski) ?
- Kakve su navike studenata u korištenju mobilnih tehnologija (uređaji i aplikacije) i da li ih već koriste za potrebe učenja ?
- Koji su ključni događaji nastali na području mobilnih tehnologija u prethodnim razdobljima (od 2012. do 2015.) ?

TABLICA I. SVJETSKI POKAZATELJI TRŽIŠNOG UDJELA MOBILNIH APLIKACIJA PREMA OPERACIJSKIM SUSTAVIMA

Godina	*Android*	*iOS*	*Windows Phone*	*Ostali*
2015	78.0%	18.3%	2.7%	1,0%
2014	81.2%	15.2%	2.5%	1,2%
2013	75.5%	16.9%	3.2%	4,4%
2012	59.2%	22.9%	2.0%	15,8%

IZVOR: [2]

TABLICA II. SVJETSKI POKAZATELJI BROJA PREUZIMANJA MOBILNIH APLIKACIJA

	Besplatna preuzimanja / milijun	*Plaćena preuzimanja / milijun*
2011	22	2,9
2012	57,33	6,65
2013	92,88	9,19
2014	127,7	11,11
2015*	167,05	12,57
2016*	211,31	13,49
2017*	253,91	14,78

IZVOR: [3]

TABLICA III. SVJETSKI POKAZATELJI UDJELA PAMETNIH TELEFONA U UKUPNOM BROJU MOBILNIH UREĐAJA, PREMA ZABILJEŽENOJ NAPLATI

	2014 god. broj u mlrd.	*2020 god. broj u mlrd.*
Obični mobilni telefoni	4,25	2,7
Pametni telefoni	**2,6**	**6,1**
Mobilna računala, dlanovnici i usmjerivači	0,25	0,4
Ukupno	**7,1**	**9,2**

IZVOR: [4]

TABLICA IV. PROCJENA KLJUČNIH PODRUČJA BUDUĆE PRIMJENE MOBILNIH APLIKACIJA NA PODRUČJU VISOKOG OBRAZOVANJU PREMA ISTRAŽIVANJU [5]

Područje primjene	*Očekivani udjel do 2017*
Ponuda različitih proizvoda i usluga za studente	25,00%
Sustavi za upravljanje učenjem (engl. Learning Managment System, LMS)	25,00%
Sustavi za razmjenu poruka i organizaciju aktivnosti	14,00%
Društvene mreže	6,00%
Uredske aplikacije	6,00%
Aplikacije za održavanje nastave	6,00%
Portali	4,00%
Konferencijski i sustavi za grupni rad	2,00%
E-učenje	2,00%
ERP sustavi	2,00%
Ostalo	19,00%

IZVOR: [5]

IV. PRIKAZ MOBILNIH APLIKACIJA U HRVATSKOM VISOKOM OBRAZOVANJU

U nastavku je prikazano nekoliko mobilnih aplikacija koje su našle primjenu kod studentske populacije u Republici Hrvatskoj. Većina aplikacija je razvijena kao plod inicijative Hrvatske akademske i istraživačke mreže (CARNET) i Sveučilišnog računskog centra Zagreb (SRCE) kao krovnih organizacija na području informacijskih tehnologija u hrvatskom obrazovnom sustavu. Može se reći kako je ponuda mobilnih aplikacija na području hrvatskog obrazovnog sustava vrlo slaba i svodi se na aplikacije opisane u ovom radu. Međutim, u novije vrijeme se pojavljuju značajne inicijative studentskih udruga kao što je eStudent koje prepoznaju potencijal razvoja mobilnih aplikacija od strane samih studenata jer im takve aktivnosti pomažu u stjecanju novih vještina i budućeg profesionalnog razvoja. Određena visoka učilišta (Sveučilište u Osijeku i Sveučilište u Rijeci) su također prepoznala značaj razvoja mobilnih aplikacija od strane samih studenata tijekom njihovog obrazovanja pa su za te namjene organizirali posebne laboratorije za razvoj mobilnih aplikacija.

A. Mobilna aplikacija mCARNet

mCARNet [6] je aplikacija za pristup mobilnim uslugama koje pruža CARNet, a dostupne su jedinstvenom prijavom putem elektroničkog identiteta u sustavu AAI@EduHr. Prijavom u ovu aplikaciju student se prijavljuje na sve CARNet usluge za koje mu je odobreno korištenje, bez potrebe za višestrukom autentifikacijom na različitim mjestima. Aplikacija omogućava i jednostavno podešavanje postavki za pristup Eduroam bežičnim mrežama, kao i upute za podešavanje

računa elektroničke pošte. Osnovno sučelje mCARNet aplikacije je prikazano na slici 1.

B. *Mobilna aplikacija CARNet mToken*

CARNet mToken [7] je aplikacija za generiranje jednokratnih lozinki potrebnih za određene CARNetove usluge ili državne e-servise kao što je npr. e-Građani. Izgled CARNet mToken aplikacije je vidljiv na slici 2.

C. *Mobilna aplikacija CARNet Meduza*

CARNet Meduza [8] je mobilna aplikacija koja omogućava korištenje obrazovnih video sadržaja pojedinačnim korisnicima, ali i uspostavu vlastitih kanala za objavu i distribuciju video sadržaja ustanova članica i njihovih korisnika. Izgled CARNet Meduza aplikacije je vidljiv na slici 3.

D. *Mobilna aplikacija CARNet Libar*

Libar [9] je aplikacija koja se koristi za preuzimanje i čitanje EPUB i PDF sadržaja kreiranog pomoću istoimene CARNetove web aplikacije (slično znatno poznatijem Amazon Kindle servisu). Prijava korisnika odvija se putem AAI@Edu.hr korisničkog računa. Korisniku je za preuzimanje dostupan sav EPUB i PDF sadržaj kojeg je on sam autor, kao i sadržaji drugih korisnika ukoliko prava pristupa to dozvoljavaju. Aplikacija je integrirana i s Lamaro Digital korisničkom policom eduKnjižare, te omogućava pregledavanje i čitanje svih sadržaja kupljenih preko Lamaro digital eduKnjižare. Izgled CARNet Libar aplikacije je vidljiv na slici 4.

E. *Mobilna aplikacija Studentski Centar Hrvatska*

Besplatna aplikacija Studentski Centar Hrvatska [10] omogućava jednostavan pregled ponude studentskih poslova svakodnevno objavljenih na stranicama studentskih centara većih sveučilišta u Republici Hrvatskoj. Na brz i lak način je moguće pregledati ponudu poslova pojedinog grada po kategorijama kao i upisom ključnih riječi u tražilicu. Izgled aplikacije Studentski Centar Hrvatska je vidljiv na slici 5.

F. *Mobilna aplikacija Zagreb Menza*

Aplikacija Zagreb Menza [11] prikazuje jelovnike zagrebačkih studentskih restorana. Omogućena su tri različita prikaza jelovnika (ovisno o danima u tjednu, odnosno radu studentskog restorana). Podaci o jelovnicima se ažuriraju prema objavljenim podacima Studentskog centra Zagreb. Izgled aplikacije je vidljiv na slici 6.

G. *Mobilna aplikacija Xica*

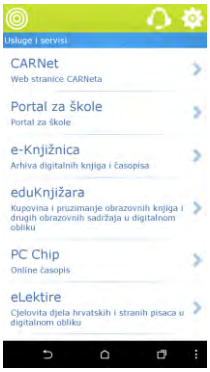
Xica [12] omogućava brzu provjeru stanja na studentskoj iskaznici (popularan naziv Xica), popis računa naplaćenih u studentskoj prehrani s detaljnim ispisom stavaka. Od verzije 0.6 dodana je funkcionalnost Xstat (matematički algoritam koji procjenjuje ostvarenu potrošnju i prikazuje očekivano stanje do kraja mjesečnog obračuna). Izgled aplikacije je vidljiv na slici 7.

H. *Mobilna aplikacija FOIapp*

FOIapp [13] je službena aplikacija Fakulteta organizacije i informatike (FOI) Varaždin. Studentima FOI-a omogućava: lakše snalaženje u gradu Varaždinu, brz pronalazak lokacija dvorana i kabineta u fakultetskim zgradama, pregled informacija o djelatnicima FOI-a (konzultacije, e-mail…),pregled informacija o kolegijima (broj ECTS-a, semestar u kojem se izvode…), čitanje vijesti s FOI web-a, unos vlastitog rasporeda sati te evidenciju izostanaka, prikaz jelovnika Varaždinskih menzi, prikaz trenutnog stanja FOI Redomata za Referadu i Računovodstvo s uzimanjem broja, prikaz stanja na studentskoj iskaznici. Izgled aplikacije je vidljiv na slici 8.

I. *Mobilna aplikacija Državna Matura*

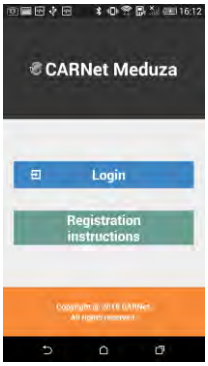
Državna Matura [14] omogućava pregled svih učilišta, pretraživanje ispita državne mature, rješavanje kviza uz objavu i prikaz rezultata, popis bitnih datuma za pojedini ispitni rok, računanje bodova za pojedino učilište i sl. Dodatne mogućnosti aplikacije su: pretraživanje visokih učilišta, ispitni zadaci s provjerama znanja, novosti sa službenih stranica NCVVO, organizacija podataka o visokim učilištima za koje je pristupnik zainteresiran. Osnovno sučelje aplikacije Državna Matura je prikazano na slici 9.



Slika 1. Mobilna aplikacija mCARNet



Slika 2. Mobilna aplikacija CARNet mToken



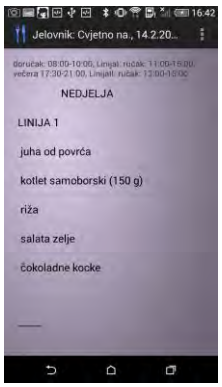
Slika 3. Mobilna aplikacija CARNet Meduza



Slika 4. Mobilna aplikacija CARNet Libar



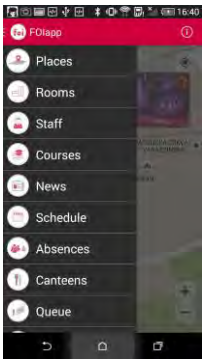
Slika 5. Mobilna aplikacija Studentski Centar Hrvatska



Slika 6. Mobilna aplikacija Zagreb Menza



Slika 7. Mobilna aplikacija Xica



Slika 8. Mobilna aplikacija FOIapp



Slika 9. Mobilna aplikacija Državna Matura

V. ZAKLJUČAK

Pregledom Google Play digitalnog servisa uočava se veliki broj dostupnih mobilnih aplikacija iz područja visokog obrazovanja. Međutim, samo mali broj takvih rješenja je namijenjen hrvatskim korisnicima i većina ih je prikazana u ovom radu. Vjerojatni razlozi takvog stanja su relativno malen broj potencijalnih korisnika (koji bi opravdali komercijalni razvoj aplikacija) i rijetki sigurni izvori podataka koje bi mobilne aplikacije obrađivale i prikazivale korisnicima. Od prikazanih rješenja posebno se ističe *FOIapp* koji u potpunosti slijedi slična rješenja dostupna na najboljim sveučilišnim kampusima u svijetu.

LITERATURA

[1] M.Rouse, „Wireless and mobile glossary“, 2013., dostupno na: http://whatis.techtarget.com/definition/mobile-app

[2] IDC, „Smartphone OS Market Share“, 2015, dostupno na: http://www.idc.com/prodserv/smartphone-os-market-share.jsp [1. ožujka 2016]

[3] Gartner, Press Release, 2015., dostupno na: http://www.gartner.com/newsroom/id/2592315 [1. ožujka 2016]

[4] Ericsson, „Mobility report 2015“, 2015., dostupno na: http://www.ericsson.com/res/docs/2015/ericsson-mobility-report-june-2015.pdf [1. ožujka 2016]

[5] B.Chen, R. Seilhamer, L. Bennett, S. Bauer, "Students' Mobile Learning Practices in Higher Education: A Multi-Year Study", EduCause Review, 2015., dostupno na: http://er.educause.edu/articles/2015/6/students-mobile-learning-practices-in-higher-education-a-multiyear-study [1. ožujka 2016]

[6] CARNET, CARNet mCarnet, 2012., dostupno na: http://www.carnet.hr/novosti/novosti?news_id=2468 [1. ožujka 2016]

[7] CARNET, CARNet mToken, 2016., dostupno na: http://www.carnet.hr/mtoken [1. ožujka 2016]

[8] CARNET, CARNet Meduza, 2016., dostupno na: https://meduza.carnet.hr [1. ožujka 2016]

[9] CARNET, CARNet Libar, 2015., dostupno na: http://www.carnet.hr/libar [1. ožujka 2016]

[10] N.Kaćunko, B.Bosnić, Mace Games, “Studentski Centar Hrvatska”, 2015., dostupno na: https://play.google.com/store/apps/details?id=com.studentservisapp.parsing&hl=hr [1. ožujka 2016]

[11] E. Sunić, “Menza Zagreb”, 2015., dostupno na: https://play.google.com/store/apps/details?id=com.zmenza.zagrebmenza&hl=hr [1. ožujka 2016]

[12] J.Šokčević, J.Stuhli, “Xica”, 2011, dostupno na: https://play.google.com/store/apps/details?id=com.tehnofobija.xica&hl=hr [1. ožujka 2016]

[13] J. Žemberi,T. Šulc, R. Turić, "FoiAPP", 2014, dostupno na: https://play.google.com/store/apps/details?id=hr.foi.foiapp.v2&hl=hr [1. ožujka 2016]

[14] V. Vidaković, “Državna matura”, 2015., dostupno na: https://play.google.com/store/apps/details?id=com.vedranvidakovic.drzavnamatura&hl=hr [1. ožujka 2016]

Metodologija izrade *E – learning* sadržaja za edukaciju o izradi Standarda zanimanja

I. Vunarić*, S. Grgić * i T. Babić *
* Visoko učilište Algebra – visoka škola za primijenjeno računarstvo, Zagreb, Hrvatska
iva.vunaric@racunarstvo.hr, silvija.grgic@racunarstvo.hr, tihana.babic@racunarstvo.hr

Sažetak – Zakonom o Hrvatskom kvalifikacijskom okviru Standard zanimanja definira se kao popis svih poslova koje pojedinac obavlja u određenom zanimanju i popis kompetencija potrebnih za uspješno obavljanje tih poslova. Zahtjev za upis u podregistar može podnijeti bilo koja pravna ili fizička osoba predajom ispunjenog propisanog obrasca. Zbog nacionalne važnosti da zapisi u podregistru budu što kvalitetnije zapisani ministarstvo nadležno za rad izradilo je smjernice za izradu standarda zanimanja. Smjernice propisuju postupke i načine procjene utemeljenosti prijedloga. Propisani postupci zahtijevaju određene vještine analiziranja i pisanja, a kako ne postoji sustavna edukacija kao odlično rješenje nameće se *E-learning*. Izradom *E-learning* sadržaja koji će biti dostupan svim dionicima znatno se približava cijela metodologija izrade standarda zanimanja, a kroz mogućnosti koje *E-learning* alati pružaju može se postići i provjeriti potrebna vještina. Cilj rada je pokazati kako se pisane smjernice za izradu standarda zanimanja mogu pretočiti u *E-learning* sadržaj. Ujedno ovaj rad pokazuje i kako se prikladnom metodologijom i alatima može olakšati razumijevanje i ponavljanje složenih analiza te stjecati vještina koja se kasnije može primijeniti u konkretnoj situaciji izrade standarda zanimanja. U završnom dijelu rada dani su važni dijelovi i ilustracije stvarnog *E-learning* sadržaja za izradu standarda zanimanja s naglaskom na prednosti i važnosti odabranog pristupa.

I. UVOD

U društvu stalnih ekonomskih i tehnoloških promjena te sve češćih zahtjeva za mobilnošću, cjeloživotno učenje nameće se kao nužnost, kako obrazovne tako i gospodarske politike. Upravo iz tih razloga, Hrvatski sabor 8. veljače 2013. godine usvojio je Zakon o Hrvatskom kvalifikacijskom okviru [6]. Ovim zakonom Hrvatski kvalifikacijski okvir (HKO) postao je temelj za uređenje sustava cjeloživotnog učenja na nacionalnoj razini s ciljem povećanja kvalitete cjelokupnog obrazovnog sustava te povezivanja istog s tržištem rada na način da daje temelj za razumijevanje postojećih kvalifikacija, njihov međusobni odnos, te razvoj novih kvalifikacija.

Stoga, za razumijevanje samog koncepta HKO-a važna je hipoteza da obrazovni programi stvaraju kvalifikacije, odnosno ishode učenja, koji su preduvjet za razvoj kompetencija nužnih za obavljanje ključnih poslova na radnim mjestima.

Također, radi osiguravanja kvalitete i transparentnosti primjene HKO-a, uspostavljen je javni Registar HKO-a koji evidentira sve usvojene standarde, a upravo dobro definirani i relevantni standardi predstavljaju glavni mehanizam održavanja zapošljivosti. Registar HKO-a sastoji se od tri podregistra [5]: Podregistra standarda zanimanja, Podregistra standarda kvalifikacija i Podregistra skupova ishoda učenja. Sam zahtjev za upis u podregistar može podnijeti bilo koja pravna ili fizička osoba predajom ispunjenog propisanog obrasca.

Bitno je napomenuti da su za većinu kvalifikacija krajnji korisnici poslodavci. Prema tome, poznavanje potreba poslodavaca i nužnih kompetencija za obavljanje poslova na radnim mjestima jedna je od glavnih zadaća obrazovnih institucija koje pripremaju mlade ili odrasle osobe za tržište rada [4]. Popis svih poslova koje pojedinac obavlja u određenom zanimanju i popis kompetencija potrebnih za uspješno obavljanje tih poslova, Zakonom o Hrvatskom kvalifikacijskom okviru [6], definira se kao Standard zanimanja.

II. STANDARD ZANIMANJA

Standard zanimanja temelji se na popisu svih poslova koje pojedinac obavlja na nekom radnom mjestu. Da bi te poslove obavljao uspješno, mora posjedovati kompetencije potrebne za njihovo obavljanje. Standardom zanimanja, do kojeg dolazimo istraživanjem i analizom, utvrđuju se potrebe poslodavaca te definiraju kompetencije koje su pojedincu potrebne da bi bio spreman na svijet rada. Ako istraživanje provedemo kvalitetno i dobro osmislimo standarde zanimanja, postavili smo temelje za izradbu relevantnih strukovnih kvalifikacija koje pripremaju na ulazak u tržište rada [7].

Spomenuti zahtjevi za upis standarda zanimanja u podregistar podnose se ministarstvu nadležnom za rad ispunjavanjem obrazaca *Zahtjev za upis standarda zanimanja* (A dio) kojim se dokazuje utemeljenost prijedloga standarda zanimanja te *Zahtjev za upis skupa kompetencija* (B dio) u kojim su prikazani rezultati Ankete o standardu zanimanja.

Anketa o standardu zanimanja daje podatke nužne za analizu kompetencija, a provodi se na uzorku poslodavaca koji zapošljavaju osobe s navedenim zanimanjem. Rezultate ankete analiziraju stručnjaci iz područja, te na taj način razvijaju i javno objavljuju standarde zanimanja.

A. Utemeljenost prijedloga standarda zanimanja

Pravilnik o Registru Hrvatskog kvalifikacijskog okvira nalaže obrazloženje zahtjeva za upisom standarda

zanimanja u Registar temeljem tri elementa. Riječ je o strateškoj, sektorskoj i analitičkoj utemeljenosti. U smjernicama za izradu standarda zanimanja opisan je svaki od elemenata [8].

Stratešku utemeljenost zahtijeva predlagatelj, a dokazuje je pozivanjem na relevantne sektorske strategije i ostale strateški relevantne dokumente poput nacionalnih, regionalnih, lokalnih, industrijskih i sektorskih strategija te akcijskih planova i mjera koje su dijelom njihove provedbe.

Sektorska utemeljenost obrazlaže se istraživanjem Profila sektora kojem pripada zanimanje za koje se predlaže standard. Pri opisu sektorske utemeljenosti, koriste se spoznaje o obuhvatu, potražnji i ostalim podacima dostupnim na HKO portalu. Sektorskom utemeljenošću prikazuje se sadašnja i buduća potražnja te trendovi koji govore o nadolazećim potrebama za nekim specifičnim znanjima i vještinama.

Naposljetku, analitička utemeljenost iskazuje se analizom trenutačne i dosadašnje ponude i potražnje za odabranim zanimanjem, odnosno projekcijom u petogodišnjem razdoblju za takvim i sličnim zanimanjima na tržištu rada te se dokazuje na temelju stručnih podloga dostupnima na HKO portalu [8].

B. Prikupljanje i obrada podataka o kompetencijama

Već je spomenuto da je glavni instrument za istraživanje kompetencija na radnom mjestu Anketa o standardu zanimanja (ASZ) [8]. Anketa se temelji na uzorku velikih, srednjih i malih poslodavaca koji zapošljavaju osobe s određenim zanimanjem, a opisuje stručne kompetencije vezane za zanimanje, temeljne kompetencije za cjeloživotno učenje, generičke kompetencije te uvjete rada na radnom mjestu.

Za potrebe provedbe Ankete razvijen je upitnik te se sa ciljem pojednostavljivanja provedbe preporuča on-line istraživanje, iako je moguće i terensko istraživanje u direktnom kontaktu s poslodavcima.

Prvi korak, nakon što je anketa provedena, predstavlja filtriranje podataka, odnosno provjera kojom razlučujemo korisne i iskoristive podatke od onih koji nisu prikladni za daljnju obradu. Sama obrada podataka odnosi se na prikaz učestalosti i postotaka pojedinih odgovora i ne zahtijeva korištenje složenih statističkih postupaka.

Specifičan postupak potreban je prilikom obrade podataka koji su bili dati kao odgovor na otvorena pitanja, a koja se tiču iskaza o ključnim poslovima na pojedinim radnim mjestima te znanjima i vještinama potrebnima za njihovo obavljanje. Postupak obrade podataka za ta pitanja zasniva se na grupiranju pojedinih odgovora poslodavaca koji upozoravaju na sadržajno isti ili sličan posao u zasebne skupine kako bi podaci bili pregledniji. Isti postupak se ponavlja i za odgovore na pitanje o potrebnim znanjima i vještinama u okviru svake definirane skupine ključnih poslova [8].

III. METODOLOGIJA IZRADE E-LEARNING SADRŽAJA ZA STANDARD ZANIMANJA

Koncept i proces izrade Standarda zanimanja relativno je nov. Iako različite ustanove u Republici Hrvatskoj kroz operativne projekte, financirane od Europskih strukturnih i investicijskih fondova (ESF), žele urediti sustav kvalifikacija u različitim područjima, još uvijek ne postoji dovoljno primjera dobre prakse te evaluiranih i odobrenih Standarda zanimanja koji bi mogli biti podloga za učenje i proučavanje u svrhu učenja.

Također, osim postojanja Zakona o Hrvatskom kvalifikacijskom okviru (NN 22/13), Pravilnika o Registru Hrvatskoga kvalifikacijskog okvira (NN 62/14) te Smjernica za izradu standarda zanimanja izdanih od strane ministarstva nadležnog za rad ne postoji sustavna edukacija o tome kako kvalitetno sastaviti, napisati i predložiti standard zanimanja.

Upravo iz svih navedenih razloga, kao najbolje rješenje za sustavnu edukaciju svih zainteresiranih dionika koji se žele i imaju interesa upustiti u izradu standarda zanimanja nameće se *E-learning.*

Prema klasifikaciji *E-learninga,* a na osnovu stupnja razlikovanja od tradicionalnih strategija učenja, čisti *E-learning* (engl. *pure E- learning*) [3] u kojem se sadržaj može savladati samostalno, pokazuje se kao najprikladniji oblik za ovakvu vrstu edukacije. Jednom izrađen sadržaj uz pomoć stručnjaka za *E-learning* i predlagatelja smjernica, verificiran od svih tijela nadležnih za kasniju evaluaciju pristiglih zahtjeva, može biti dugoročno rješenje za edukaciju. Čisti *E-learning* u kontekstu dugoročnosti upotrebe sadržaja te efikasnosti edukacije postavlja kao uvjet da se izradi takvog sadržaja pristupi studiozno, uz odabir pravilne metodologije.

A. Kategorizacija znanja i vještina

Kategorizacija očekivanih znanja i vještina radi se uz pomoć Bloomove taksonomije. Benjamin Bloom (1913. – 1999.) dijeli kognitivnu domenu učenja na 6 odvojenih razina usvojenosti znanja [1]: 1. znanje – prepoznavanje informacija, 2. razumijevanje – shvaćanje informacija, 3. primjena – primjena znanja u rješavanju problema, 4. analiza – razdvajanje informacija na sastavne dijelove, 5. sinteza – primjena informacija i 6. vrednovanje – prosuđivanje korisnosti. Prelazak u svaku višu razinu omogućen je isključivo ukoliko je zadovoljena razina znanja propisana svim prethodnim razinama.

Bloomova taksonomija prikladna je za ovu vrstu *E–learning* sadržaja iz razloga što potiče izgrađivanje znanja više razine na temeljima postojećih znanja nižih razina. Sagledamo li osnovni koncept izrade Standarda zanimanja, koji je opisan u prethodnom poglavlju, uočit ćemo da on u svojoj osnovi upravo polazi od znanja i razumijevanja osnovnih pojmova HKO-a, kvalifikacija, ishoda učenja kompetencija, zahtjeva i potreba poslodavaca, tržišta rada itd., a bez kojih je primjena i analiza dobivenih rezultata u konkretnim situacijama i primjerima vrlo teška, odnosno gotovo nemoguća. Ujedno, dvije najviše razine; sinteza i vrednovanje, koje se trebaju doseći za potpuno postizanje cilja, postavljaju kao uvjete savladane sve niže razine. Kao dokaz tvrdnji

možemo uzeti analitičku utemeljenost, opisanu u prethodnom poglavlju, koja zahtijeva da osoba prosudi korisnost određenog zanimanja obzirom na trenutačnu ponudu i potražnju, no bez znanja i razumijevanja o ponudi i potražnji te analize i sinteze istog, veoma je teško napraviti kvalitetno vrednovanje.

Svaka razina Bloomove taksonomije povezana je s ključnim glagolima. Glagoli omogućuju definiranje ciljeva odnosno ishoda učenja na osnovu kojih polaznici mogu pokazati usvojena znanja i vještine. Odabir pravilnog glagola određene razine omogućuje kreatoru *E-learning* sadržaja jasno definiranje polazniku što se od njega očekuje da će savladati kroz određenu temu te osigurava dolazak do cilja.

Ako uzmemo za primjer temu koja obrađuje upotrebu HKO portala za dobivanje informacija o profilu sektora, koja za cilj postavlja dokazivanje sposobnosti korištenja tog portala, tada tema u kojoj se polaznika traži da zna koji profili sektora postoje i koji se podaci tamo mogu pronaći, što je razina 1 (znanje – prepoznavanje informacija) prema Bloomovoj taksonomiji, ne bi u potpunosti dovela do ispunjenja zadanog cilja.

U slučaju da bi se, na primjer, polazniku pokazao postupak kako se odabire neki sektor te kako se dolazi do detaljnijih, potrebnih podataka uz dodatnu vježbu u kojoj se od polaznika očekuje da ponovi radnju za neki drugi sektor, što je razina 3 (primjena – primjena znanja u rješavanju problema), polaznik to ne bi znao napraviti u konkretnoj situaciji bez savladane razine 1 i 2 (znanje i razumijevanje profila sektora).

Znanja koji svi profili sektora postoje i kako se određuje pravilni profil sektora te koji su to podaci koji se tamo mogu pronaći relevantni su preduvjet za savladavanje vještine pravilne upotrebe HKO portala. Ovim primjerom pokazuje se važnost pravilne upotrebe taksonomije.

B. Instrukcijski dizajn

Svakako, svoju važnost u kvaliteti sadržaja imaju i nastavni događaji koje je potrebno pravilno definirati za svaku cjelinu koja se obrađuje. Instrukcijski dizajn Roberta Gangea (1916.-2002.) postavlja teoriju da svaki cilj zahtijeva drugačiju vrstu nastave te predlaže 9 nastavnih događaja [2] koji, ako se prate, mogu osigurati ujednačenost i kvalitetu sadržaja.

Skretanje pažnje, kao prvi nastavni događaj Gangeovog instrukcijskog dizajna, važan je kako bi se privukla pažnja polaznika i kako bi se polaznik usredotočio na ono što će učiti. Skretanje pažnje, ovisno o cjelini i temi, može biti demonstracija mogućnosti, zanimljiva priča ili jednostavno očito pogrešan postupak (za koji će se kasnije pokazati pravilan način) ili u krajnjem slučaju samo naglašavanje značaja gradiva.

Drugi nastavni događaj je upoznavanje s ciljem koje pomaže polazniku da organizira svoje misli i usredotoči se na ono što će vidjeti, čuti i činiti.

Treći nastavni događaj vezan je za prisjećanje na prethodno naučeno. Taj događaj ima svoju posebnu važnost ukoliko se instrukcijski dizajn kombinira s Bloomovom taksonomijom jer upravo u tom koraku se osigurava nadogradnja znanja i vještina.

Četvrti nastavni događaj vezan je za aktivno sudjelovanje. Edukativna aktivnost ne samo da mora biti prikladna postavljenom cilju već mora i angažirati polaznika. Da bi se poboljšale vještine polaznika važno je da on aktivno sudjeluje u učenju kroz aktivnosti koje treba povremeno napraviti. Ujedno, aktivnim sudjelovanjem u kojem polaznik treba ponoviti ili odraditi određene radnje možemo provjeriti jesu li zadovoljene određene razine taksonomije, a time i postavljenih ciljeva nastavnih događaja, te je moguć prijelaz na višu razinu. Povratna informacija koju tim putem dobivamo daje nam mogućnost da bolje i efikasnije usmjerimo učenje.

Davanje referenci za učenje kao peti nastavni događaj olakšava pamćenje informacija jer se upravo te reference pozivaju na poznate činjenice.

Šesti nastavni događaj daje važnost prezentiranju informacija koje se prenose, a koje je potrebno rascjepkati kako ne bi prenatrpali kratkoročno pamćenje polaznika i na taj način ugrozili učenje. Opće je poznato da ljudi uče na različite načine, stoga je važno da se informacije koje se prenose prezentiraju korištenjem različitih kanala i načina.

Povratna informacija i provjera naučenog kao sedmi i osmi nastavni sadržaj omogućuju analizu polaznikove uspješnosti. Kroz provjeru naučenog, odnosno kratko testiranje, možemo jednostavno utvrditi je li cilj zaista savladan.

Deveti nastavni događaj koji navodi instrukcijski dizajn jest transfer znanja. Transfer znanja kod ovakve vrste sadržaja od iznimne je važnosti jer su smjernice za izradu standarda zanimanja, a koje su podloga za *E-learning* sadržaj, pisane veoma općenito s obzirom na to da imaju za cilj biti podloga za standarde zanimanja iz različitih područja. Pravilnim informiranjem polaznika o sličnim situacijama, specifičnim stručnim savjetima, primjerima i mogućnostima primjene naučenog gradiva, daje se dodatna vrijednost cjelokupnom učenju.

Upotreba primjera koji pokrivaju različita područja, također je veoma važna jer kod polaznika pobuđuje interes za sadržaj zato što ga može smjestiti u realnu situaciju i predočiti si konkretnu upotrebu, a samim time neće smatrati da je predočen sadržaj nepotreban.

IV. PREGLED *E-LEARNING* SADRŽAJA ZA EDUKACIJU O IZRADI STANDARDA ZANIMANJA

E - learning sadržaj u svojoj osnovi sastoji se od nastavnih jedinica, koje se opet sastoje od poglavlja, a unutar njih su teme kao najmanje jedinice i kao takav sasvim odgovara svim zahtjevima ovakve vrste edukacije.

Kroz teme kao najmanje edukativne elemente postavlja se cilj učenja u obliku izjave koja opisuje nova znanja i vještine koje će kandidat steći. Nastavni objekt kojim se kandidata poučava i na kraju, ali veoma važna, provjera naučenih znanja i vještina potvrđuju da je polaznik postigao zadani cilj.

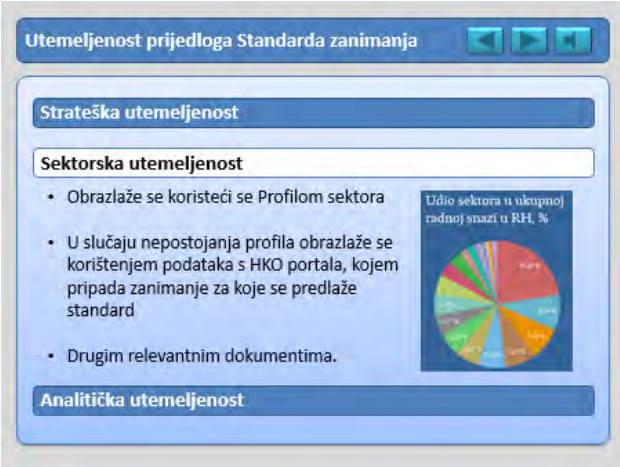
A. *Organizacija sadržaja*

Planirani *E- learning* sadržaj za izradu standarda zanimanja, koji se izrađuje kroz projekt „Potpora radu HKO sektorskih vijeća i ostalih dionika u procesu provedbe HKO-a“ Ministarstva znanosti, obrazovanja i sporta, podijeljen je u 3 nastavne jedinice koje se sastoje od 19 cjelina, a svaka cjelina od određenog broja tema.
Broj nastavnih jedinica i cjelina usklađen je s ključnim dijelovima smjernica za izradu standarda zanimanja.

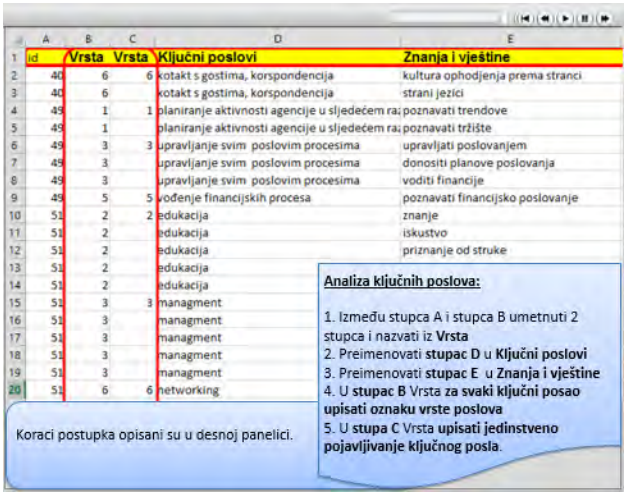
Svaka nastavna jedinica ima jasno popisan cilj. Prva nastavna jedinica naziva Hrvatski kvalifikacijski okvir upoznaje korisnike s teorijskim okvirom HKO-a te daje uvid u osnovne pojmove nužne za razumijevanje koncepta HKO-a. Cilj ovog poglavlja je postizanje razine 1 (znanje) i razine 2 (razumijevanje) Bloomove taksonomije zbog čega sam *E-learning* sadržaj započinje na ovaj način uređenom nastavnom jedinicom. Na temelju stečenih znanja mogu se graditi više razine, a već druga nastavna jedinica, koja u svom uvodnom dijelu također ima sadržan teoretski uvod razine 1 (znanje) i 2 (razumijevanje), u svojoj osnovi upravo traži od polaznika sve ostale više razine. Cilj druge nastavne jedinice je edukacija polaznika o metodi ispunjavanja A dijela obrasca prijedloga standarda zanimanja odnosno način dokazivanja utemeljenosti prijedloga standarda zanimanja.. Cilj treće nastavne jedinice je ispunjavanje B dijela istog obrasca koji daje popis ključnih poslova i skupova kompetencija standarda zanimanja i većinom od polaznika traži razinu 3 (primjena) i razinu 4 (analizu).

Zbog zahtjevnosti cjelokupne materije E- learning sadržaj kroz sva poglavlja organiziran je tako da daje kratak teoretski uvod (razine 1 i 2) nakon kojeg slijedi svladavanje viših razina, a naposljetku i provjera znanja. Kroz sve oblike pokušava se maksimalno angažirati polaznika da sudjeluje u učenju.

Teoretski dio *E- learning* sadržaja svodi se na prikaz pojmova i definicija koje bi polaznik trebao upamtiti. Informacije i pojmovi sustavno su i logički jasno organizirani te je, radi boljeg razumijevanja, svaki od pojmova dodatno opisan primjerima, a gdje je to prikladno, grafički popraćen slikom koja omogućuje vizualno povezivanje. Primjer jednog takvog teorijskog sadržaja prikazan je na slici 1.



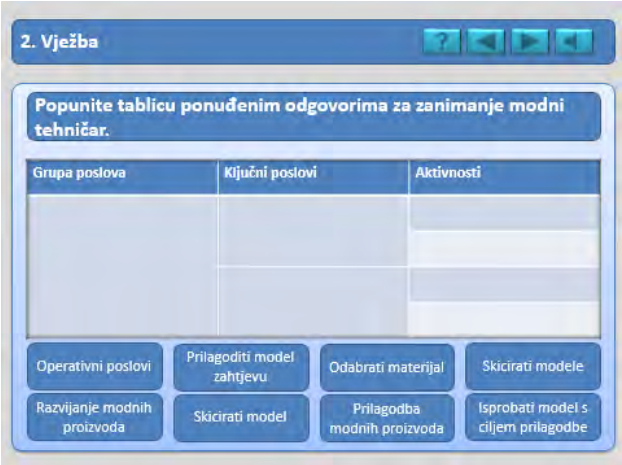
Slika 1: Primjer teoretskog sadržaja



Slika 2: Primjer praktičnog sadržaja

Na slici 2. prikazan je sadržaj kojem je cilj savladavanje razine 3 (primjena), konkretno u ovom primjernu za cilj se postavlja da polaznik zna napraviti analizu podataka dobivenih iz Anketa o standardu zanimanja (ASZ) odnosno da stečena znanja zna primijeniti u realnoj situaciji analize svoje ankete. Da bi se omogućilo potpuno razumijevanje postupka i zatim primjena sadržaj prati korake instrukcijskog dizajna. Zbog važnosti ovog sadržaja, dodatno se skreće pažnja i upoznaje polaznika s ciljem i svrhom ove teme te prisjeća na prethodno naučeno. Prikazane informacije rascjepkane su na ključne dijelove te se polazniku prikazuju u obliku filma. Film daje prikaz koraka koje treba odraditi, a osim u tekstualnom obliku prenose se i zvučnim putem. Vremenska linija u ovakvom obliku sadržaja omogućuje interaktivnu kretnju kroz određeni vremenski period događanja što dodatno omogućuje bolje svladavanje gradiva jer se polaznik u bilo kojem trenutku može vratiti na dio koji želi ponoviti.

Kao što je rečeno, aktivnom sudjelovanju polaznika dan je veliki značaj. Dodatnim prikazom ispravnih i krivih rješenja problemskog zadatka postignuto je jasnije, zanimljivije i učinkovitije učenje. Kroz cjeline i teme, osim što se od polaznika traži da ponovi neku radnju, provjerava se i naučeno te se na taj način korisnika dodatno potiče na aktivno sudjelovanje. Na kraju svake nastavne jedinice nalazi se skupina zadataka u obliku vježbi (slika 3.) kojima se provjerava jesu li zadovoljeni ciljevi postavljeni na početku nastavne jedince. Konkretno, u primjeru prikazanom na slici 3., od polaznika se zahtijeva da pokaže znanje i razumijevanje pojmova kroz primjenu u realnoj situaciji. Sve vježbe na kraju nastavnih jedinica osmišljene su tako da objedinjuju više razina usvojenosti znanja. To ih čini kompleksnijim, a ujedno i vrlo dobrim pokazateljem polaznikovog usvajanja svih potrebnih razina te potvrdom da će ih znati primijeniti u realnoj situaciji. Cilj ovog *E- learning* sadržaja nije da polaznik nauči samo neke pojmove i zasebne postupke, nego da se usvoji cjelovito znanje za pravilno ispunjavanje, jer je Standard zanimanja jedna cjelina i upravo ovakvim načinima provjera znanja to se želi provjeriti. Sustav sam bilježi napredovanje polaznika te se točnim rješavanjem zadanih vježbi, polaznicima omogućuje prelazak na sljedeće poglavlje.



Slika 3: Primjer provjere znanja

V. ZAKLJUČAK

Hrvatski kvalifikacijski okvir, između ostalog, kao cilj postavlja osiguravanje uvjeta za kvalitetno obrazovanje i učenje te stjecanje kvalifikacija koje su u skladu s potrebama gospodarstva. Kako bi on u konačnici bio ispunjen, jedan od koraka u cjelokupnom procesu je izrada kvalitetnih standarda zanimanja koji će se temeljiti na popisu svih poslova, znanja i vještina koje pojedinac treba posjedovati, a koje su u skladu s potrebama poslodavaca i zahtjevima tržišta rada. Zahtjev za upis Standarda zanimanja u podregistar može podnijeti bilo koja pravna ili fizička osoba. Kako bi se osigurala kvaliteta, ujednačenost i dosljednost zapisa, a zbog nepostojanja sastavne edukacije kroz projekt „Potpora radu HKO sektorskih vijeća i ostalih dionika u procesu provedbe HKO-a“ izrađuje *se E-learning* sadržaj kojem je cilj osposobiti svaku zainteresiranu osobu za izradu Standarda zanimanja.

Po završetku projekta *E-learning* sadržaj opisan kroz ovaj članak bit će javno dostupan svim zainteresiranim osobama. Da bi takav *E–learning* sadržaj u potpunosti postigao svoj cilj izradi sadržaja pristupilo se veoma studiozno i kao adekvatno rješenje kreatorima sadržaja postavila se Bloomova taksonomija i Gangeovih devet koraka instrukcijskog dizajna. Cjeline vezane uz određenu nastavnu jedinicu, raspoređene su tako da prate smjernice taksonomije i logični poredak aktivnosti dok sama obrada teme prati instrukcijski dizajn. Time je osigurano da nastavni sadržaj bude prikladan postavljenom cilju te da svakog polaznika, bez obzira na njegovu početnu razinu znanja, dovede do krajnjeg cilja. Dodatni naglasak u sadržaju dan je vježbama i provjerama znanja koje nemaju za cilj ispitati fragmente već cjelovitost znanja, jer se samo na taj način polazniku može dati potvrda da je sadržaj ispunio svoj cilj, da ima sve informacije te je osposobljen pristupiti izradi standarda zanimanja.

LITERATURA

[1] Bloom, B.S., (1956) Taxonomy of educational objectives, Longman, NY

[2] Gagne, R. M, Wager, W. W., Golas, K. C. & Keller, J. M. (2005). *Principles of Instructional Design, 5th ed*. Belmont, CA: Wadsworth/ Thomas Learning

[3] http://ahyco.uniri.hr/metodika/e_learning.htm#Klasifikacija , 5.2.2016

[4] *Metodologija za izradu i tumačenje profila sektora.* www.kvalifikacije.hr/fgs.axd?id=805, 6.2.2016.

[5] Ministarstvo znanosti obrazovanja i sporta (2014.). *Pravilnik o Registru HKO-a*

[6] Ministarstvo znanosti obrazovanja i sporta (2003.). *Zakon o Hrvatskom kvalifikacijskom okviru*

[7] *Priručnik za razvoj strukturnih standarza zanimanja, kvalifikacija i kurikuluma.* http://www.asoo.hr/UserDocsImages/projekti/kvalifikacije/eu%20knjige/4%20Prirucnik.pdf, 6.2.2016.

[8] *Smjernice za izradu standarda zanimanja.* http://www.kvalifikacije.hr/fgs.axd?id=922, 6.2.2016.

Uloga IKT u razvoju financijske pismenosti djece

Ivana Ružić
I. osnovna škola Čakovec,
Ulica kralja Tomislava 43, Čakovec, Hrvatska
ivana.ruzic@skole.hr

Ključne kompetencije za cjeloživotno učenje su kombinacija znanja, vještina i stavova koja odgovara kontekstu. One su posebno potrebne za osobno ispunjenje, razvoj i socijalnu uključenost aktivnog građanina. Navedene kompetencije su neophodne u društvu znanja, one jamče veću fleksibilnost omogućujući pojedincu da se brže prilagoditi stalnim promjenama u sve povezanom svijetu. One su također glavni čimbenik u inovacijama, produktivnosti i konkurentnosti, a doprinose motivaciji i zadovoljstvu pojedinca i kvaliteti rada. Ključne kompetencije trebaju steći mladi ljudi tijekom njihovog obveznog obrazovanja i osposobljavanja, pripremanja za život odraslih, a odrasli ljudi tijekom svog života, kroz proces izgradnje i usavršavanje vještina. Ovaj rad prikazuje razvoj digitalnih kompetencija i financijske pismenosti učenika u osnovnom obrazovanju, njihovih roditelja i učitelja kao dio međunarodnog projekta Erasmus+.

I. UVOD

U današnje doba svaka organizacija nastoji ostvariti što bolju stratešku prednost kako bi postala prepoznatljiva i uspješnije funkcionirala. U obrazovnim ustanovama u nastavu se sve više uvodi projektni način rada njegujući timski duh među učenicima i nastavnicima te priprema učenika za budućnost tj. kako bi učenici ostvarili što veću konkurentnost na tržištu rada te se adekvatno pripremili za buduće karijere i vlastiti poslovni i osobni razvoj, razvili znanja, vještine i sposobnosti za stabilnu i perspektivnu budućnosti, te temelje za cjeloživotno učenje.

Keep Invest Donate Spend je Erasmus+ projekt iz područja međunarodna školska partnerstva sa ciljem razvoja i osnaživanja suradnje, inovacija i razmjene iskustava među partnerima, financiran sredstvima Europske komisije. U projektu sudjeluje pet osnovnih škola iz Turske, Portugala, Rumunjske, Poljske i Hrvatske uz snažnu potporu ministarstva obrazovanja u Turskoj i Narodnu banku Rumunjske, sa ciljem da se razmjene primjeri dobre prakse i zajednički razvijaja niz obrazovnih sadržaja za financijsko opismenjavanje učenika, njihovih roditelja i učitelja.

Projektne aktivnosti prilagođene su dobnim skupinama djece, obuhvaćaju razvoj kreativnosti, suradnje, komunikacijskih vještina, timskog rada, kritičkog i problemskog načina razmišljanja uz podršku IKT kao temeljnog alata u razvoju financijske pismenosti, digitalnih kompetencija i kompetencija za cjeloživotno učenje.

Kako bi osigurali ostvarivanje projektnih ciljeva, projektne aktivnosti obuhvaćaju i edukaciju učitelja da bi oni bili osposobljeni za edukaciju o financijskoj pismenosti djece i aktivnosti za roditelje kako bi roditelji kao primarni edukatori i partneri u obrazovanju razvili svijest o važnosti razvoja financijske pismenosti njihove djece.

II. FINANCIJSKA PISMENOST DJECE

Navika čuvanja i spremanja dobara, tj. štednje je vrlo važan dio našeg života, poznata je kao financijska disciplina ili financijska pismenost. Financijska pismenost je sposobnost korištenja znanja i vještina kako bi učinkovito donosili odluke o upravljanju novcem.[1] Osobna financijska pismenost obuhvaća niz tema poput: što je novac, povijest novca, što su financijske institucije i sl., do balansiranja između vlastite potrošnje i dugoročnog planiranja za odlazak u mirovinu. Opća pismenost je temeljni dio obrazovnog sustava, dok je financijska pismenosti često izostavljena u obrazovanju.

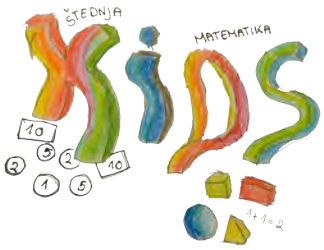
Iako se financijska pismenost nastoji uvesti u obrazovne institucije, roditelji i staratelji su primarni edukatori kada je u pitanju podučavanje djece vještinama potrebnim za razvoj snažnih temelja za cjeloživotnu financijsku pismenost.

Poput drugih provokativnih tema, novac je nešto o čemu će djeca čuti izvan kuće - u školi, na sportskim aktivnostima ili u domovima prijatelja. Djeca mogu dobiti pogrešnu poruku o novcu ako informacije dobivaju od svojih vršnjaka. Važno je djecu u ranoj dobi naučiti da bogatstvo nije izvor zadovoljstva i sreće. Čak ako roditelji ne znaju razliku između naknade za rad i doprinosa mirovinskim stupovima, roditelji mogu dati točne informacije, predstaviti ideje, razvijati interes i svijesti, pomoći i osnažiti svoju djecu da preuzmu kontrolu nad svojim financijskim životom.

Promatramo li ranije generacije, različiti oblici štednje su vrlo usko povezani s njihovom svakodnevnom životom. Za prošle generacije štednja je bila najvažniji aspekt života, a omogućila im je sigurnust u starosti. Međutim, s modernom ekonomijom i globalizacijom, mlade generacije su prisiljene trošiti više novaca u svakodnevnom životu, često trošimo iznad svojih mogućnosti, ne možemo se kontrolirati u okolini koja promiće raskošan život.

Ako na vrijeme ne razvijemo i usvjojimo naviku štednje i racionalnog korištenja svih oblika i vrsta dobara kojima raspolažemo, tu naviku nećemo steći ni u kasnijoj fazi svog života. Sve što naučimo i usvojimo u djetinjstvu, prati nas kroz cijeli život. Iz tog razloga bismo trebali djecu učiti o važnosti štednje, poticati i razvijati navike racionalnog korištenja dobara. Financijska disciplina i neovisnost je jedan od najvažnijih područja koje zajedno roditelji i učitelji moraju razviti kod djece.

Na žalost, mnoga istraživanja pokazuju da je financijska pismenost građana veoma niska čak i u razvijenim zemljama.



Slika 1. Logo Erasmus+ projekta Keep Invest Donate Spend

Keep Invest Donate Spend je međunarodni projekt koji ima cilj razvijati financijsku pismenost djece, podizanjem svijesti o važnosti vlastitog ponašanja i raspolaganja dobrima, omogućuje djeci da stječu naviku štednje, osmisle, predlažu i provode kampanju podizanja svijesti u svojoj školi, susjedstvu i široj zajednici, koristeći crteže, plakate, računalne igre i obrazovne programe, video i druge sadržaje i aktivnosti kroz koje će prenijeti snažne poruke na temu financijske pismenosti. Ovaj projekt uključuje istraživanje i ispitivanje razine financijske pismenosti učenika, učitelja i roditelja na početku i kraju projekta, prate ga obrazovni materijali prilagođeni za djecu i mlade prema dobnim skupinama, obrazovni materijali za učitelje i roditelje.

III. IKT KAO PODRŠKA U RAZVOJU FINANCIJSKE PISMENOSTI

Učenici su uz potporu učitelja, samostalno vodeći se vlastitim interesima i sklonostima osmislili i proveli niz aktivnosti za ostvarenje postavljenih ciljeva projekta Erasmus+ Keep Invest Donate Spend koristeći IKT.

IKT nastoji se smisleno i sa svrhom koristiti u svakoj projektnoj aktivnosti, razvijajući digitalne kompetencije učenika. Dok digitalne tehnologije postupno obuhvaćaju sve dijelove svakodnevnog života, 60% učenika nikada ne koriste digitalnu opremu u svom razredu. Puni potencijal za poboljšanje obrazovanja kroz IKT-a u Europi i dalje tek treba otkriti, a to je razlog zašto je Europska komisija razvija politiku i potporu istraživanja kako bi učenici odgovaraju za život u 21. stoljeću i rad.[2]

Aktivnosti učenika:

- izrada vlastitog loga programa,

- korištenje online alata Socrative za istraživanje i ispitivanje sudionika na početku i kraju projekta,

- izrada proširene stvarnosti uz alata Aurasma, a može se koristiti na pametnim telefonima i tabletima koje pokreće iOS,

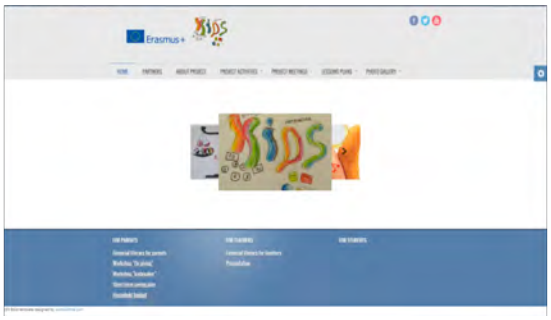
- izrada obrazovne računalne igre uz korištenje robota Apple Sphero,

- oblikovanje 3D modela vlastitih novčića i ispis na 3D pisaću,



Slika 3. 3D ispis novčića

- izrada mrežnih stranica projekta koristeći CMS sustav Joomla,



Slika 3. Prikaz mrežnih stranica projekta

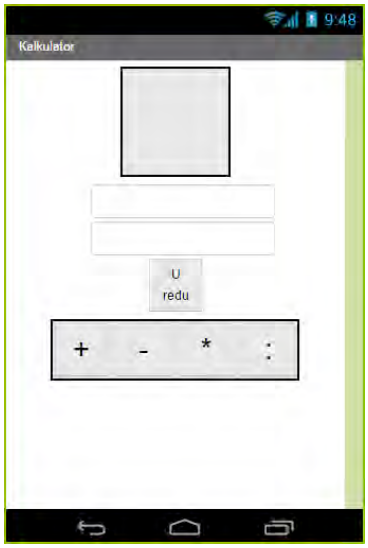
- izrada obrazovnih računalnih igara koristeći alate Game Maker, Blender i Construct, a koje se mogu koristi na osobnim računalima, tabletima i pametnim telefonima na operacijskim sustavima Windows, Android i iOS,



Slika 4. Prikaz računalne igre na tablet uređaju

- izrada kalkulatora sa osnovnim matematičkim operacijama u alatu Visual Basic 2015., a koji se može koristi na osobnim računalima, tabletima i pametnim

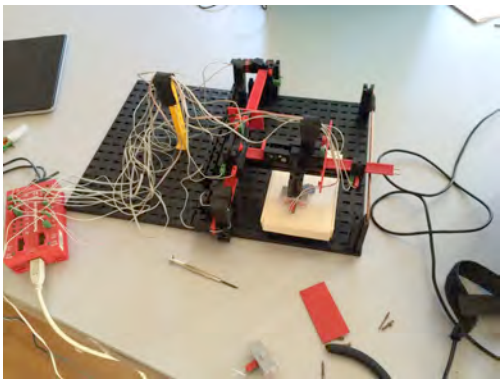
telefonima na operacijskim sustavima Windows, Android i iOS,



Slika 5. Prikaz kalkulatora na tablet uređaju

- izrada kalkulatora koji omogućuje konverziju željenog iznosa u različite valute, kalkulator je izrađen u alatu Visual Basic 2015., tečajna lista se automatski ažurira sa tečajem HNB, a može se koristiti na osobnim računalima, tabletima i pametnim telefonima na operacijskim sustavima Windows, Android i iOS,

- izrada robota koji prepoznaje i odvaja novčiće iz hrpe različitih predmeta, robotski program izrađen je u alatu RoboPro,



Slika 6. Robot koji prepoznaje i odvaja novčiće

- vođenje dnevnika štednje u online alatu Bankaroo – virtualnoj banci za djecu koja je dostupna na svakom uređaju spojenom na Internet, omogućuje djeci učenje o pametnom raspolaganju novcem, upravljanju prihodima i rashodima, postavljanje ciljeva štednje i potporu nagrađivanja različitim značkama[3],

- izrada elektroničke knjige, videa, kvizova, interaktivnih postera i slikovnica, animacija i drugih multimedijalnih obrazovnih sadržaja.

U navedenim aktivnostima IKT je dodatno motivirala učenike za uključivanje u projektne aktivnosti s obzirom da je učenicima ponuđeno istraživanje, korištenje i razvoj vlastitih sadržaja koristeći suvremenu tehnologiju.

Razvoj digitalnih vještina postaje dio obrazovanja o financijskoj pismenosti kroz:

- traženje, upravljanje i preuzimanje podataka i informacija,

- ostvarivanje međusobne komunikacije i suradnje,

- problemsko promatranje i rješavanje zadataka,

- korištenje gotovih digitalnih alata,

- stvaranje i objavu vlastitih digitalnih sadržaja.

Na taj način učenje postaje usmjereno prema učeniku, njegovim interesima i sposobnostima. Učenjem razvijamo kreativnost, inovativnost učenika, suradničko učenje, pozitivan odnos prema radu i učenju.

Posebna pažnja posvećena je programiranju kao pismenosti 21. stoljeća. Programiranje pojačava kreativnost, uči ljude da surađuju, da rade zajedno bez obzira na fizička ograničenja i zemljopisne granice, ono je univerzalni jezik koji nam omogućuje komunikaciju. Tehnologija je dio našeg života, a programiranje nas povezuje i zbližuje![4]

Navedene aktivnosti provodile su se kao niz radionica na nastavi i u sklopu izvannastavnih aktivnosti te kao radionice u vrijeme odmora za učenike. U dijelu radionica sudjelovali su i članovi Društva distrofičara i cerebralne paralize iz Čakovca sa ciljem aktivnog sudjelovanja u projektnim aktivnostima svih članova lokalne zajednice, razvijanja empatije i suradnje kod učenika škole. Humanitarnom prodajom radova, prikupljena su sredstva koja su donirana socijalnoj samoposluzi u Vukovaru te kao pomoć za liječenje učenika naše škole koji boluju od teško izliječivih i rijetkih bolest. Projekt je predstavljen i na državnoj Smotri projekata iz građanskog odgoja i obrazovanja, na različitim državnim i međunarodnim seminarima, konferencijama, edukacijama i sličnim događajima. Potaknuo nas je i na uključivanje u projekt „Young Entrepreneurs in Croatia" kojem je cilj razvoj pet glavnih kompetencija: proaktivno učenje, poduzetnički stav, emocionalna inteligencija, društvena odgovornost i globalno razmišljanje.

Projekte aktivnosti završavaju krajem lipnja 2016. godine i tada će se provesti završno istraživanje o razini financijske pismenosti učenika, učitelja i roditelja. Do sada je uočeno je da su svi sudionici projekta socijalno osjetljiviji i rado se uključuju u različite humanitarne aktivnosti, velika pažnja se pridaje recikliranju i stvaranju novih vrijednosti, postavljanjem jasnih ciljeva i vođenjem dnevnika štednje učenici su motiviraniji za štednju i racionalno trošenje džeparca.

IV. ZAKLJUČAK

Međunarodni projekti obogaćuju cjelokupan nastavni proces te omogućuju razvoj obrazovnih sadržaja, razmjenu iskustava i primjera dobre prakse među partnerima. Odabirom projekte teme, ciljeva, sadržaja i

aktivnosti, omogućuju upotpunjivanje školskih kurikuluma, nastavnih planova i programa temama koje koje budućim generacijama doprinose cjelokupan razvoj u profesionalnom i privatnom smislu. IKT kod učenika dodatno razvija interes, a omogućuje uvođenje inovativnosti u obrazovnom procesu.

LITERATURA

[1] http://www.investopedia.com/university/teaching-financial-literacy-kids/

[2] https://ec.europa.eu/digital-agenda/en/skills-jobs

[3] https://www.bankaroo.com

[4] https://ec.europa.eu/digital-agenda/en/coding-21st-century-skill

Informacijsko-komunikacijske znanosti u nastavi - digitalizirani materijali za učenje

Tihana Babić, Ana Ogrin i Manuela Babić*
Visoko učilište Algebra – visoka škola za primijenjeno računarstvo, Zagreb, Hrvatska
*Knjižnice grada Zagreba, Zagreb, Hrvatska
tihana.babic@algebra.hr, ana.ogrin@algebra.hr, manuela.babic@gmail.com

Sažetak - U suvremenom multimedijalnom vremenu događa se pomak od učenja kroz klasične tiskane izvore kao što su knjige i članci, prema novim oblicima učenja putem digitaliziranih materijala. Digitalizirani nastavni materijali definiraju se kao informacije potrebne za učenje i poučavanje, a koje su dostupne u digitalnom obliku i koriste se na računalu ili nekom sličnom uređaju. Studenti Visokog učilišta Algebra na kolegijima imaju obveznu literaturu u klasičnoj tiskanoj formi priručnika ili knjiga koje posuđuju iz knjižnice. Također, studentima su dostupni nastavni materijali preko sustava Digitalne referade – Infoeduke, kao što su prezentacije s predavanja, zadaci za vježbu, upute za uspješno svladavanje gradiva te dodatna literatura. Očekuje se još veći pomak od klasičnih oblika učenja prema različitim oblicima e-učenja, pa je istraživanje imalo je za svrhu istražiti stavove studenata o korisnosti digitaliziranih materijala za učenje u usporedbi s klasičnom, fizičkom literaturom. Istraživanjem je obuhvaćen uzorak od 52 studenata upisanih na preddiplomski i diplomski studij Digitalnog marketinga u akademskoj godini 2015./2016. Cilj rada je prikazati prikupljene podatke istraživanja, a kao zaključak rada prikazane su prednosti i nedostaci klasičnih oblika učenja korištenjem tradicionalnih izvora, kao i prednosti i nedostaci e-učenja korištenjem digitaliziranih izvora informacija i znanja.

I. UVOD

Knjige su ljudima već stoljećima izvor informacija, a svoj tradicionalni oblik koji se sastoji od korica i knjižnog bloka dobile su nakon 1440. godine kada je Johannes Gutenberg izumio prvi tiskarski stroj. Razvojem društva i tehnologije, a osobito informacijsko-komunikacijskih tehnologija, uvode se inovacije i u sektoru knjiga, pa je tako knjiga dobila svoj elektronički odnosno tzv. e-oblik. Pojava ove vrste digitalnog materijala označila je pomak prema novim oblicima učenja.

Projekti masovne digitalizacije knjiga započeli su 1971. godine Projektom Gutenberg. Osnivač tog projekta bio je tadašnji student Sveučilišta Illinois Michael Stern Hart koji je najvećom vrijednošću svog računala smatrao upravo njegovu sposobnost slobodnog pristupa, pronalaženja, pretraživanja i pohrane onoga što se nalazi u knjižnicama i koji je izumio prvu e-knjigu [1]. Dakle, u toj kratkoj, ali burnoj povijesti pojavljivanja masovne digitalizacije knjiga, odnosno uz uporabu napredne i suvremene tehnologije te uz velika ekonomska ulaganja digitaliziraju se materijalne tiskane knjige [1] Svaka digitalizirana tiskana knjiga smatra elektroničkom ili digitalnom knjigom, skraćenog naziva e-knjiga. No, naziv digitalni materijali podrazumijeva i filmove, glazbu, crteže, grafikone, pa tako digitalni nastavi materijali mogu biti i prezentacije, interaktivni multimedijalni plakati, digitalni stripovi, web priručnici, interaktivne animacije, online provjere znanja i sl. [2].

Studenti Visokog učilišta Algebra – visoke škole za primijenjeno računarstvo na kolegijima imaju obveznu literaturu u klasičnoj tiskanoj formi priručnika ili knjiga koje posuđuju iz knjižnice. Također, studentima su dostupni digitalni nastavni materijali preko sustava Digitalne referade – Infoeduke, kao što su prezentacije s predavanja, zadaci za vježbu, upute za uspješno svladavanje gradiva te dodatna literatura.

Istraživanje koje smo proveli na uzroku studenata Visokog učilišta Algebra - visoke škole za primijenjeno računarstvo imalo je za svrhu istražiti stavove studenata o korisnosti digitalnih materijala za učenje u usporedbi s klasičnom, fizičkom literaturom.

II. METODOLOGIJA ISTRAŽIVANJA KORIŠTENJA DIGITALNIH NASTAVNIH MATERIJALA U USPOREDBI S KLASIČNOM LITERATUROM ZA UČENJE

A. Sudionici istraživanja

U istraživanju koje je provedeno na Visokom učilištu Algebra - visokoj školi za primijenjeno računarstvo obuhvaćen je uzorak od 52 studenata. Sudjelovao je ukupno 51 student 1. godine preddiplomskog stručnog studija Digitalnog marketinga te 1. godine diplomskog studija Primijenjenog računarstva, a 1 anketni listić izuzet je iz rezultata istraživanja obzirom da se odgovori studenta/ice mogu definirati kao nevezani za cilj istraživanja i prema tome nevažeći.

U Tablici I. prikazana je struktura sudionika s obzirom na studij i spol sudionika.

TABLICA I. STRUKTURA SUDIONIKA

Naziv studija	Primijenjeno računarstvo	Digitalni marketing
Broj sudionika po studiju	20	31
Broj sudionika po spolu	19 – M	19 – M
	1 – Ž	12 – Ž
Sveukupno sudionika (N)	51	

B. Istraživački instrumenti

U postupku anketiranja ispitanicima su podijeljeni anketni listići s pitanjima koja su imala ponuđen višestruki izbor unaprijed određenih odgovora, no postojala je i mogućnost odabira odgovora otvorenog tipa bez unaprijed ponuđene opcije. Nakon što je sudionicima bila objašnjena svrha ispitivanja, studenti su zamoljeni da iskreno odgovore na pitanja. Dobiveni podaci obrađeni su kvantitativnom i kvalitativnom analizom odgovora ispitanika.

Anketni upitnik sastojao se od sedam pitanja od čega su tri pitanja bila zatvorenog tipa, a četiri pitanja bila su kombinacija zatvorenog tipa i otvorenog tipa, odnosno na četiri pitanja sudionici su mogli birati jedan ili više unaprijed ponuđenih odgovora i/li dati svoj vlastiti alternativni odgovor. Na ostala tri pitanja unaprijed su bili ponuđene tri odnosno četiri mogućnosti odabira odgovora, a ispitanici su se mogli odlučiti za odabir više opcija.

C. Rezultati istraživanja

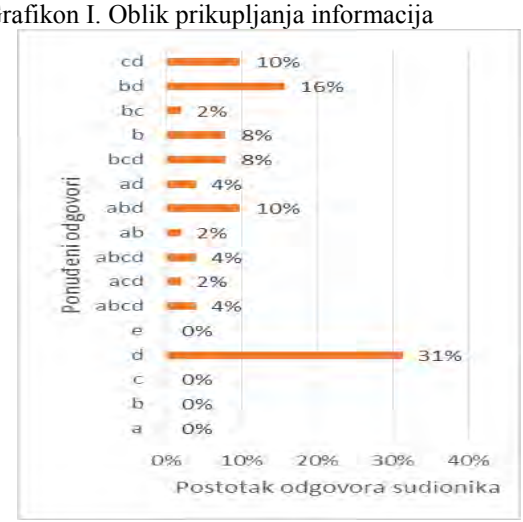
a) Oblik prikupljanja informacija

Studentima je postavljeno pitanje koji oblik prikupljanja informacija najčešće koriste prilikom izvršavanja akademskih zadataka (pisanje seminara, spremanje ispita, pripremanje prezentacija). U upitniku su na to pitanje unaprijed ponuđena četiri odgovora te dodatna kategorija „drugo“ koja je bila koncipirana kao otvoreno polje za unos konkretnog odgovora. Bilo je ponuđeno pet kategorija: a) tiskane izvore (knjige, članci u časopisima); b) digitalne/elektroničke izvore; c) specijalizirane Internetske portale; d) Internetske pretraživače; e) drugo. Sudionici su mogli birati više odgovora: odabrati jednu, dvije, tri, četiri, pet ili neku od kombinacija ponuđenih opcija.

Za četiri od pet osnovnih pitanja sudionici se nisu odlučili kao samostalan izbor već su odabirali višestruke odgovore čije su kombinacije prikazane navedenim grafikonom. Grafikon broj I. prikazuje postotak odgovora sudionika.

Grafikonom I. prikazano je šesnaest kombinacija odgovora od kojih je pet osnovnih pitanja.

Grafikon I. Oblik prikupljanja informacija



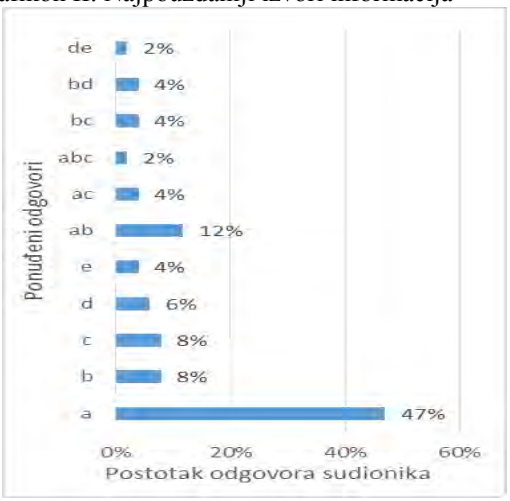
Iz odgovora ispitanika vidljivo je da tiskane izvore za prikupljanje informacija kao što su knjige, časopisi ili bilo koji drugi oblik „klasične literature“ studenti najrjeđe koriste, odnosno koriste u kombinaciji s drugim izvorima navedenim u upitniku. U prikupljanju informacija studenti se češće koriste digitalnim materijalima do kojih dolaze preko digitaliziranih knjiga ili časopisa, specijaliziranih portala te Internetskim pretraživačima općenito kako bi pronašli informaciju koju žele. Na pitanje otvorenog tipa nije ponuđen niti jedan odgovor, odnosno studenti ne navode druge oblike prikupljanja informacija. Zanimljivo je da se zapravo 31% ispitanih studenata prvotno okreće Internetskim pretraživačima kao jedinom izvoru prikupljanja informacija prilikom izvršavanja svojih akademskih zadataka što pokazuje znatan odmak od prakse korištenja obvezene i dodatne literature u tiskanom obliku.

b) Najpouzdaniji izvori informacija

Sudionicima ispitivanja postavljeno je pitanje koje izvore informacija smatraju najpouzdanijima. Sudionicima je ponuđeno pet identičnih odgovora kao i na pitanje koje izvore koriste najčešće. Rezultat je zanimljiv jer pokazuje kako studenti najviše povjerenja imaju u tiskane izvore, odnosno njih 47% odgovara kako ih smatra najpouzdanijim izvorima prikupljanja informacija. Za 8% ispitanika najrelevantniji izvori prikupljanja informacija su digitalni/elektronički izvori (digitalne knjige, časopisi, članci). 8% sudionika ispitivanja misli da su to specijalizirani portali dok 4% ispitanika smatra da je to nešto drugo iako ne navode izrijekom druge izvore. Samo njih 6% Internetske pretraživače smatra najpouzdanijim izvorima, a istovremeno njih 31% iste njih koristi kao prvotni oblik prikupljanja informacija. Također, sudionici ispitivanja višestrukim su odabirom dva ili tri ponuđena odgovora također prednost dali tiskanim izvorima u kombinaciji s digitalnim pa tako njih 12% smatra da pouzdane informacije može dobiti iz oba izvora.

Grafikonom broj II. prikazani su odgovori sudionika prema postotnoj frekvenciji odabranog odgovora.

Grafikon II. Najpouzdaniji izvori informacija



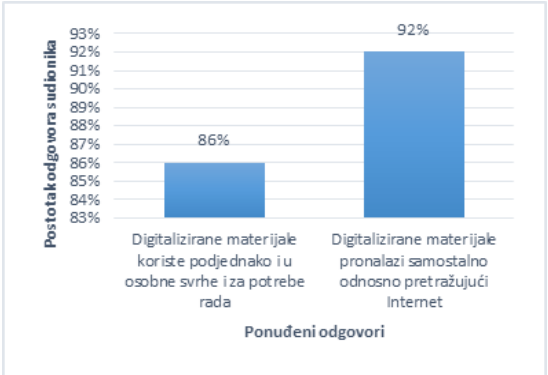
c) Upotreba i pronalazak digitaliziranih materijala

Sudionicima su postavljena dva pitanja kojima se tražilo odgovor na pitanje za što se konkretno koriste digitaliziranim materijalima te gdje te materijale nalaze. Za potrebe rada 6% sudionika istraživanja koristi digitalne odnosno elektroničke izvore, dok njih 3% iste izvore koristi za osobne potrebe. Većina ispitanika, njih 86% odgovorilo je da digitalne materijale koriste podjednako; kako u osobne svrhe tako i za potrebe rada.

Na pitanje gdje pronalaze najviše digitalne/elektroničke građe potrebne za rad, 8% ispitanika navodi visokoškolsku knjižnicu, konkretno sustav digitalne referade. 4% ispitanika navodi gradsku knjižnicu, dok je uvjerljivo najveći postotak sudionika, odnosno njih 92% ustvrdilo da digitalne materijale potrebne za (prethodnim pitanjem utvrđene svrhe) rad ili osobne potrebe, pronalazi samostalno odnosno pretražujući Internet. Grafikon III. prikazuje upotrebu digitalnih materijala.

Grafikonom broj III. su prikazani usporedni visoki postotci odgovora sudionika istraživanja o upotrebi digitalnih materijala odnosno visokoj razini samostalnosti prilikom upotrebe digitalnih materijala bilo u osobne svrhe bilo u svrhe rada. Sudionici ispitivanja odgovorima na ova dva pitanja pokazuju visoku razinu informatičke pismenosti i istovremeno učestalu upotrebu računala u svom obrazovanju kao i svakodnevnom životu.

Grafikon III. Upotreba digitalnih materijala



d) Digitalna građa na visokoškolskoj ustanovi

TABLICA II. DIGITALNA GRAĐA NA VISOKOŠKOLSKOJ USTANOVI

Pitanje/Odgovori	Da	Ne	Ne znam	Nešto drugo
Jeste li na studiju bili upoznati s referentnim digitalnim izvorima i kako se njima služiti?	31%	41%	28%	0%
Smatrate li da bi Vam pomoglo u studiju kada biste imali pristup većem broju digitalizirane građe na Vašem fakultetu?	72%	8%	16%	4%

Sudionicima istraživanja postavljena su dva pitanja o izvorima digitalne građe na Visokom učilištu Algebra – visokoj školi za primijenjeno računarstvo. 31% sudionika ispitivanja odgovara da je adekvatno upoznato s referentnim digitalnim izvorima te kako se njima služiti tijekom studija, dok njih 41% smatra da nije. 28% ispitanika ne zna odgovor na ovo pitanje.

Na pitanje „Smatrate li da bi Vam pomoglo u studiju kada biste imali pristup većem broju digitalizirane građe na Vašem fakultetu?“ ponuđeno je slobodno polje za odgovore te tri opcije odgovora koje su prikazane u Tablici II. zajedno s odgovorima na prethodno postavljeno pitanje.

72% ispitanika odgovara da smatra kako bi im pomoglo u studiju imati veći pristup digitaliziranoj građi. Njih 8% smatra kako im više digitalne građe nije potrebno za studij dok njih 16% na ovo pitanje odgovara „ne znam“. 4% sudionika je pod „drugo“ navelo odgovore „možda i vjerojatno“, dakle mogli bismo njihove odgovore svrstati pod „ne znam“.

Iz ovih je rezultata vidljiva nužnost jasnijeg obavještavanja te upućivanja studenata u načine dolaženja do digitalnih materijala na studiju. Prvenstveno se ovdje misli na visokoškolsku ustanovu općenito, dok se pitanje pojedinih kolegija nije postavljalo. Naime, studenti imaju pristup digitalnim materijalima u obliku prezentacija i sažetaka tijekom kolegija koji su im dostupni kroz sustav Digitalne referade – Infoeduke, ali i pristup Digitalnom katalogu knjižnice kroz isti sustav preko kojeg mogu vidjeti svu raspoloživu literaturu za posudbu. Također, iz odgovora sudionika na drugo navedeno pitanje vidljivo je kako je potrebno povećati broj raspoloživih i dostupnih digitalnih materijala na ustanovi općenito. Obzirom da su svi podaci nominalni, može se izračunati samo mod odnosno dominantna vrijednost te izračuni statističkih mjera centralne tendencije i varijabilnosti nisu bili mogući. Dominantna vrijednost je najuočljivija upravo iz grafikona.

III. PREDNOSTI I NEDOSTACI DIGITALNIH NASPRAM TISKANIH MATERIJALA U OBRAZOVANJU

Informacije potrebne za učenje i poučavanje, a koje su dostupne u digitalnom obliku i koriste se na računalu ili nekom sličnom uređaju nazivaju se digitalni nastavni materijali.

Digitalni materijali namijenjeni su za korištenje u nastavi jednako kao i tiskani materijali, a pohranjeni mogu biti u digitalnom formatu na elektroničkom mediju, računalu ili pak na internetu. Također, važno je reći da se digitalni materijali jednako vrednuju kao i tiskani u visokoškolskom sustavu obrazovanja Republike Hrvatske stoga nastavnici sami mogu birati koji će od oblika nastavnih materijala koristiti.

Vrlo često nastavnici odabiru oba oblika kada imaju dvojbu oko odabira tiskanog ili digitalnog materijala[4]. Važno je stoga razmotriti prednosti i nedostatke digitalnih naspram tiskanih materijala.

A. DIGITALNI NASTAVNI MATERIJALI

a) Prednosti digitalnih nastavnih materijala

Jedna od glavnih prednosti digitalnih materijala naspram tiskanih je cijena iz razloga što kod digitalnih materijala nema troškova tiskanja i korištenja papira što je ujedno prednost i s ekološkog gledišta. Obzirom na način pohrane ovim se oblikom materijala štedi i fizički prostor jer se veliki broj digitalnih materijala može pohraniti na jednom elektroničkom mediju. Uz navedeno, mogućnost brzog kopiranja, kao i cijena distribuiranja, prednosti su digitalnih materijala. Osim toga, velika je prednost što je digitalni oblik pogodan za pretraživanje po ključnoj riječi što doprinosi efikasnosti pri učenju i uštedi vremena kod potrage za važnim informacijama.

Također, digitalni materijali omogućuju postavljanje linkova koji mogu voditi do nekog drugog izvora informacija. Na taj se način linearna struktura tiskanih materijala gubi jer linkovi omogućuju daljnje proširivanje učenja i istraživanja nastavnog gradiva. U odnosu na tiskane materijale, digitalni se ne istroše s vremenom pa je i to jedna od prednosti kao i jednostavnost nadopune i izmjene sadržaja.

b) Nedostaci digitalnih nastavnih materijala

Kod digitalnih nastavnih materijala kao glavni nedostatak često se javlja pitanje autorskih prava kao glavni nedostatak. Nastavnici koji izrađuju digitalne materijale ne žele da se njihovi radovi kopiraju te predstavljaju kao tuđi rad ili pak neovlašteno mijenjaju pri čemu često dolazi do krivih informacija. Pitanje autorskog prava repozitoriji rješavaju na način da uz svaki nastavni materijal u pravilu budu navedeni uvjeti korištenja. Ukoliko se informacije i nastavni materijali traže na internetu tada je potrebno proučiti uvjete korištenja te ukoliko je kopiranje dozvoljeno navesti izvor.

Uz spomenutu prednost digitalnih materijala te korištenja linkova važno je spomenuti i nedostatak istog, a to su zastarijevanje sadržaja kao i samo gašenje linkova na internetu. Kako bi se ovo izbjeglo, potrebno je redovito provjeravati aktualnost linka kao i točnost sadržaja.

Spremanje sadržaja u digitalnom obliku zahtijeva pažljivo spremanje dokumenta i rukovanje što je nedostatak, a javlja se i nekompatibilnost spremljenih oblika dokumenata sa novim softverima i hardverima. Neki od drugih nedostataka su i nepreglednost korištenja digitalnih materijala na elektroničkim medijima radi čega pak velik broj ljudi i dalje preferira tiskane materijale.

B. TISKANI NASTAVNI MATERIJALI

a) Prednosti tiskanih nastavnih materijala

Jedna od glavnih prednosti tiskanih materijala je što za njihovo korištenje nije potrebno imati elektroničke uređaje, računalo kao ni pristup izvoru električne energije i Internetu. Sukladno tome tiskani se materijali mogu koristiti na različitim mjestima, a važno je spomenuti da se mogu koristiti i kada postoji određeno oštećenje što je kod digitalnih materijala nemoguće.

Knjižnice, čitaonice i knjižare gdje se tiskani materijali koriste dobro su organizirani i kvalitetno organizirani sustavi dok se elektroničke mreže i virtualne knjižnice kao sustavi za služenje digitalnim materijalima još razvijaju te stoga još nisu jednako zastupljene i valorizirane u društvu [5].

b) Nedostaci tiskanih nastavnih materijala

Kao što je prednost tiskanih materijala da se mogu koristiti na različitim mjestima, tako je i nedostatak istih što su, kada imamo veći broj materijala, teži za prenošenje i zauzimaju više mjesta.

Nedostatak je i teži pristup fizičkoj jedinici kao što je knjiga jer se knjige mogu prestati tiskati pa ih je teže nabaviti, a u knjižnici mogu biti ili posuđene ili korištenje dopušteno s ograničenjem rada na knjižnicu pa se materijalu može pristupiti samo u radno vrijeme knjižnice.

Cijena tiskanih materijala velik je nedostatak u odnosu na digitalne, a obzirom da se koristi papir u velikim količinama javlja se pitanje ekologičnosti, kao što je navedeno. Neki od inih nedostataka su i što se tiskani materijali ne mogu preuzeti s Interneta te ih je teško kopirati.

C. INTERNET KAO ALAT U NASTAVNOM PROCESU

Internet kao alat poučavanja služi za izravno svladavanje nastavnih sadržaja. Internet pomaže u prevladavanju nedostataka 'klasičnog' prenošenja znanja jer omogućuje pristup dodatnim informacijama, a što naposljetku pridonosi uspješnom i kvalitetnom svladavanju nastavnih sadržaja.

Za prenošenje znanja putem Interneta potrebno je:

- omogućiti komunikaciju na računalima za sve sudionike uključene u nastavni proces
- napraviti redoslijed prenošenja, prezentiranja i usvajanja sadržaja

- upravljati učenjem o administrativnim podacima studenata te obrazovnom sadržaju
- znanje učvrstiti kroz testove za samoprovjeru
- omogućiti spremanje nastavnih sadržaja na serveru odnosno web poslužitelju kao i povezivanje istih te pokretanje diskusija i stvaranje testova za provjeru znanja[5].

a) Prednosti učenja putem Interneta

Jedna od glavnih prednosti ovakvog oblika učenja jest fleksibilnost odnosno sloboda u samostalnom stvaranju vremenskog rasporeda učenja. Također mogućnost dodatnog pretraživanja informacija na Internetu pomaže kod lakšeg usvajanja i proširivanja znanja. Jedna od prednosti je i mogućnost komunikacije s profesorima bez fizičkog susreta što omogućuje brže prevladavanje mogućih prepreka u svladavanju gradiva.

b) Nedostaci učenja putem Interneta

Neki od nedostataka učenja putem Interneta su različite inicijalne razine znanja u služenju elektroničkim uređajima i programima, pa tako i pretraživanja Interneta, a što može stvarati razlike u uspješnosti svladavanja gradiva kod ovog oblika učenja. Također ovakav oblik učenja zahtijeva određene uvjete kao što su pristup električnoj energiji i kvalitetnoj Internetskoj vezi čemu nije uvijek tako pa dolazi do tehničkih poteškoća ukoliko je veza spora, a također može doći i do gubitka uređenog ili mogućnosti rada ako dođe prekida veze. Važno je naglasiti i da nemaju svi pristup internetu kod kuće, a fakultet ili neka druga ustanova poput knjižnica posjeduju određen broj računala pa učenje putem Interneta nije dostupno svima u svakom trenutku što je velik nedostatak ovakvog oblika učenja. Financijski troškovi posjedovanja priključka na Internet kao i posjedovanja računala još su jedan od nedostataka Interneta.

IV. E-UČENJE

E-učenje u odnosu na klasične oblike učenja za prednost ima mogućnost učenja na daljinu što uključuje i stalno usavršavanje znanja tzv. 'long life learning', komuniciranje i diskutiranje s drugim sudionicima e-učenja, usavršavanje znanja i vještina korištenja različitih tehnologija kao i neovisnost o prostoru i vremenu odnosno fleksibilnost kod usvajanja znanja. Osim toga ističe se kvalitetnija interaktivnost između nastavnika i studenata, a što vodi uspješnijem svladavanju gradiva čime se podiže kvaliteta nastave. E-učenje pruža mogućnost jednostavnijeg organiziranja nastave kao i uključivanje svjetskih stručnjaka u nastavu kroz videokonferencijske prijenose jer ne postoji potreba za putovanjem, bilo navedenih stručnjaka, nastavnika ili učenika.

No, uz brojne navedene prednosti postoje i neki nedostaci e-učenja kao što su nedostatak ljudskog kontakta odnosno neverbalne komunikacije što može voditi do nejasnoća i nerazumijevanja. Problemi se isto tako javljaju zbog nepoznavanja tehnologija potrebnih u e-obrazovanju. Također postoje područja u obrazovanju koja se ne mogu obrađivati isključivo online odnosno e-učenjem. Važno je istaknuti kako je teško održati motiviranost kod e-učenja zbog čega može doći do odustajanja.

V. ZAKLJUČAK

Obzirom da se u budućnosti očekuje se još veći pomak od klasičnih oblika učenja prema različitim oblicima e-učenja u radu su razmotrene prednosti i nedostaci tiskanih materijala, kao i prednosti i nedostaci digitalnih materijala. Osim toga, Internet kao poseban medij nudi nove mogućnosti kod usvajanja i proširivanja nastavnih sadržaja te se u sve većoj mjeri očekuje uključivanje u sustav odgoja i obrazovanja kroz prevladavanje nedostataka koji su još uvijek prisutni prilikom služenja ovim alatom u obrazovanju.

Može se zaključiti da „digitalni obrazovni sadržaj (DOS) potiče aktivno učenje na inovativan, učinkovit, motivirajući i pojedincu prilagođen način, te omogućuje korištenje raznolikih metoda poučavanja“, što je prepoznala Hrvatska akademska i istraživačka mreža – Carnet te je s tim u skladu organizirala javnu raspravu na temu „Prijedlog kriterija za izradu kvalitetnih digitalnih obrazovnih sadržaja“ [6] koja je bila otvorena do 8. siječnja 2016. godine, a za svrhu je imala definiranje načela izrade kvalitetnih digitalnih nastavnih materijala, kao i kriterije za ocjenjivanje digitalnog obrazovnog sadržaja u osnovnim i srednjim školama. Pilot projekt „e-Škole: Uspostava sustava razvoja digitalno zrelih škola“ [7] koji sufinancira Europska unija, u Hrvatskoj se provodi kroz Cjelovitu kurikularnu reformu u osnovnim i srednjim školama. Shodno tome, u budućnosti se očekuje da će sve navedeno imati svoj odraz i u izradi, razvoju i upotrebi još kvalitetnijih digitalnih nastavnih materijala na visokim učilištima u Republici Hrvatskoj, a osobito obzirom da se, kao posljedica dubokih promjena u obrazovanju i obrazovnim procesima, „repozitoriji digitalnog obrazovnog materijala kao sastavnica kvalitete suvremenih koncepta obrazovanja“ [8] nameće kao nužnost.

LITERATURA

[1] Grafički fakultet – Sveučilište u Zagrebu, http://eprints.grf.unizg.hr/1533/1/DB285_Sestak_Antonija.pdf. Pristupljeno: 25.01.2016. A. Šestak, Razina interaktivnosti i potencijal e-knjige, te implementacija u sustave e-obrazovanja, Zagreb, 2013, 1-85.

[2] M. Šapro-Ficović, Masovna digitalizacija knjiga: utjecaj na knjižnice, Vjesnik bibliotekara Hrvatske, vol.54, no.172, lipanj 2011, str. 216-250.

[3] CARNet - Hrvatska akademska i istraživačka mreža, http://www.carnet.hr/upload/javniweb/images/static3/91307/File/Digitalni_nastavni_materijali_-_prezentacija.pdf Pristupljeno: 25.01.2016. Digitalni nastavni materijali, Carnet - Hrvatska akademska i istraživačka mreža, 2015, prezentacija 1-27.

[4] Tiskarstvo 08 - Prof. Vilko Žiljak, http://www.ziljak.hr/tiskarstvo/tiskarstvo08/Radovi08/ZA%20WEB/TiskaniIliDigitalni113.html/ Pristupljeno: 25.01.2016. B. Vukelić, I.Pogarčić, M. Gligora Marković, Tiskani ili digitalni nastavni materijal?, Tiskarstvo 08, FS d.o.o., Zagreb, 2008.

[5] Knjižnica Filozofskog fakulteta - Sveučilište u Zagrebu, http://darhiv.ffzg.unizg.hr/23/1/TomanaBurger.pdf. Pristupljeno: 25.01.2016. T. Burger, Percepcija i stavovi studenata prema “učenju putem Interneta, Zagreb, 2003.

[6] CARNet - Hrvatska akademska i istraživačka mreža, http://www.carnet.hr/upload/javniweb/images/static3/92862/File/PrijedlogkriterijaipreporukaDOS-zajavnuraspravu.pdf. Pristupljeno: 25.01.2016. Prijedlog kriterija i preporuka za izradu kvalitenih digitalnih nastavnih sadržaja – za javnu raspravu, Carnet - Hrvatska akademska i istraživačka mreža, 2015, 1-14.

[7] CARNet - Hrvatska akademska i istraživačka mreža, http://www.carnet.hr/upload/javniweb/images/static3/92862/File/DOS%20-%20javna%20rasprava.pdf. Pristupljeno: 25.01.2016. S. Lušić Radošević (voditeljica radne skupine za definiranje kriterija za digitalne obrazovne sadržaje), Kriteriji i preporuke za izradu kvalitetnih digitalnih obrazovnih sadržaja, e-Škole: Uspostava sustava razvoja digitalno zrelih škola (pilot projekt), 2015, prezentacija 1-46.

[8] E-LIS repository - rclis, http://eprints.rclis.org/6524/1/edupoint_33_lasic_banek_spiranec.pdf. Pristupljeno: 01.02.2016. J. Lasić-Lazić, S. Špiranec, M. Banek Zorica, Repozitoriji digitalnog obrazovnog materijala kao sastavnica kvalitete suvremenih koncepta obrazovanja, Edupoint V, 2005.

Istraživanje stavova i očekivanja studenata prilikom upisa na studij kao metoda povećanja kvalitete usluge u visokom obrazovanju

Tihana Babić, Silvija Grgić i Eugen Rajković (student)
Visoko učilište Algebra – visoka škola za primijenjeno računarstvo, Zagreb, Hrvatska
tihana.babic@racunarstvo.hr, silvija.grgic@racunarstvo.hr, eugen.rajkovic@racunarstvo.hr

Sažetak - Globalizacija ima velik utjecaj na suvremeno poslovno okruženje i u takvom okruženju obrazovanje sve više postaje usluga. Visoka učilišta moraju kontinuirano podizati kvalitetu usluge, kako bi mogle privući i zadržati što veći broj studenata. Kvaliteta usluge jedan je od glavnih faktora konkurentnosti, a svaki pružatelj usluga treba poznavati svoje korisnike kako bi mogao pružiti uslugu po mjeri korisnika. Kroz rad su prikazani stavovi i očekivanja studenata prilikom upisa na studij, a podaci su prikupljeni metodom anketiranja. Anketa je bila dobrovoljna, a odvijala se u sklopu i tijekom nastave. Istraživanje je provedeno na uzorku 137 novoupisanih studenata na 1. godinu preddiplomskih studija Visokog učilišta Algebra u školskoj godini 2015./2016., kao budućih korisnika usluge.
Svrha ovog istraživanja jest identificirati motive upisanih studenata za upis studija na Visokom učilištu Algebra te stavove i očekivanja studenata vezanih uz njihovu uključenost u nastavne procese, izvannastavne aktivnosti, nastavak obrazovanja i razvoj karijere. Rezultati provedenog istraživanja imat će utjecaj na planiranje budućih aktivnosti koje će izravno i neizravno doprinijeti zadovoljstvu studenata i povećanju kvalitete obrazovnog procesa.

I. UVOD

Proces globalizacije, neovisno kako ga definirali, ima sve veći utjecaj na svakidašnji život i na društvene prilike, a shodno tome i na suvremeno poslovno okruženje.

Rastući globalni trendovi i promjene koje se događaju na tržištu, utječu na poduzeća i ona su prisiljena ne samo pronalaziti nova tehnička i tehnološka rješenja, već prilagođavati svoju organizacijsku strukturu novonastalim uvjetima. Ti uvjeti odražavaju se i na područje obrazovanja, a da bi zadržale konkurentsku prednost visokoškolske ustanove sve više postaju pružatelji „usluge“ obrazovanja.

U okruženju u kojem obrazovanje odnosno izvođenje studijskih programa sve više postaje usluga, obrazovne ustanove moraju se kontinuirano razvijati i podizati ljestvicu kvalitete usluge koju pružaju kako bi mogle privući i zadržati što veći broj studenata, ali i da bi se mogle natjecati za upis najboljih kandidata, baš kao i poduzeća koja pružaju profesionalne i poslovne usluge u realnom sektoru.

Kvaliteta postaje jedan je od glavnih faktora konkurentnosti, a svaki pružatelj usluga treba poznavati svoje korisnike kako bi mogao mobilizirati svoje potencijale i povećati kvalitetu svojih usluga.

A. Kvaliteta

U ekonomiji 21. stoljeća utemeljenoj na znanju i kompetencijama, u uvjetima snažne konkurencije, kvaliteta obrazovanja ima najznačajniju ulogu. Trenutno ne postoji konsenzus ili jedinstvena definicija kvalitete u visokom obrazovanju, kao što ne postoji jedinstveni sistem niti skup indikatora koji bi prikazivali kompletnu, točnu i dubinsku sliku kvalitete institucija u visokom obrazovanju. Jedna od tradicionalnih definicija kvalitete kaže da je kvaliteta „Mogućnost proizvoda da ispunjavaju zahtjeve kupca“.

Kvaliteta nije slučajan događaj, već je rezultat kontinuiranog planiranja i unapređivanja poslovnih procesa. Planiranje kvalitete odnosi se prvenstveno na identificiranje koji su standardi kvalitete relevantni za projekt i utvrđivanje kako ih zadovoljiti [1].

Plan upravljanja kvalitetom treba obuhvatiti niz čimbenika. Neovisno o definicijama kvalitete, postignut je konsenzus na razini međunarodne akademske zajednice o kvaliteti, a kao najvažniji čimbenici i akteri u procesu provođenja kvalitete iskristalizirali su se: uspješno rukovođenje odnosno uprava koja točno percipira što njihovi potrošači žele, angažman zaposlenika odnosno osoblje koje stručno i pozitivno usmjereno na korisnike, jasno definirani, razumljivi procesi i aktivnosti koje međusobno povezane dovode do provedbe zajedničkih ciljeva, razvoj i stalno unapređivanje postojećih razina provedbe koje onda omogućavaju bržu i lakšu prilagodbu novim vanjskim i unutarnjim uvjetima kao i stvaranje novih prilika i mogućnosti, donošenje odluka koje temelj imaju na objektivnim dokazima i znanju kao što su činjenice i analize podataka, upravljanje odnosima sa svim zainteresiranim stranama, a osobito s partnerima, dobavljačima i društvom u cjelini.

No, kako navodi internacionalna organizacija ISO 9000, primarni princip upravljanja kvalitetom jest fokus na korisnicima i obzirom da organizacije ovise o svojim korisnicima, trebaju razumjeti njihove trenutne i buduće potrebe, upoznati njihove zahtjeve i težiti da ispune njihova očekivanja [2].

Uspješnost provođenja poslovnih procesa stoga „uvelike ovisi o zahtjevima koje postavljaju korisnici“ [3]. Imajući uspješno poslovanje kao cilj, obveza tvrtke je ispunjavati zahtjeve svojih klijenata odnosno korisnika.

Sukladno tome, svaki sustav, proces, usluga ili proizvod mora se prilagoditi zahtjevu korisnika, smanjiti razinu

eksternih ili internih neuspjeha te pružati i povećavati kvalitetu, kako god je poslovanje ili korisnik vidjeli [4].

B. Upravljanje kvalitetom usluga

Usluga se može definirati i tumačiti na više načina, no u svjetskoj literaturi dominira određenje usluge kao bilo koje aktivnosti ili koristi koju jedna stranka nudi drugoj i koja je u osnovi neopipljiva i ne rezultira vlasništvom nad nečim. Njezina proizvodnja može ali i ne mora biti vezana za kakav fizički proizvod [5].

Uzevši visoko obrazovanje kao pružanje usluge ono se može kategorizirati kao „čista usluga“ jer je glavni cilj povećanje obrazovanja i edukacija pojedinca, a ne fizički opipljiv proizvod [6].

U svrhu poboljšanja kvalitete, moraju se identificirati ključne odrednice odnosno glavni i ključni kriteriji na temelju kojih korisnici prosuđuju o kvaliteti usluge, koja su očekivanja ciljnih korisnika i kako klijenti ocjenjuju uslugu tvrtke u odnosu na svoja očekivanja [5].

Kvaliteta usluge ima izuzetno bitnu ulogu prilikom donošenja odluke o kupovini i kao takva prilično utječe na zadovoljstvo kupca [7]. Upravo iz tog razloga nužno je kontinuirano pratiti, analizirati i prilagođavati razne aspekte kvalitete obrazovne usluge kako bi se efikasno mogli pratiti zahtjevi tržišta i glavnih korisnika. Problemi koji se često javljaju u tom procesu su na koji način mjeriti kvalitetu obrazovne usluge i subjektivno shvaćanje kvalitete od strane različitih sudionika procesa evaluacije [8].

Unatoč tome obrazovne institucije mogu mjeriti kvalitetu na način da odrede kako klijenti usluge percipiraju kvalitetu. Neke studije sugeriraju da su ocjene klijenata o kvaliteti usluga rezultat uspoređivanja vlastitih očekivanja s vlastitim iskustvom vezanim uz uslugu. Svaki raskorak u tome predstavlja „raskorak u kvaliteti“ [5]. Prema tome, sve što klijent, u ovom slučaju student, smatra da je realno i jest realno; kvaliteta je sve što klijent smatra da je kvaliteta.

U nastalim uvjetima u kojima više nije moguće izbjeći globalizaciju, a biti uspješan, mijenja se, dakle, i uloga studenata. Studenti sve više postaju korisnici usluge, a visoka učilišta kao pružatelji usluga trebaju (u)poznavati svoje korisnike kao preduvjet korištenja svojih potencijala u svrhu povećanja kvalitete svojih usluga odnosno kako bi mogle ispuniti najvažniji postulat u poslovnom smislu, a to je pružiti uslugu po mjeri korisnika.

Potaknuti tim spoznajama, autori su pokušali istražiti stavove, mišljenja i očekivanja studenata prilikom upisa na studije Visokog učilišta Algebra – visoke škole za primijenjeno računarstvo kako bi u tome mogli uspjeti.

II. METODOLOGIJA ISTRAŽIVANJA MOTIVA UPISANIH STUDENATA ZA UPIS STUDIJA NA VISOKOM UČILIŠTU ALGEBRA TE STAVOVA I OČEKIVANJA STUDENATA VEZANIH UZ NJIHOVU UKLJUČENOST U NASTAVNE PROCESE, IZVANNASTAVNE AKTIVNOSTI, NASTAVAK OBRAZOVANJA I RAZVOJ KARIJERE

A. Opis uzorka

Istraživanje je provedeno na Visokom učilištu Algebra – visokoj školi za primijenjeno računarstvo u prvom semestru školske godine 2015./2016. U dobrovoljnoj originalnoj anketi sudjelovali su studenti upisani na 1. godinu preddiplomskih studija Primijenjenog računarstva, Multimedijskog računarstva i Digitalnog marketinga u školskoj godini 2015./2016., kao budući korisnici usluga Visokog učilišta Algebra. Anketu je popunilo 137 studenata od upisanih 194 studenata.

U Tablici I. prikazana je struktura sudionika s obzirom na upisani studij, spol i status studija.

TABLICA 1. STRUKTURA SUDIONIKA S OBZIROM NA UPISANI STUDIJ, SPOL I STATUS STUDIJA

Upisani preddiplomski studij	**Digitalni marketing**	**Multimedijsko računarstvo**	**Primijenjeno računarstvo**
Broj studenata (N)	40	23	74
Spol			
Muški	25	21	67
Ženski	15	2	7
Status			
Redoviti	37	20	56
Izvanredni	3	3	18

B. Ciljevi istraživanja

Visoko učilište Algebra – visoka škola za primijenjeno računarstvo u provedenom istraživanju u svrhu pružanja kvalitete usluge po mjeri studenata, imalo je sljedeće ciljeve:

- Identificirati motive upisanih studenata za upis studija na Visokom učilištu Algebra.
- Upoznati očekivanja studenata prilikom upisa na studij, a vezano uz njihovu uključenost u nastavne procese i izvannastavne aktivnosti.
- Istražiti stavove studenata u svezi s nastavkom obrazovanja i razvojem karijere.

C. Korišteni instrumenti u istraživanju

Podaci su prikupljeni ispitivanjem pomoću upitnika koji je dijelom sadržavao unaprijed ponuđene odgovore zatvorenog tipa. Na skali od 5 stupnjeva Likertovog tipa (uvijek, često, povremeno, rijetko, nikad) sudionici su odgovarali na pitanja u svezi s očekivanjima i stavovima, te pitanja višestrukog izbora u svezi s motivima upisa na studij. Studentima je objašnjena svrha ispitivanja, nakon čega su dobili anketne upitnike te su zamoljeni da iskreno odgovore na pitanja.

Odgovori ispitanika obrađeni su kvantitativnom i kvalitativnom analizom.

III. REZULTATI ISTRAŽIVANJA

Rezultati ovog istraživanja nisu pokazali otklon obzirom na spol ili status studija, tako da nisu prema tim varijablama predočeni u prikazu, a obzirom da se radi o kvalitativnim podacima istraživanja, nije bilo moguće izračunati statističke mjere centralne tendencije i varijabilnosti.

A. Motivi za upis studija na Visokom učilištu Algebra

Studentima je postavljeno pitanje kako su se odlučili za studij na Visokom učilištu Algebra. U upitniku su unaprijed ponuđeni mogući motivi koji se u javnosti često ističu kao općenite prednosti prilikom odabira studija te dodatni motivi koji se često ističu kao prednost upisa na studije na Visokom učilište. Sudionici su mogli birati više odgovora, a zamoljeni su da odaberu najvažnijih pet.

U Tablici II. prikazane su frekvencije odgovora, ukupan broj odgovora je 593.

TABLICA II. MOTIVI IZBORA STUDIJA (N=137)

Motivi za upis	
Dobri poslovi i lakše zapošljavanje.	13%
Visoka zapošljivost u digitalnoj industriji/ekonomiji.	13%
Dobra akademska reputacija Algebre.	12%
Algebra je po ocjenama kvalitete u samom vrhu.	11%
Odnos cijene i kvalitete studija.	8%
Društvena/medijska zastupljenost Algebre.	6%
Zbog informacija dostupnih na web stranici Algebre.	6%
Posjetio sam ranije prostore Algebre.	4%
Savjet nekoga tko radi u digitalnoj/IT struci.	4%
Zbog financijske pomoći/stipendije.	3%
Zbog savjeta studenta koji već studira na Algebri.	3%
Visoki postotak studenata završava ovaj fakultet.	3%
Zbog aktivnosti Algebre u kojima sam sudjelovao/la.	3%
Zbog veličine fakulteta (broja studenata).	3%
Mogućnost on-line praćenja nastave.	2%
Nisam uspio/uspjela upisati državni fakultet.	2%
Poslodavac mi je predložio studij na Algebri.	1%
Prijatelj ili član obitelji je upisao/la Algebru.	1%
Roditelji/bliski rođaci su željeli da upišem Algebru.	1%
Na Algebri se lako dolazi do diplome.	0%
Preporuka savjetnika u srednjoj školi.	0%
Prvi izbor fakulteta bio mi je financijski nepodnošljiv.	0%
Tako me savjetovao nastavnik u srednjoj školi.	0%
Zbog ostvarivanja studentskih prava.	0%
Zbog sportskih i drugih studentskih aktivnosti.	0%

Pet najčešćih motiva koje su studenti naveli kao svoje motive za upis na studij Visokog učilišta Algebra – visoke škole za primijenjeno računarstvo, što je vidljivo iz podataka navedenih u Tablici II., su dobri poslovi i lakše zapošljavanje nakon završetka studija, visoka zapošljivost u digitalnoj industriji/ekonomiji, dobra akademska reputacija visokog učilišta, vrhunske ocjene kvalitete visokog učilišta i odnos cijene i kvalitete studija koje smatraju usklađenima.

Niti jedan student nije kao motiv upisa naveo pretpostavku da se na Visokom učilištu Algebra lako dolazi do diplome, što je u skladu s najčešće odabranim motivima te se također može povezati s dobrom akademskom reputacijom visokog učilišta i vrhunskim ocjenama kvalitete.

Dodatno se iskristaliziralo da studenti nisu studij upisali zbog ostvarivanja studentskih prava, jer im je neki drugi studij bio financijski nepodnošljiv ili jer su im tako savjetovali nastavnici i savjetnici u srednjim školama te zbog sportskih i drugih studentskih aktivnosti.

Društvena i medijska zastupljenost i informacije dostupne na web stranicama Visokog učilišta Algebra svrstavaju se u srednje zastupljene, ali nikako zanemarive i irelevantne poticaje za upis studija.

B. Očekivanja studenata vezano uz njihovu uključenost u nastavne procese i izvannastavne aktivnosti.

Studentima su postavljena pitanja koja s odnose na njihovo sudjelovanje u nastavnom procesu, kao i izvannastavnim aktivnostima. U upitniku su unaprijed ponuđeni odgovori intenziteta: uvijek, često, ponekad, rijetko i nikad. Sudionici su mogli birati jedan odgovor koji najviše pobliže opisuje njihovo očekivanje vezano uz njihovo sudjelovanje u navedenoj aktivnosti.

a) Očekivano sudjelovanje u nastavnim procesima

Pitanja koja se tiču sudjelovanja u nastavnim procesima, a temeljeno na očekivanom sudjelovanju studenata, glasila su:

Pitanje 1. Koliko često očekujete da ćete sudjelovati u diskusijama na nastavi?

Pitanje 2. Koliko često očekujete da ćete samostalno učiti?

Pitanje 3. Koliko često očekujete da ćete učiti u grupi s drugim studentima?

Pitanje 4. Koliko često očekujete da ćete voditi bilješke u nastavi?

Pitanje 5. Koliko često očekujete da ćete pisati zadaće i seminarske radove?

Pitanje 6. Koliko često očekujete da ćete koristiti knjižnicu?

Na navedenih 6 pitanja iskristalizirali su se odgovori koje u postotnom omjeru navodimo u Tablici III.

TABLICA III. OČEKIVANO SUDJELOVANJE U NASTAVNIM PROCESIMA (N=137)

	Uvijek	**Često**	**Ponekad**	**Rijetko**	**Nikad**
Pitanje 1.	10%	35%	46%	8%	0%
Pitanje 2.	23%	64%	14%	0%	0%
Pitanje 3.	4%	19%	40%	34%	4%
Pitanje 4.	26%	34%	31%	8%	1%
Pitanje 5.	37%	35%	24%	3%	0%
Pitanje 6.	2%	11%	54%	33%	1%

Iz navedenih rezultata vidljivo je da gotovo polovina studenata očekuje da će barem ponekad sudjelovati u diskusijama na nastavi, a više od trećine to čak očekuje često, dok nema studenata koji to nikad ne očekuju.

Ukupno 87% studenata očekuje da će samostalno učiti često, odnosno uvijek, a 14% očekuje da će to činiti samo ponekad. Može se uočiti da studenti ne očekuju da to neće činiti nikad ili rijetko.

Učenje u grupi studenti očekuju ponekad ili rijetko, ukupno njih 74%, a mali postotak, točnije njih 4% to ne očekuje nikad.

Vođenje bilježaka 60% studenata namjerava nečim što očekuju redovno raditi, 31% ih namjerava voditi ponekad, a 1% nikad.

Pisanje zadaća i seminarskih radova također je aktivnost koju studenti smatraju čestom, njih 35%, povremenom ju smatra 24%, a najveći postotak odnosno 37% studenata smatra da će tu aktivnost uvijek obavljati.

Što se tiče korištenja knjižnice, najviše studenata smatra, ukupno 87%, smatra da će ju koristiti tek ponekad ili rijetko.

b) Očekivano sudjelovanje u izvannastavnim aktivnostima

Pitanja koja se odnose na izvannastavne aktivnosti, a u svezi s očekivanim sudjelovanju studenata, bila su:

Pitanje 1. Koliko često očekujete da ćete sudjelovati na predavanjima gostujućih predavača?
Pitanje 2. Koliko često očekujete da ćete sudjelovati u projektima, organiziranju događaja, natjecanjima?
Pitanje 3. Koliko često očekujete da sudjelovati u promotivnim aktivnostima (npr. Smotra sveučilišta, javne konferencije)?
Pitanje 4. Koliko često očekujete da ćete sudjelovati u sportskim aktivnostima?
Pitanje 5. Koliko često očekujete da ćete sudjelovati u radu studentskih organizacija?
Pitanje 6. Koliko često očekujete da ćete razgovarati o studiju sa stručnim službama i odgovornim osobama Visokog učilišta Algebra?
Pitanje 7. Koliko često očekujete da ćete razgovarati o razvoju vlastite karijere sa stručnim službama i odgovornim osobama Visokog učilišta Algebra?

U Tablici IV. prikazane su frekvencije odgovora studenata koji se odnose na očekivano sudjelovanje u izvannastavnim aktivnostima.

TABLICA IV. OČEKIVANO SUDJELOVANJE U IZVANNASTAVNIM AKTIVNOSTIMA (N=137)

	Uvijek	Često	Ponekad	Rijetko	Nikad
Pitanje 1.	20%	43%	29%	7%	1%
Pitanje 2.	4%	27%	37%	29%	4%
Pitanje 3.	4%	17%	35%	34%	10%
Pitanje 4.	6%	7%	31%	32%	24%
Pitanje 5.	3%	5%	25%	37%	30%
Pitanje 6.	2%	21%	55%	20%	2%
Pitanje 7.	8%	32%	45%	11%	4%

Rezultati pokazuju da ukupno 63% studenata očekuje da će vrlo često sudjelovati na predavanjima gostujućih predavača, što nije u skladu s onim što je pokazala praksa prethodnih godina, obzirom da su se na predavanjima gostujućih predavanja studenti pojavljivali u veoma malom broju.

27% studenata namjerava vrlo često sudjelovati u projektima, organiziranju događaja i natjecanjima dok ih ukupno 66% u navedenim aktivnostima očekuje sudjelovati ponekad ili rijetko.

Ukupno 69% studenata promotivne aktivnosti visokog učilišta, kao što su na primjer Smotra sveučilišta i javne rasprave, smatra nečim u čemu se sudjelovati rijetko ili ponekad, a samo 4% studenata namjerava u tim aktivnostima sudjelovati uvijek.

Sudjelovanje u sportskim aktivnostima studenti najvećim dijelom, njih ukupno 87% smatra aktivnostima u kojima očekuju sudjelovati ponekad, rijetko ili nikad.

Rad u studentskim organizacijama, odnosno sudjelovanje u radu studentskih udruga pokazalo se u istraživanju također kao aktivnost koja studente u velikom postotku, čak 92%, zanima tek ponekad, rijetko ili ih uopće ne zanima.

Više od polovice ispitanika odnosno 55% očekuje da će barem ponekad tijekom studija razgovarati sa stručnim službama i odgovornim osobama visokog učilišta, 23% očekuje to često ili uvijek, a 22% rijetko ili nikad.

Također, studenti očekuju da će o razvoju vlastite karijere povremeno tijekom studija razgovarati sa stručnim službama i odgovornim osobama, čak 40% smatra da će to činiti često ili uvijek, a 15% to uopće ne očekuje činiti.

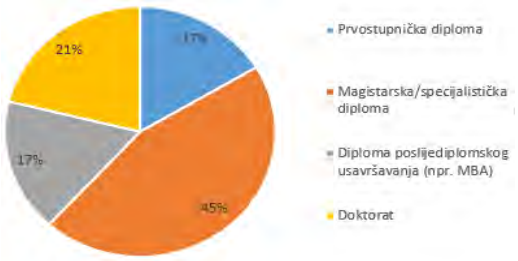
C. Stavovi studenata u svezi s nastavkom obrazovanja i razvojem karijere

a) Nastavak obrazovanja

U anketnom upitniku studentima je bilo postavljeno pitanje koji je njihov krajnji obrazovni cilj, a kao odgovor ponuđene su im opcije: Prvostupnička diploma, Magistarska/specijalistička diploma, Diploma poslijediplomskog usavršavanja (npr. MBA) i Doktorat.

U Grafikonu I. navedena je frekvencija pojedinog krajnjeg obrazovnog cilja koji studenti očekuju.

Grafikon I. KRAJNJI OBRAZOVNI CILJ (N=137)



Rezultati pokazuju da samo 17% studenata očekuje prvostupničko obrazovanje kao krajnji cilj, a najveći broj ispitanika, njih 45%, kao obrazovni cilj imaju uspješan završetak diplomskog studija. Poslijediplomsko usavršavanje, kao što je npr. MBA, za krajnji cilj svog obrazovanja navelo je 17% studenata.

No, posebno indikativan je podatak da čak 21% studenata namjerava upisati i završiti doktorski studij, što trenutačno u Republici Hrvatskoj nije moguće učiniti zbog neprohodnosti studenata stručnih studija prema poslijediplomskim sveučilišnim doktorskim studijima Nameće se kao zaključak da će navedeni studenti svoj krajnji obrazovni cilj pokušati ostvariti u nekoj drugoj europskoj zemlji gdje će im to biti omogućeno.

b) Razvoj karijere

Kako bi utvrdili što ispitanici očekuju u razvoju svoje karijere nakon završetka studija, ispitanicima je postavljeno pitanje što očekuju da će dobiti završetkom studija na Visokom učilištu Algebra. Bilo je ponuđeno 11 unaprijed

definiranih odgovora, a sudionici istraživanja su mogli birati više njih na postavljeno pitanje, odnosno one koji najviše opisuju njihov stav i mišljenje o budućem razvoju karijere.

Tablica V. prikazuje mogućnosti odgovora i učestalost odabira pojedinog odgovora na navedeno pitanje, a ukupni broj odgovora bio je 590.

TABLICA V. RAZVOJ KARIJERE NAKON ZAVRŠETKA STUDIJA (N=137)

Što očekujete da ćete dobiti završetkom studija na Visokom učilištu Algebra?	
Stručna znanja i vještine.	18%
Nova znanja za bolje radno mjesto.	16%
Nove ideje i mogućnosti za osobni razvoj.	14%
Upoznavanje zanimljivih ljudi.	13%
Odskočnu dasku za inozemstvo.	10%
Nastavak obrazovanja.	10%
Bolji platni koeficijent na poslu.	8%
Društveni status.	7%
Diplomu u što kraćem vremenu.	2%
Nešto drugo.	2%
Nemam nikakva očekivanja.	0%

Sudionici su kao najvažnija očekivanja od završetka stručnog studija naveli usvajanje novih stručnih znanja i vještina koja će im omogućiti kandidiranje za bolja radna mjesta u budućnosti što su primarna očekivanja sudionika, kao što je vidljivo prema ukupno 34% posto odabranih opcija odgovora. Iz 14% odabranih odgovora može se zaključiti da nove ideje i mogućnosti osobnog razvoja studenti također smatraju najvećim dobitkom tijekom i nakon završetka studija. Zanimljiv podatak je da se čak 13% odgovora odnosi na upoznavanje ljudi kao važno ostvarenje. Podjednak postotak odnosno 10% odabranih odgovora ukazuje da ispitanici završetak studija smatra kao odskočnu dasku za inozemstvo, kao i mogućnošću za nastavak obrazovanja. 8% odabranih odgovora ukazuje na bolji platni koeficijent kao očekivanje sudionika, a 7% odgovora očekuje određen društveni status. Najmanji postotak tj. 2% odgovora odnosi se na diplomu u što kraćem vremenu ili nešto drugo, dok nema ispitanika koji ništa ne očekuje od završetka studija na Visokom učilištu Algebra.

IV. ZAKLJUČAK

Planiranje kvalitete od iznimne je važnosti. U sklopu planiranja važno je identificirati studente određenog studija, spola, statusa kao eventualnih specifičnih korisnika kao i identificirati njihove zahtjeve, saznati koji su bitni ili manje bitni te u konačnici organizirati sustav obrazovanja koji će studentima odnosno korisnicima usluge pružiti kvalitetnu uslugu po mjeri.

Rezultati ovog istraživanja pokazuju da su najčešći motivi upisa povezani s očekivanjima tijekom i nakon studija, i iskristaliziralo se da su to stjecanje novih stručnih znanja i vještina te osobni razvoj kako bi u budućnosti kako bi nastavili svoje obrazovanje u navedenom području i/li ostvarili prednost prilikom kandidiranja za radna mjesta i zapošljavanja, a postigli bolji standard i stekli društveni status i nova poznanstva. Visoko učilište Algebra – visoka škola za primijenjeno računarstvo to im približava na način da im omogućuje studiranje na studijima ICT-a, koji je trenutno u svijetu jedan od sektora u kojem se zapošljava najviše novih stručnjaka, ali i stoga što sustavno ima fokus na svojim korisnicima, razvija se i nastoji pružiti uslugu koja ne samo zadovoljava, nego i nadmašuje očekivanja korisnika.

No, obzirom da je samo mjerenje kvalitete u obrazovanju veoma složeno, zbog njezinih specifičnih karakteristika kao kompleksne usluge koja sadrži niz materijalnih, ali i nematerijalnih komponenti, osobito je važno voditi računa da se percepcija usluge može kod korisnika razlikovati, pa iako se radi o istoj razini kvalitete usluge korisnici ju mogu različito doživljavati ovisno o svojim osobnim preferencijama.

Ovo istraživanje nosilo je i svoje ograničenje, kao što je na primjer nemogućnost mjerenja koliko su početna očekivanja ispitanika ispunjena, stoga bi svakako trebalo provesti istraživanje na kraju njihovog studija koje bi pokazalo u kolikoj mjeri smatraju da su njihovi zahtjevi postavljeni prilikom upisa na studij zaista bili ispunjeni tijekom njihovog školovanja. Također, obzirom da dimenzije kvalitete usluge visokog obrazovanja široko variraju, „modele upravljanja i osiguranja kvalitete neophodno je kontinuirano inovirati i prilagođavati novonastalim uvjetima“ [9].

LITERATURA

[1] R. Kenneth, “Project quality management: why, what and how” J.Ross Publishing, Inc., Florida, 2005, 173.

[2] ISO 9000 quality management - ISO, http://www.iso.org/iso/pub100080.pdf. Pristupljeno: 31.01.2016. Quality management principles“, International Organization for Standardization, Geneva, Switzerland, 2015, ISBN 978-92-67-10650-2, 1-20.

[3] D. Delić, D. Bebek, “Razumijevanje kvalitete u području upravljanja projektima”, neobjavljen.

[4] Ž. Mikić, N. Momčilović, D. Jovičić, “Planiranje kvalitete u projektima”, neobjavljen.

[5] P. Kottler, V. Wong, J. Saunders, G. Armstrong, „Osnove marketinga“, Mate d.o.o., Zagreb, 2006.

[6] P. Kottler, K. L. Keller, „Upravljanje marketingom“, 12. izdanje, Mate d.o.o., Zagreb, 2008.

[7] A. Caruana, A. H. Money, P. R. Berthon, „Service quality and satisfaction – the moderating role of value“, European Journal of Marketing, vol. 34, No. 11/12, str. 1338-1353.

[8] M. Leko Šimić, H. Štimac, „Kvaliteta obrazovne usluge – slučaj Ekonomskog fakulteta u Osijeku“, Zagreb, Ekonomski fakultet, 2013.

[9] SeMA | Srpsko udruženje za marketing | Serbian Marketing ..., http://www.sema.rs/repository/download/marketing-vol-45-no-3.pdf. Pristupljeno: 15.01.2016. M. Vranješ, D. Gašević, D. Drinić, „Analiza elemenata kvaliteta usluga u visokom obrazovanju“, Marketing, SeMa, Beograd, 2014, 45/03, 213-222.

E-obrazovanjem do fleksibilnog modela učenja

M. Bo̧urić, R. Bogut i M. Tretinjak
Elektrotehnička škola, Zagreb, Hrvatska
mario.bozuric@gmail.com, romanabogut@gmail.com, mario.tretinjak@gmail.com

Sažetak - U školskoj godini 2010/2011 Elektrotehnička škola iz Zagreba i Elektrostrojarska škola iz Varaždina sudjelovale su u realizaciji i implementaciji IPA projekta pod nazivom With Flexible Learning Approach and Curricula Innovations towards the Labour Market podržanog od Hrvatske agencije za strukovno obrazovanje i osposobljavanje odraslih. Pet godina nakon završetka projekta autori ovog rada žele ukazati na prednosti i slabosti uvođenja fleksibilnog modela učenja u strukovne škole, a za čiju provedbu su neophodne metode e-učenja i m-učenja. Ovaj rad prikazuje konkretne učinke uvođenja e-obrazovanja u obrazovni sustav kako bi se održao i razvio fleksibilni model učenja u strukovnim školama. Autori su uvjereni da iskustva opisana u ovom radu mogu pomoći pri planiranju i provođenju sličnih projekata u drugim strukovnim školama.

I. UVOD

Ujednačeni školski kurikulum predstavlja temelj za sve strukovne škole. Takav obrazovni sustav nastavnicima ne omogućava dovoljno fleksibilnosti u nastavnom procesu s ciljem prepoznavanja i poticanja individualnih sposobnosti i vještina učenika. Kako bi se to promijenilo potrebno je razvijati obrazovni pristup koji će se temeljiti na fleksibilnosti, prepoznavanju i isticanju pojedinačnih vrijednosti, znanja i vještina svakog pojedinog učenika [1].

Elektrotehnička škola iz Zagreba i Elektrostrojarska škola iz Varaţdina sudjelovale su u zajedničkom projektu "With Flexible Learning Approach and Curricula Innovations towards the Labour Market" s ciljem implementacije fleksibilnog modela učenja u klasični oblik nastave, te razvoja dodatnih školskih aktivnosti pokretanjem karijernog kutića pomoću kojeg bi učenici mogli ostvariti kontakte s tvrtkama iz područja elektrotehnike i računarstva. Sudionici projekta smatrali su da će navedenim projektom učenici biti bolje pripremljeni za zahtjeve trţišta rada [2].

Vezano uz navedeni projekt, grupa autora je provela istraţivanje i napisala članak pod nazivom „Flexible Learning Approach and Curricula Education towards the Labour Market in Secondary Technical Education: A Case Study". Autori su istraţivali izazove provedbe fleksibilnog učenja u Elektrotehničkoj školi u Zagrebu, s osvrtom na osposobljavanje i prilagodbu nastavnika, kao i prihvaćanje drugačijeg pristupa u nastavi od strane učenika [3].

Implementacijom fleksibilnog modela učenja učenici mogu lakše samostalno organizirati potrebno vrijeme za učenje iz pojedinih predmeta prema vlastitim ţeljama, interesu i kompetencijama. Europa ne poznaje jednoznačno tumačenje modela fleksibilnog učenja, ali kad govorimo o strukovnim školama elektrotehnike i računarstva, on podrazumijeva unaprjeĎenje školskog kurikuluma, primjenu meĎunarodno priznatih certifikata, te napredno korištenje e-alata u učenju i poučavanju [4][5]. U fleksibilnom modelu učenja nastavnik ima ulogu trenera koji daje smjernice za učenje, priprema i preporučuje nastavni materijal i na kraju provjerava usvojenost i primjenu naučenog gradiva. Za novi, fleksibilni obrazovni pristup koji mora biti prilagoĎen individualnom pristupu i potrebama svakog pojedinog učenika nastavnik mora biti pripremljen i osposobljen. U ovom članku prikazat ćemo utjecaj provedbe navedenog projekta u Elektrotehničkoj školi u Zagrebu, odnosno kako su nastavnici prihvatili i implementirali aktivnosti i znanja stečena sudjelovanjem u tom projektu te kako i koliko danas koriste različite metode za fleksibilnog učenja te kako to utječe na njihov rad i rad učenika.

II. OPIS PROJEKTA

Projekt se je odvijao u tri dijela (rujan 2010.- rujan 2011.):

- U prvom dijelu projekta (rujan 2010. – listopad 2010.) razvijao se model fleksibilnog učenja u skladu sa specifičnim potrebama partnerskih škola i zahtjevima nacionalnog strateškog i zakonodavnog okvira u sustavu strukovnog obrazovanja.
- U drugom dijelu projekta (listopad 2010. - oţujak 2011.) provedena je edukacija nastavnika na temu provedbe i implementacije fleksibilnog modela učenja, osposobljavanje za neurolingvističko programiranje [6], edukacije za predavače: Training the Trainer – „Solarteur" i „Konneks" te osnovni i napredni New Adobe eLearning Suite trening, u okviru kojih su nastavnici naučili i usavršili izradu digitalnih materijala za učenje, utvrĎivanje i provjeru znanja.
- U trećem dijelu projekta (oţujak 2011. - rujan 2011.) nabavljena je i instalirana oprema, a zatim je proveden prilagoĎeni model fleksibilnog učenja. Planirane i provedene aktivnosti za treći dio projekta su voĎenje nastave po modelu fleksibilnog učenja koje su realizirane po specifičnim zahtjevima i potrebama partnerskih škola s pilot grupama od 30 – 40 učenika, uveden je strukovni karijerni kutić u sektoru elektrotehnike i računarstva, te pilot implementacija obrazovanja koje koristi e-učenje u programima s meĎunarodno priznatim certifikatima.

Sudjelovanjem u projektu Elektrotehnička škola iz Zagreba dobila je novu opremu, dok su njeni djelatnici sudjelovali na mnogo edukacija. Za uspješnu implementaciju fleksibilnog učenja potrebno je i jedno i drugo, čime se stvaraju pretpostavke za moderno strukovno obrazovanje koje je nositelj razvoja i rasta tržišnog gospodarstva. Uvedene su inovacije u postojeće školske programe, osigurana dostupnost međunarodno certificiranih programa kako bi se zadovoljili interesi učenika u skladu sa potrebama domaćeg i europskog tržišta rada. Nastavnicima je pružena prilika za stjecanje međunarodnih certifikata Solarteur i Konnex, specijaliziranim certificiranim programima koji su ponuđeni kao dodatni dijelovi školskog kurikuluma implementirani u model fleksibilnog učenja. Ciljane skupine bile su 30 strukovnih nastavnika i 30 učenika iz Elektrotehničke škole [2].

III. EVALUACIJA PROJEKTA

A. *Evaluacija projekta (2011.g.)*

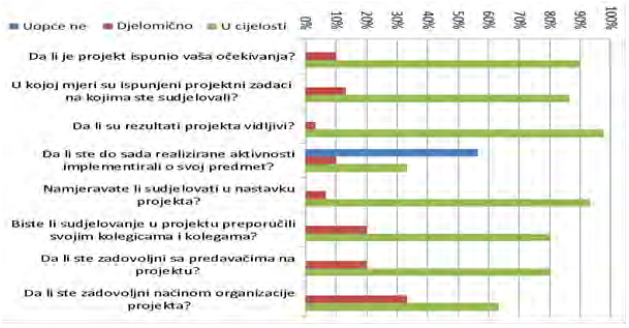
Prva evaluacija projekta provedena je u stručnom radu s međunarodnom recenzijom pod nazivom „Uvođenje fleksibilnog učenja u srednjoškolsko strukovno obrazovanje: komparativna studija dvije škole“ 2011. godine [7].

Istraživanjem su bili obuhvaćeni nastavno osoblje, administrativno osoblje i učenici. Korištene su ankete, a kako bi se dodala interpretativna dimenzija procjene projekta ankete su dopunjene intervjuom i fokus grupom.

Iz rezultata ankete (slika 1.) bilo je vidljivo je da projekt uglavnom ispunio očekivanja nastavnika kao i to da su projektni zadaci u tadašnjem tijeku projekta izvršeni. Većina nastavnika vidjela je rezultate projekta u svojoj školi, a sudjelovanje u projektu preporučili bi i ostalim kolegicama i kolegama. Jedan dio nastavnika (33,34%) nije bio zadovoljan s organizacijom projekta, odnosno terminima u kojima su se pojedine aktivnosti održavale. Najveći problem bio je uočen kod implementacije projektnih aktivnosti u obrazovni proces jer je iz ankete vidljivo da je samo 10 ispitanika (33,33%) u potpunosti primjenjivalo rezultate projektnih aktivnosti u svom razredu. Troje ispitanika (10%) to je činilo djelomično, a 17 ispitanika (56,7%) uopće nije primjenjivalo rezultate projekta u nastavi. U razgovoru s nastavnicima taj problem bio je objašnjen nedovoljnim poznavanjem softvera ili opreme potrebne za primjenu novih metoda u nastavnom procesu.

Evaluacija projekta nad 26 učenika koji su bili podijeljeni u dvije grupe po 13 učenika provedena je metodom fokus grupe.

Provedena rasprava pokazala je da su svi koristili nastavni sadržaj ponuđen preko Interneta. Rado su prihvaćali ovakav način nastave jer je to za njih novi oblik nastave, koji im pomaže pri savladavanju nastavnog sadržaja i pripremanju za provjere znanja.



Slika 1: Rezultati ankete Elektrotehničke škole u 2011. g.

Učenici s boljim uspjehom rekli su da mogu unaprijed proučavati dostupni nastavni materijal i nakon toga pristupiti testovima za provjeru što im omogućava lakše praćenje redovne nastave. Onima s lošijim uspjehom, dodatni materijali za vježbu pomažu da što prije i lakše poprave svoj uspjeh. Također, istaknuli su želju za većim brojem testova u e-učenje sustavu za provjeru znanja, koji bi bili pokazatelj usvojenosti gradiva iz pojedinih predmeta. Učenici su predlagali da se ovakav model nastave uvede u većinu školskih predmeta, te da se omogući unos rezultata tih testova u učenički imenik kao jedan od pokazatelja učenikova postignuća [7].

B. *Evaluacija projekta (2015.g.)*

Evaluacijom su obuhvaćeni učenici i nastavno osoblje Elektrotehničke škole iz Zagreba. Među njima su bili i nastavnici koji nisu sudjelovali u provođenju projekta. Kod nastavnog osoblja koristili smo metodu intervjua i razgovora, a kod učenika metodu fokus grupe.

U provođenju intervjua i u razgovoru s nastavnicima postavili smo sljedeća pitanja:

- Jeste li sudjelovali u IPA projektu "With Flexible Learning Approach and Curricula Innovations towards the Labour Market"?
- Jeste li u nastavu implementirali opremu i alate iz projekta?
- Koriste li učenici „karijerni kutić“ kao pripremu za tržište rada?
- Jesu li u nastavu implementirali znanja i vještine stečene projektnim aktivnostima?
- Primjenjujete li neki oblik fleksibilnog učenja u nastavi?
- Primjenjujete li u nastavi neki oblik e-učenja ili m-učenja?
- Je li vam potrebna dodatna edukacija za uvođenje e-učenja u obrazovni proces?
- Smatrate li korisnim uvođenje e-obrazovanja u klasični oblik nastave?
- Jeste li nastavili učiti kako stvarati i razvijati nove materijale za e-učenje?

Nešto manje od polovine ukupnog broja nastavnog osoblja, njih 30 sudjelovalo je u provođenju projekta i nekih projektnih aktivnosti. Jedna od projektnih

aktivnosti bila je uvođenje „karijernog kutića“, za koji nastavnici ističu da ga učenici uglavnom ne koriste, jer ne postoji kontinuirano praćenje i interakcija između škole i tržišta rada. Ukazuju na potrebu podrške šire društvene zajednice kako bi se uspostavila što bolja suradnja strukovnih škola i srodnih područja iz gospodarstva. Nastavnici koji su prošli obuku za specijalizirane certificirane programe Solarteur i Konnex, implementirali su iste u nastavni proces.

Polovina sudionika projekta koristi neki oblik e-učenja ili m-učenja u svom nastavnom procesu, najčešće koristeći aplikaciju Edmodo, dok neki koriste Moodle i Google Disk. Većina nastavnika koji koriste neki oblik e-učenja su nastavnici struke, dok samo njih nekoliko iz ostalih predmeta (matematika, fizika, hrvatski jezik) primjenjuje e-učenje kao pomoć u nastavi. Za dodatnom edukacijom u primjeni e-alata u nastavi zanimanje pokazuju uglavnom nastavnici koji već djelomično provode neki oblik e-nastave. Izraziti interes za uvođenje e-učenja u redoviti nastavni proces pokazuju mladi nastavnici koji su iskustva o implementaciji i primjeni e-učenja stekli na fakultetima. E-učenje pomaže nastavnicima u interakciji s učenicima te ovakav oblik nastave postaje izrazito korisnim, ali zahtjeva dodatan napor u provođenju i realizaciji nastavnog programa. Ovakav oblik nastave olakšava pripremu učenika za natjecanja, pripremu za maturu kao i svaki drugi oblik dodatne i dopunske nastave. No nije jednostavno pripremiti odgovarajuće materijale za e-učenje: za tzv. prosječne učenike, darovite učenike i za one kojima je potreban individualiziran pristup. Potreban je dodatan napor za provjeru i ispravak stavljenih učeničkih radova, zadaća i vježbi u aplikaciju, jer ako se te zadaće i radovi ne provjere, onda oni nemaju smisla. Iz dosadašnjeg iskustva vidljivo je da su konačni rezultati bolji kod učenika koji su redovito dostavljali zadaće ili zadatke u aplikaciju (ili na Disk) i koji su prihvaćali sugestije. Pojedini učenici su uz redovito pisanje domaćih zadaća također postizali dobre rezultate iako nisu koristili sustav e-učenja.

Korištenjem e-učenja i m-učenja učenicima je pružena dodatna mogućnost praćenja i provjere nastavnih sadržaja, a njihovo korištenje ne bi smjelo biti jedini i obavezni izvor informacija. Osim u nastavi, upotreba e-učenja omogućava i roditeljima pregled literature, zadataka i ostalih nastavnih sadržaja, a time i interakciju između roditelja i nastavnika pojedinih predmeta. Primjena e-učenja i znanja potrebna za upotrebu informacijskih tehnologija pomogli su nastavnicima u prilagodbi i korištenju e-dnevnika.

S učenicima je provedena metoda fokus grupe, a razgovaralo se s najviše 13 učenika u grupi. Prilikom razgovora vodilo se računa da se prilikom postavljanja pitanja ne sugerira ili usmjerava učenika na određeni odgovor. Učenicima su postavljana sljedeća pitanja:

- Primjećujete li model fleksibilnog učenja u nastavi?
- Koristite li neku aplikaciju za e-učenje?
- Koristite li i koliko često dostupne online materijale i testove?
- Koristite li ih paralelno sa redovnom nastavom? Ako ne, zašto?
- Pomaže li vam e-učenje u lakšem praćenju redovne nastave?
- Pomažu li vam testovi u e-učenje sustavu pri rješavanju ispita znanja na redovnoj nastavi?
- Olakšava li vam e-učenje ili m-učenje komunikaciju s nastavnikom izvan redovne nastave?
- Biste li željeli koristiti neki model e-učenja u svim nastavnim predmetima?

Učenici su vrlo pažljivo i zainteresirano sudjelovali u raspravi koja je pokazala da većina koristi aplikacije koje im omogućuju online nastavne materijale. Prihvaćaju ovakav oblik nastave jer im pomaže pri savladavanju redovnog nastavnog programa. Prilikom usmenih i pisanih provjera znanja e-učenje im omogućuje lakšu i bolju pripremu. Učenici s boljim predznanjem i izraženim sposobnostima ističu kako pojedine dijelove nastavnog kurikuluma ne moraju pratiti zajedno sa svojim vršnjacima jer im e-učenje omogućuje dodatne i napredne nastavne sadržaje. Ovakav oblik nastave omogućuje im bolju pripremu za natjecanja i Državnu maturu. Učenicima sa slabijim predznanjem, a time i vjerojatnije lošijim uspjehom, dodatni materijali pomažu u lakšem razumijevanju i savladavanju obaveznog nastavnog sadržaja. Želja im je da se nastava temeljena na e-učenju, kao dodatni oblik nastave, primjenjuje u svim nastavnim predmeta gdje za to postoji mogućnost, kako bi imali pristup nastavnim sadržajima tijekom cijeloga dana. E-učenje omogućuje fleksibilnost u nastavi, jer fleksibilni oblik nastave pruža nastavnicima mogućnost individualnog pristupa svakom pojedinom učeniku prema njegovim željama, interesu i sposobnostima za pojedini dio gradiva.

Osim e-učenja, uporaba m-učenja učenicima omogućuje interakciju s predmetnim nastavnikom i međusobnu komunikaciju izvan vremena redovne nastave, pa čak i 24 sata. Zadovoljstvo im pruža mogućnost međusobne razmjene edukativnih materijala, kao i grupnog rješavanja problemskih zadataka. Puno im znači brz odgovor nastavnika ili komentar na neki predani zadatak ili zadaću. Žao im je što se ocjene iz online testova i aktivnosti na aplikacijama za e-učenje ne mogu vrednovati kao i klasični oblik provjere znanja.

IV. RASPRAVA

Iz provedenog intervjua i dobivenih odgovora može se uočiti da većina nastavnika strukovnih predmeta primjenjuje neki oblik e-učenja implementirajući ga u klasični oblik nastave. Oprema, edukacija i pozitivno

ozračje pozitivno su utjecali na nastavnike u implementaciji novih e-alata u nastavni proces. Na osnovu opreme (za Solarteur i Konnex) i vlastitog usavršavanja, kombinirajući klasični oblik nastave i e-učenje, nastavnici su uz pomoć stečenih znanja nove sadržaje implementirali u postojeće predmete, izborne predmete i završne radove. Zbog dobrog poznavanja i praćenja razvoja informacijskih tehnologija, nastavnicima struke bilo je puno lakše implementirati i u klasičnoj nastavi iskoristiti sve prednosti i potencijale aplikacija za e-učenje.

Upravo su nastavnici struke, ali i vodstvo škole, istaknuti kao motor pokretanja i uvođenja novih tehnologija u nastavni proces. Oni su zaslužni za stvaranje modernog, digitalnog ozračja u nastavnom procesu. Korištenje digitalnih materijala, osim toga, potiče nastavnike na sve napredniju primjenu istih jer stalno traže i otkrivaju nove alate ili načine za razvoj i napredak obrazovnog pristupa. Nastavnici stalno razvijaju nove i proširuju svoje postojeće edukativne materijale. Primjenom novih tehnologija i e-alata u obrazovanju, kako bi se u potpunost razvio i održao fleksibilni obrazovni model, ukazala se potreba za cjelodnevnom i trenutnom komunikacijom između nastavnika i učenika izvan učionice. U tu svrhu nastavnici, koji to žele; uglavnom koriste mobilne telefone tj. m-učenje kao sredstvo za dobivanje i davanje kratkih i hitnih informacija ili globalan pregled nekog rada.

Pokazalo se da su svi nastavnici koji koriste neki oblik e-učenja bez poteškoća prihvatili i savladali izazove vezane uz uvođenje e-dnevnika. Nastavnici koji nastavu drže na tradicionalni način, uglavnom su željeli zadržati razrednu knjigu - klasični način evidencije i praćenja učenika. Veliki interes i zalaganje za drugačiji oblik nastave, potpomognut upotrebom nekih oblika e-alata pokazuju mladi nastavnici, educirani primjenom modernih nastavnih tehnologija. Oni svoje znanje i iskustvo o primjeni i implementaciji novih e-alata prenose na ostale nastavnike. Primijećeno je da osim nastavnika struke, jedan dio, iako manji broj, nastavnika ostalih predmeta (matematika, fizika, biologija, …) koristi napredne tehnologije e-učenja u nastavi. To su uglavnom nastavnici koji su sudjelovali na dodatnim edukacijama i seminarima o upotrebi informacijskih tehnologija kao alatu za primjenu e-učenja ili su sami učili, opet korištenjem digitalnih materijala.

Ako se analizira dobna granica onda se može uočiti da nastavnici s nekoliko godina do mirovine, iako ima iznimaka, najtеže prihvaćaju nove obrazovne metode. Oni u pravilu provode klasičan, tradicionalan oblik poučavanja, koristeći pritom ploču, kredu i udžbenik kao osnovno pomagalo u nastavi. Razlog za to je nerado sudjelovanje na edukacijama zbog nesnalaženja na novim sklopovskim i programskim tehnologijama.

V. ZAKLJUČAK

Iako je u Elektrotehničkoj školi od završetka projekta prošlo više od pet godina, evaluacija je pokazala da su rezultati njegove provedbe ostavili značajan trag u odvijanju obrazovnog procesa. Kako bi rezultati projekta ostali vidljivi, uprava škole i nastavno osoblje stvarali su potporu i pozitivno okruženje za permanentnu edukaciju i prihvaćanje novih oblika nastave. Rezultati takvog pristupa vidljivi su kroz poboljšane materijalne resurse, nove nastavne programe i e-učenje kao potporu klasičnoj nastavi. Pokazalo se da poboljšanje i modernizacija opreme i ostalih materijalnih resursa u kombinaciji sa stručnom edukacijom na nastavnike djeluje motivirajuće, potičući ih da i dalje prate nove tehnološke izazove, te iste nastoje primijeniti u nastavi. Primjer implementacije novih aplikacija je uvođenje e-dnevnika. Pokazalo se da nastavnici koji su već implementirali neki oblik e-učenja u svojoj nastavi puno brže prihvaćaju i usvajaju promjene koje nezaustavljivo mijenjaju dosadašnji obrazovni pristup.

Fleksibilni model učenja, e-učenje i m-učenje pokazali su veliku međusobnu korelaciju. Gotovo je nemoguće primijeniti fleksibilni obrazovni model bez značajne podrške e-učenja, za što je potrebno pripremiti i educirati nastavno osoblje. Za potpunu implementaciju fleksibilnog učenja, osim promjena u školi, potrebno je izvršiti društvene i gospodarske promjene i prilagodbe kako bi obrazovni sustav, posebno strukovne škole, bio pripremljen za izazove koje pred njega postavlja zahtjevno i dinamično tržište rada. Između škola i tržišta rada trenutno ne postoji dovoljna interakcija, za što je potrebna potpora šire društvene zajednice.

Primjer Elektrotehničke škole pokazuje da će škole s iskustvom u implementaciji novih obrazovnih alata za provedbu, praćenje i vrednovanje obrazovnih postignuća puno lakše prihvaćati nove oblike nastave u budućnosti.

LITERATURA

[1] Vlada Republike Hrvatske, MZOŠ: Strategija razvoja strukovnog obrazovanja u Republici Hrvatskoj 2008. – 2013., Zagreb, Hrvatska, srpanj 2008.

[2] IPA Projekt, dostupno na http://www.ss-elektrotehnicka-zg.skole.hr/ipa (pristupljeno 03. veljače 2016.)

[3] Znanstveni rad s međunarodnom recenzijom dostupan na https://bib.irb.hr/prikazi-rad?&rad=529348 (pristupljeno 20. siječnja 2016.)

[4] W. Wade Flexible Learning in Higher Education, London and Philadelphia: Kogan Page, 1994.

[5] P. Race: 500 Tips for Open and Flexible Learning, London: Kogan Page, 1998.

[6] NLP Hrvatska, dostupno na http://www.nlp.com.hr/ (pristupljeno 04. veljače 2016.)

[7] Stručni rad s međunarodnom recenzijom dostupan na https://bib.irb.hr/prikazi-rad?&rad=534314 (pristupljeno 05. veljače 2016.)

Preporuke i primjeri dobre prakse e-učenja u hrvatskom visokom školstvu

Dino Junaković*, Inka Paćelat* i Frane Urem*
* Veleučilište u Šibeniku/Odjel menadžmenta, Šibenik, Hrvatska
neo_nexus@msn.com, pacelat.inka@gmail.com, frane.urem@vus.hr

Sažetak - Danas postoje brojna programska rješenja za podršku e-učenju. Iako su informacijske tehnologije vrlo bitne u uspješnoj primjeni takvih sustava, važno je i pravovremeno donošenje obrazovnih strategija i smjernica koje osiguravaju širu primjenu e-učenja i vrednovanje studijskih programa koji se u potpunosti izvode učenjem na daljinu. U ovom radu je dan pregled određenih preporuka i smjernica za e-učenje kao i najčešćih programskih rješenja za e-učenje koja se koriste u hrvatskom visokom obrazovanju.

I. UVOD

Nagli razvoj tehnologija otvorenog računarstva je omogućio različite primjene e-učenja u obrazovanju. Danas je e-učenje u svjetskom i hrvatskom visokom obrazovanju uobičajena nastavna metoda i u obliku učenja na daljinu se sve češće kombinira s klasičnim oblicima održavanja nastave. Određena istraživanja [1] i preporuke [2] u prostoru europskog visokog obrazovanja naglašavaju potrebu donošenja nacionalnih strategija za e-učenje kako bi se što veći broj nastavnika i studenata uključio u takve oblike obrazovanja. Ministarstvo znanosti obrazovanja i sporta (MZOS) je prepoznalo potencijal e-učenja i uvrstilo ga je u prioritete Strategije obrazovanja, znanosti i tehnologije [5]. U ovom radu je također naveden pregled nekoliko programskih sustava za e-učenje koji su posebno popularni na hrvatskim visokim učilištima. Neka od navedenih rješenja se uspješno koriste i na Veleučilištu u Šibeniku.

II. OSNOVNI POJMOVI

Pojam e-učenje općenito predstavlja bilo kakav oblik učenja ili podučavanja kod kojeg se koriste informacijske i komunikacijske tehnologije (ICT) [1]. Međutim, danas se e-učenje često izjednačava s tzv. mješovitim učenjem (engl. blended learning) koje podrazumijeva obrazovne programe u kojima se nastava izvodi uobičajenim postupcima u kombinaciji s tzv. učenjem na daljinu (engl. online learning) koje nastavne sadržaje isporučuje putem interneta (udaljena predavanja, digitalizirani nastavni materijali dostupni na web sjedištima itd.) [1]. Također, e-učenje sve više postaje prostor besplatnih tzv. MOOC (engl. Massive Open Online Course) edukacijskih programa koji su najčešće besplatni i dostupni velikom broju polaznika jer se u potpunosti izvode učenjem na daljinu [1]. Ovakvi programi obrazovanja postaju novi izazov visokoškolskim ustanovama u smislu priznavanja tako ostvarenih ishoda učenja.

III. E-UČENJE U EUROPSKOM VISOKOM ŠKOLSTVU

Stanje primjene e-učenja u europskom visokom školstvu je detaljno opisano u studiji [1] koju je izradila Europska udruga sveučilišta (engl. European University Association – EUA), a provedena je tijekom 2013. i 2014. godine na uzorku od 249 visokih učilišta iz 36 europskih država. U navedenom uzorku anketirano je 228 sveučilišta (engl. university) i 21 veleučilište (engl. university of apllied sciences). U anketi su sudjelovala dva hrvatska sveučilišta (Sveučilište u Zagrebu i Sveučilište u Rijeci). Rezultati studije su objavljeni u prosincu 2014. godine.

U studiji [1] je posebno razmatrano stanje nacionalnih politika za e-učenje i utvrđeno je kako relativno mali broj europskih država ima već utvrđene strategije za e-učenje u visokom školstvu (samo 16% od ispitanih) iako su preporuke za izradu takvih strategija dane od strane Europske komisije 2010. godine [2]. Međutim, u većem broju država su takve strategije u postupku izrade i očekuje se njihova puna primjena do 2020. godine (40% od ispitanih). U nedostatku nacionalnih politika, visoka učilišta su ipak donosila vlastite politike za e-učenje pa ih 49% ispitanih ustanova već primjenjuje, a 26% je u postupku donošenja strategija. Aktivan pristup europskih visokih učilišta očito daje dobre indikatore primjene e-učenja jer čak 82% ispitanih ustanova studentima nudi različite oblike učenja na daljinu, a 91% primjenjuje model mješovitog učenja. Posebno je značajan pokazatelj prema kojem 40% ispitanih ustanova primjenjuje nekakav oblik e-učenja u suradnji s drugim visokoškolskim ustanovama. Međutim, ovako dobre pokazatelje je potrebno uzeti s određenom rezervom jer kada se prema [1] analizira stvaran broj studenata koji su uključeni u različita područja e-učenja, pokazatelji su nešto lošiji. Ukupni rezultati iz [1] upućuju na zaključak da 55% ispitanih ustanova procjenjuje kako manje od 50% njihovih studenata ima pristup e-učenju. Prema [1] najbolje rezultate u uključenosti većeg broja studenata bilježe Njemačka, Švicarska i Velika Britanija. Također se uočava povezanost između uspostavljene strategije i raširenosti pristupa e-učenju od strane studenata jer 45% ustanova s aktivnom strategijom bilježi uključenost više od 50% studenata u proces e-učenja.

Određene visokoškolske ustanove iz [1] dodatno definiraju ciljne skupine studenata kojima su obrazovni programi za učenje na daljinu namijenjeni : studenti koji ne mogu prisustvovati nastavi iz opravdanih razloga, studenti na poslijediplomskim studijskim programima,

studenti na razmjenama unutar programa mobilnosti, studenti koji studiraju uz rad, ali i bivši studenti te nezaposleni koji žele usavršavati svoje vještine kako bi bili konkurentniji na tržištu rada.

IV. SMJERNICE UVOĐENJA E-UČENJA U HRVATSKOM VISOKOM ŠKOLSTVU

Republika Hrvatska (RH) je prvi puta mogućnost održavanja studijskih programa učenjem na daljinu regulirala člankom 7. i člankom 79. stavkom 2. Zakona o znanstvenoj djelatnosti i visokom obrazovanju [2]. Evaluacija takvih studijskih programa je prepuštena Nacionalnom vijeću za znanost i visoko obrazovanje koje je na svojoj 95. sjednici održanoj 13. veljače 2013. donijelo Kriterije za vrednovanje online studija [3]. Iste godine Vlada RH pokreće izradu Strategije obrazovanja, znanosti i tehnologije koja u radnoj verziji [4] u točki 1.10 definira mjere, nadležnosti i rokove provedbe za razvoj digitalnih obrazovnih sadržaja, alata i metoda korištenja ICT tehnologija u učenju i poučavanju. U posebnom poglavlju Strategije [4] se dodatno definira provedba cjelovite informatizacije odgojnoobrazovnih procesa s potrebnom komunikacijskom infrastrukturom. Sabor RH je u listopadu 2014. godine usvojio konačnu verziju Strategije obrazovanja, znanosti i tehnologije koja je dostupna na [5].

V. PRIMJERI PROGRAMSKIH RJEŠENJA ZA E-UČENJE U HRVATSKOM VISOKOM ŠKOLSTVU

Većina visokoškolskih ustanova u RH primjenjuje mješoviti model učenja u kojem se kombinira standardna nastava i učenje na daljinu. Područje osnivanja MOOC tečajeva zasada predvodi samo CARNET kroz tzv. ISE (engl. Inspiring Science Education) MOOC program koji je namijenjen osposobljavanju nastavnika koji žele takve programe realizirati na vlastitim ustanovama prvenstveno iz STEM (engl. Science, Technology, Engineering, and Mathematics) područja.

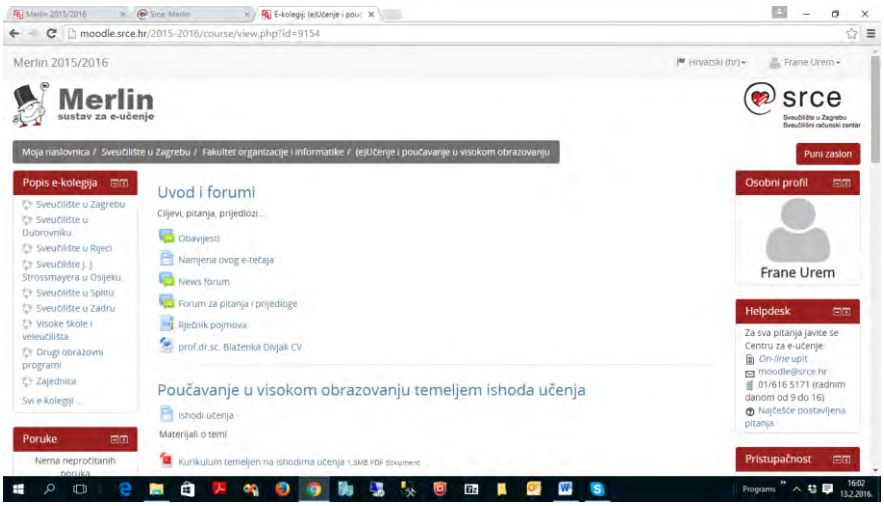
Veliki doprinos razvoju učenja na daljinu u hrvatskom visokom školstvu ostvaruje Centar za e-učenje koji je formiran unutar Sveučilišnog računskog centra (Srce) u Zagrebu. Iako Srce svojom djelatnošću osigurava snažnu ICT infrastrukturu prvenstveno za potrebe Sveučilišta u Zagrebu, tijekom godina su unutar centra razvijeni i određeni sustavi za potrebe cjelokupne akademske i istraživačke zajednice u Hrvatskoj. Jedna od takvih je platforma za e-učenje poznatija pod nazivom Merlin [6] koji se zasniva na sustavu otvorenog koda (engl. open source) Moodle [7]. Različite verzije Moodle sustava koriste se na više od 65 tisuća web sjedišta u 224 zemlje [7]. Broj ukupnih korisnika prelazi 80 milijuna, od čega je preko milijun nastavnika [7]. Sustav za e-učenje Merlin sastoji se od osnovnog sustava za e-učenje, sustava za webinare (engl. Webinar – web based seminar) i tzv. e-portfolio sustava. Merlin sustav trenutno u RH koristi više od 46000 korisnika, a na sustavu je aktivno više o 2200 e-kolegija [6].

Osnovni Merlin sustav za e-učenje sadrži alate koji omogućavaju da nastavnik sve nastavne materijale (tekstne, slikovne i audiovizualne) jednostavno postavlja u sustav i tako ih čini uvijek raspoloživima studentima. Sučelje osnovnog Merlin sustava za e-učenje je prikazano na slici 1. Dodatno su omogućeni moduli za komunikaciju među studentima i nastavnicima kao i alati za provjeru znanja i ocjenjivanje studenata. Merlin sustav za webinare [8] omogućava održavanje seminara, prezentacija, predavanja ili radionica koje se prenose pomoću interneta u stvarnom vremenu, a uključuje video, audio i tekstovnu komunikaciju između sudionika. Merlin e-portfolio [9] je sustav je utemeljen na sustavu otvorenoga koda Mahara [10]. Predstavlja sustav socijalnoga umrežavanja za pripadnike akademske zajednice (studenti, nastavnici i visokoškolske ustanove) na kojem je moguća izrada osobnih stranica, prikupljanje radova, projekata, dokumenata, slika te prikaz takvih aktivnosti na razini čitave visokoškolske ustanove.

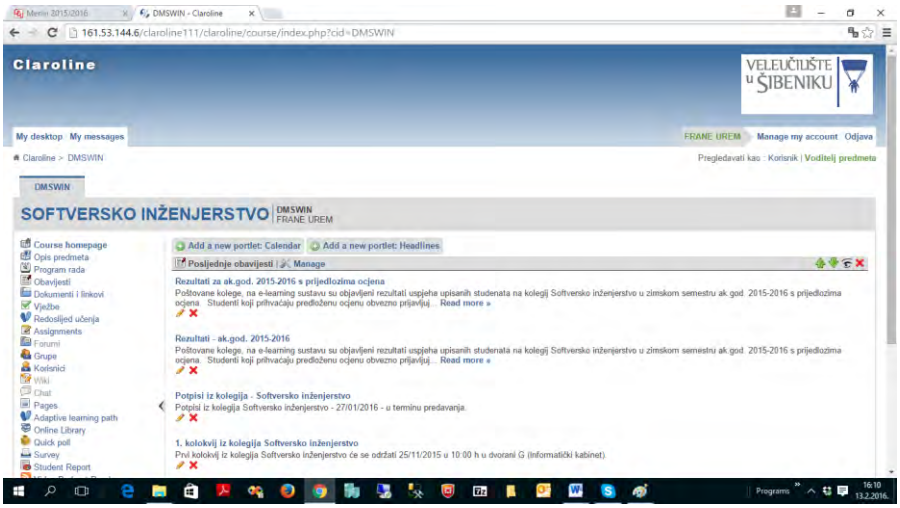
Osim rješenja zasnovanih na Moodle sustavu, među korisnicima iz visokog školstva u RH je popularan sustav za e-učenje Claroline [11] koji se koristi i na Veleučilištu u Šibeniku. Sustav je početno razvijen 2001. godine u Belgiji na Katoličkom sveučilištu u Louvainu. Glavna prednost ovog sustava (u usporedbi sa sličnim rješenjima) je jednostavno i intuitivno sučelje za rad koje nastavniku omogućava postavljanje nastavnih i ispitnih materijala na sustav. Na Veleučilištu u Šibeniku se Claroline sustav koristi na svim kolegijima (objave rezultata kolokvija i ispita, informacije u vezi održavanja nastave, objave nastavnih materijala i izdanja Veleučilišta u Šibeniku, izrada i korištenje ispita za provjeru znanja), ali i za internu objavu svih dokumenata i aktivnosti u sklopu uvedenog sustav kvalitete u visokom obrazovanju ISO 9001:2008. Claroline je preveden na 35 jezika i trenutno se koristi u više od 100 država diljem svijeta [11]. Osnovno sučelje Claroline sustava primjenjenog na Veleučilištu u Šibeniku je prikazano na slici 2.

Posebno se može izdvojiti sustav e-učenja LMS Canvas by Infrastructure [14] koji se koristi na Sveučilištu u Rijeci, a detaljnije je njegova primjena opisana u [15].

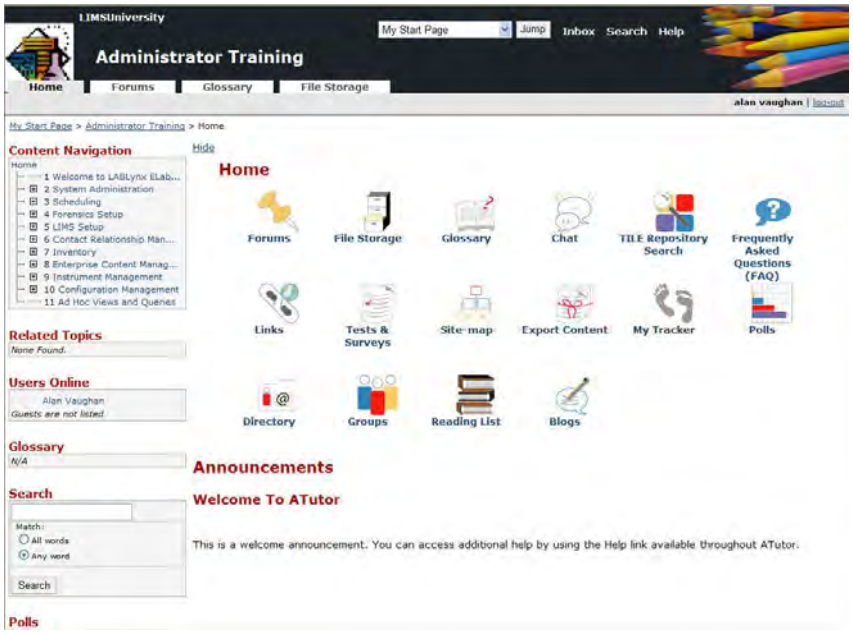
Poseban izazov u uvođenju sustava za učenje na daljinu predstavlja sudjelovanje studenata s invaliditetom, a sustav ATutor [12] predstavlja jedno od najboljih rješenja za osobe s invaliditetom koje je zasnovano na GPL (engl. General Public License) besplatnoj licenci. Prema izvoru [13] osobe s invaliditetom mogu uspješno kreirati vlastite obrazovne sadržaje u ATutoru što ih čini aktivnim sudionikom u razvoju sustava za udaljeno učenje. Također, sustav ATutor podržava stroge norme pristupačnosti elektroničkih informacija za osobe s invaliditetom kao što su: W3C WCAG 1.0 , W3C WCAG 2.0, W3C ATAG 2.0, US Section 508, Stanca Act, ISO FDIS 24751. Navedene norme se odnose na pristupačnosti sadržaja na mrežnim stranicama, pristupačnosti alata za kreiranje sadržaja, individualizacije sučelja i formata dokumenata i informacija, inter-operabilnosti meta-podataka kao i na direktne zahtjeve zakona u SAD i Italiji koji reguliraju pitanja pristupa nosti mrežnih stranica. Veleučilište u Šibeniku planira paralelno uvođenje ATutor programskog rješenja s postojećim Claroline sustavom za udaljeno učenje kako bi olakšalo studij studentima s invaliditetom od akademske godine 2016/2017. Osnovno sučelje ATutor sustava je prikazano na slici 3.



Slika 1. Osnovno sučelje Merlin sustava za e-učenje



Slika 2. Nastavničko sučelje za rad na Claroline sustavu Veleučilišta u Šibeniku



Slika 3. Administiranje kolegija na ATutor sustavu

VI. ZAKLJUČAK

Hrvatsko visoko školstvo ostvaruje dobre rezultate u primjeni ICT tehnologija za potrebe učenja na daljinu. U takvim aktivnostima hrvatska visoka učilišta imaju veliku infrastrukturnu pomoć od strane referentnih centara za informacijsku potporu kao što su CARNET i Srce. Usvajanjem Strategije obrazovanja, znanosti i tehnologije takva potpora će u budućnosti sigurno jačati. U aktivnostima e-učenja je potrebno prepoznati osobe s invaliditetom i promicati uvođenje sustava za udaljeno učenje koja im olakšavaju pristup digitalnim sadržajima, ali i omogućavaju da aktivno stvaraju obrazovne sadržaje. Poseban izazov u budućnosti će predstavljati izvođenje MOOC programa jer omogućavaju i pristup nekim skupinama ljudi koji nisu u mogućnosti pohađati klasičnu nastavu. U takvim aktivnostima će dostupnost odgovarajućih tehnoloških rješenja biti najmanji problem jer je većina takvih sustava već danas besplatna za korištenje i primjenjiva na postojećoj infrastrukturi u viskokoškolskim ustanovama. Međutim, ostaju otvoreni problemi kvalitete takvih nastavnih sadržaja, provedbe ispita u kontroliranim uvjetima i izjednačavanje s klasičnom nastavom (praktična nastava, mentoriranje).

LITERATURA

[1] M. Gaebel, V. Kupriyanova, R. Morais, E. Colucci, "E-learning in European Higher Education Institutions", European University Association, EUA Publications 2014, 2014., dostupno na: http://www.eua.be/Libraries/publication/e-learning_survey [1. ožujka 2016.]

[2] Communication by the European Commission, "Opening up Education", 2013, dostupno na http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52013DC0654&from=EN [1. ožujka 2016.]

[3] Narodne novine, "Zakon o izmjenama i dopunama zakona o znanstvenoj djelatnosti i visokom obrazovanju", br. 139, 2013., dostupno na: http://narodne-novine.nn.hr/clanci/sluzbeni/2013_07_94_2132.html [1. ožujka 2016.]

[4] Vlada RH, "Strategija obrazovanja znanosti i tehnologije (radni materijal)", 2013, dostupno na: https://vlada.gov.hr/UserDocsImages/ZPPI/Strategije/Strategija%20obrazovanja,%20znanosti%20i%20tehnologije%20(radni%20materijal).pdf [1. ožujka 2016.]

[5] Narodne novine, " Strategija obrazovanja znanosti i tehnologije", br. 124, 2014, dostupno na: http://narodne-novine.nn.hr/clanci/sluzbeni/2014_10_124_2364.html [1. ožujka 2016.]

[6] Srce – Sveučilišni računski centar Sveučilišta u Zagrebu, "Merlin sustav za e-učenje", 2015, dostupno na: http://moodle.srce.hr/2015-2016/ [1. ožujka 2016.]

[7] Moodle Pty Ltd, „The moodle Project“, 2015., dostupno na: https://moodle.org/ [1. ožujka 2016.]

[8] Srce – Sveučilišni računski centar Sveučilišta u Zagrebu, "Sustav za webinare", 2014, dostupno na: http://www.srce.unizg.hr/webinari-i-videokonferencije/webinari/ [1. ožujka 2016.]

[9] Srce – Sveučilišni računski centar Sveučilišta u Zagrebu, "Merlin e-portfolio", 2014, dostupno na: https://moodle.srce.hr/eportfolio [1. ožujka 2016.]

[10] Mahara.Org ,"Open source eportfolios", 2015, dostupno na: https://mahara.org [1. ožujka 2016.]

[11] Claroline Connect, 2016, dostupno na: http://www.claroline.net

[12] ATutor, "Learning Management Tools", dostupno na: http://www.atutor.ca/index.php [1. ožujka 2016.]

[13] V.Vučić, "Inicijative za promicanje pristupačnosti elektroničkim informacijama osobama s invaliditetom", Hrvatska revija za rehabilitacijska istraživanja 2009, Vol 45, br. 2, dostupno na: http://hrcak.srce.hr/file/79516 [1. ožujka 2016.]

[14] LMS Canvas by Infrastructure ,2016, dostupno na https://www.canvaslms.com/k-12 [1. ožujka 2016.]

[15] E Kusen, N Hoic-Bozic, "Use of new technology in higher education: A migration to a cloud-based learning platform", The 2nd International Workshop on Learning for Education in Cloud, 2014

Ilustracija primjene novog Kurikuluma iz predmeta Informatika i to domene - Računalno razmišljanje i programiranje na primjeru metode Početnica Mema za prvi razred osnovne škole

M. Čičin-Šain*, S. Babić ** and L. Kralj ***
* Društvo Kibewrnetičara, Rijeka, Hrvatska
** Veleučilište u Rijeci, Rijeka, Hrvatska
*** Osnovna škola Veliki Bukovec, Veliki Bukovec, Hrvatska
marinacs@efri.hr

Sažetak - **Računalno razmišljanje i programiranje dio je informatičkog kolegija koji se predlaže za sva četiri razreda osnovne škole prema novom Kurikulumu. Metoda Početnica Mema pruža mogućnost da se ta nastava odvija sasvim prilagođena uzrastu djece uz veliku potporu učiteljici ili učitelju. U ovome radu dana je ilustracija primjene novog Kurikuluma na primjeru metode Početnica Mema za prvi razred osnovne škole.**

I. UVOD

Rijetko kada su nastavnici na svim razinama, od učitelja ili učiteljica osnovne škole do profesora na fakultetu bili u prilici na tako direktan način sudjelovati u kreiranju novih kurikuluma kao što je to upravo sada u Hrvatskoj slučaj. Kurikulumi bi trebali biti okviri koji nastavnike usmjeravaju što i kako podučavati i što trebaju očekivati kao rezulat podučavanja. Nažalost, često se događa da su kurikulumi dobrom nastavniku teret, jer silom prilika zastarjevaju, pa okviri postaju pretijesni, a sadržaji besmisleni, a lošem nastavniku isprika, da svoj posao odradi rutinski bez trunka kreativnosti. Kada se radi o informatici onda je to posebno jako istaknuto, jer se u tom području stvari mijenjaju brže nego u svim ostalim predmetima zajedno. Zato veseli da je u prijedlogu novog Kurikuluma iz informatike [1] za niže razrede osnovne škole ostavljeno dosta prostora da se nastava može odvijati na način koji je za učiteljice ili učitelje prihvatljiv, a za učenike koristan. Radi se o modernom pristupu koji nastavu informatike nudi kroz četiri domene [1]:

- E-društvo;
- Digitalna pismenost i komunikacija;
- Informacije i digitalna tehnologija;
- Računalno razmišljanje i programiranje

koje su prikazane na slici 1.



Slika 1. Četiri domene predmeta Informatika [1]

Od svih domena, Računalno razmišljanje i programiranje je poseban izazov. Učenici su silno motivirani kada se računalom komunicira ili se pomoću njega izrađuju zgodni materijali poput plakata, čestitaka ili video uradaka. Ali kada se počne programirati u nekom od programskih jezika nastaju problemi.

Nastavnici koji predaju programiranje na svim razinama sreću se sa problemom da neki učenici ili studenti programiranje teško prihvaćaju i razumiju. Pokušavaju se naći razni pristupi, koji bi učenike i studente motivirali za programiranje [6], a oni opet iz raznih razloga propadaju, jer do sada nijedna metoda programiranje nije učinila lakšim. Uvijek doduše postoje talentirani đaci i njihovi požrtvovni nastavnici koji pobjeđuju na našim, pa i svjetskim natjecanjima, ali što je

sa preostalom većinom? Danas svako dijete u školi nauči čitati i pisati, računati i crtati. Zašto svako dijete ne nauči programirati barem na osnovnoj razini?

Ono što programiranje može učiniti dostupnim svakom prosječnom čovjeku je upravo ono što se danas predlaže novim Kurikulumom iz informatike [1]: s programiranjem treba početi već od prvog razreda osnovne škole. Kako i autorice metode pomoću Meme istu stvar tvrde već decenijama, rado se pristupilo pokušaju da se upravo pomoću Meme ilustrira primjena novog Kurikuluma. Možda za slijedećih četrdeset godina programiranje postane i obavezan predmet za osnovne škole.

II. METODA POČETNICA MEMA

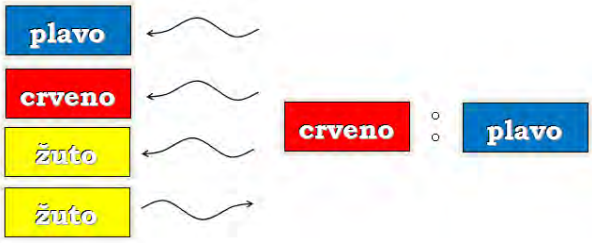
Metoda koja koristi igračku Memu, koja se koristi od 1980 [2] i koja je opisana još 1990 [3] predložena je za upotrebu u Kurikulumu za kolegij informatike za osnovne škole [4]. Interesantno je da je metoda koja je predložena pred četrdeset godina danas još uvijek prihvatljiva. Kako je to moguće, kada je poznato da u informatici stvari brzo zastarjevaju? Radi se o tome da se mijenjaju tehnička rješenja i sučelja, kako se računalu pristupa, a osnovni principi se ne mijenjaju još od Charlesa Babbagea iz XIX stoljeća. Kad se radi o programiranju te osnovne invarijante su:

- Program;
- Naredba;
- Redoslijed naredbi;
- Varijabla;
- Adresa varijable;
- Sadržaj varijable;
- Promjena sadržaja varijable;
- Konstanta;
- i malo kasnije: Odluka.

U metodi Početnica Mema pomoću Meme predviđeno je da nastavu izvodi učitelj ili učiteljica razredne nastave, sve upute su detaljno dane u metodi [2] [3]. Koristi se igračka Mema slika 2. i programi dani grafičkim pseudokodom (slika 3).



Slika 2. Igračka Mema za prvi razred osnovne škole (po ideji M. Čičin –Šain)



Slika 3. Program za dijeljenje dan grafičkim pseudokodom (po ideji M.Čičin-Šain)

Neke metode, npr. programski jezik Logo ili edukativne igrice poput Code Combat ili Lighbot pomažu u razumijevanju što je program, što je naredba i da je redoslijed naredbi važan, ali varijablu i sadržaj varijable ne objašnjavaju, a primjerice Lightboat ima i ograničenje na broj naredbi u procedurama, koje pravi programski jezici nemaju, pa im je svrha upitna.

Prednost metode koja koristi Memu je u tome da na intuitivan način objašnjava što se događa sa varijablama, a naredbe se ponašaju kao i prave naredbe u pravim programima. Djeca igraju ulogu procesora, a da bi ju odigrala moraju usvojiti način na koji on radi.

III. PRIMJERI PRIMJENE

Prema Kurikulumu informatike [1] za prvi razred osnovne škole, domena Računalno razmišljanje i programiranje, očekuju se ishodi učenja prikazani u tablici 1 i 2 [1].

TABLICA 1. ISHOD UČENJA B.1.1. I RAZINE ZNANJA [1]

Ishod učenja B.1.1.: Nakon prve godine učenja predmeta informatika u domeni računalno razmišljanje i programiranje-učenik rješava jednostavan logički zadatak
Razrada ishoda: Učenik prepoznaje problem i smišlja mogućnosti rješavanjajednostavnoga logičkog zadatka. Uz učiteljevu pomoć rješava jednostavne logičke zadatke. Pokazuje znatiželju i propituje putove dolaska do rješenja jednostavnoga logičkog zadatka. Potiče sebe i druge na ustrajnost u pronalaženju rješenja.
Razina usvojenosti zadovoljavajuća: Učenik uz pomoć učitelja prepoznaje problem, slijedeći upute rješava jednostavni logički zadatak.
Razina usvojenosti dobra: Uz pomoć učitelja otkriva mogućnosti rješavanja jednostavnoga logičkog zadatka.
Razina usvojenosti vrlo dobra: Učenik uz povremenu pomoć učitelja otkriva mogućnosti rješavanja jednostavnoga logičkog zadatka.
Razina usvojenosti iznimna: Učenik samostalno preispituje putove dolaska do rješenja jednostavnoga logičkog zadatka. Potiče sebe i druge na ustrajnost u pronalaženju rješenja.

TABLICA 2. ISHODI UČENJA B.1.2. I RAZINE ZNANJA [1]

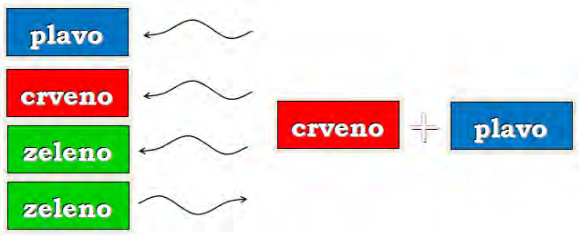
Ishod učenja B.1.2.: Nakon prve godine učenja predmeta informatika u domeni računalno razmišljanje i programiranje-učenik prati i prikazuje slijed koraka potrebnih za rješavanje nekog jednotavnog zadatka.
Razrada ishoda: Učenik prepoznaje jednostavni zadatak (problem) iz svakodnevnoga života. Analizira zadatak (različitinačini rješavanja), opisuje korake njegova rješavanja. Prikazuje korake rješavanja zadatka (slikom, riječima).
Razina usvojenosti zadovoljavajuća: Učenik uz pomoć učitelja prepoznaje problem i opisuje kako bi ga riješio.
Razina usvojenosti dobra: Uz povremeneu pomoć učitelja otkriva korake rješavanja jednostavnog zadatka i prikazuje rješenje.
Razina usvojenosti vrlo dobra: Samostalno prikazuje i prati slijed odvijanja koraka potrebnih za rješavanje nekoga jednostavnog zadatka.
Razina usvojenosti iznimna: Učenik uz pomoć učitelja analizira rješenje jednostavnoga zadatka i vrednuje ga.

Dakle, učenik će biti osposobljen da:

- rješava jednostavan logički zadatak (ishod učenja B. 1. 1. [1])
- prati i prikazuje slijed koraka potrebnih za rješavanje nekoga jednostavnog zadatka (ishod učenja B. 1. 2. [1]).

Na primjer, ishodi B.1.1. i B.1.2. mogu se postići primjenom metode Početnica Mema na sljedeći način.

A. Primjer 1. za realizaciju definiranog ishoda B1.2.: Što radi sljedeći program prikazan na slici 4.?



Slika 4. Program za zbrajanje dan grafičkim pseudokodom (po ideji M.Čičin-Šain)

Ovdje se radi o zadatku osnovanom na principu prepoznavanja. Prve dvije naredbe traže da se pretinci Meme pune bilo kojim brojem po izboru učenika. *Pitanje:* O kojim se pretincima radi? *Odgovor:* Radi se o plavom i crvenom pretincu. Treća naredba traži da se neki broj stavi u zeleni pretinac. *Pitanje:* Koji broj treba staviti u zeleni pretinac? *Odgovor:* Onaj broj koji se dobije kada se zbroje sadržaji crvenog i plavog pretinca. *Pitanje:* Da li se crveni i plavi pretinac trebaju isprazniti? *Odgovor:* Ne, sadržaj u njima se ne mijenja. *Pitanje:* Što radi četvrta naredba? *Odgovor:* Ispisuje se nekamo (na ploču ili bilježnicu) sadržaj zelenog pretinca. *Pitanje:* Da li se pri tom zeleni pretinac prazni? *Odgovor:* Ne, sadržaj ostaje nepromijenjen.

B. Primjer 2. za realizaciju definiranog ishoda B1.1.: Nacrtaj naredbu kojom se prazni zeleni pretinac.

Pitanje: Možemo li ovom naredbom (slika 5.) isprazniti zeleni pretinac?



Slika 5. Naredba za ispis sadržaja zelenog pretinca (po ideji M.Čičin-Šain)

Odgovor: Ne. Ova naredba ispisuje sadržaj zelenog pretinca, a sadržaj pretinca se ne mijenja.

Učiteljica ili učitelj potiče učenike na ustrajnost u pronalaženju rješenja. Samo u iznimnom slučaju može se očekivati da učenik ili učenica samostalno dođu do rješenja ovog zadataka i u tom slučaju vjerojatno se radi o talentu za programiranje.

Pitanje: Kako isprazniti pretinac? Odgovor se daje slikom (slika 6.):



Slika 6. Naredba za pražnjenje sadržaja zelenog pretinca (po ideji M.Čičin-Šain)

C. Primjer 3. za realizaciju definiranog ishoda B1.2.: Kako bi izgledao postupak kojim bi pomoću Meme izračunali koliko smo platili 5 kg jabuka ako je poznata cijena 1 kg?

Postupak rješavanja zadatka:

Riječima: Najprije je u neki pretinac Meme potrebno unijeti cijenu jednog kg jabuka.

Pitanje: Kako to nacrtati? Odgovor je prikazan na slici 7.



Slika 7. Naredba za unos cijene 1 kg jabuka (po ideji M.Čičin-Šain)

Pitanje: Da li se mora upotrijebiti zeleni pretinac? *Odgovor:* Ne, mogao se upotrijebiti bilo koji pretinac ali u nastavku se mora koristiti onaj koji je odabran.

Pitanje: Da li je u zeleno trebalo odmah upisati i neki broj? *Odgovor:* Ne, jer želimo da postupak vrijedi za svaku cijenu jabuka.

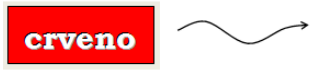
Pitanje: Kako izračunati cijenu za 5 kg jabuka? *Odgovor:* Potrebno je cijenu jednog kg pomnožiti sa brojem 5.

Pitanje: Kako to nacrtati? *Odgovor:* slika 8.



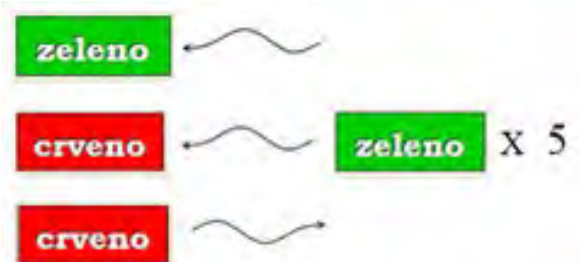
Slika 8. Naredba za izračunavanje cijene za 5 kg jabuka
(po ideji M.Čičin-Šain)

Pitanje: Da li se mora upotrijebiti crveni pretinac?
Odgovor: Ne, može se upotrijebiti bilo koji pretinac.
Pitanje: Da li se umjesto broja 5 mogao na to mjesto nacrtati neki pretinac?
Odgovor: Samo ako smo prethodno u taj pretinac unijeli broj 5.
Pitanje: Da li sada korisnik zna koliko košta 5 kg jabuka?
Odgovor: Ne, jer je rezultat u crvenom pretincu još nije ispisan.
Pitanje: kako ispisati sadržaj crvenog pretinca?
Odgovor: slika 9.



Slika 9. Naredba za ispis sadržaja crvenog pretinca
(po ideji M.Čičin-Šain)

Pitanje: Da li je ta naredba ispraznila crveni pretinac?
Odgovor: Ne, samo ga je ispisala.
Pitanje: Nacrtaj kompletan program?
Odgovor: slika 10.



Slika 10. Postupak kojim bi pomoću Meme izračunali koliko smo platili 5 kg jabuka ako je poznata cijena 1 kg?
(po ideji M.Čičin-Šain)

Pitanje: Koliko će koštati 5 kg jabuka ako 1 kg košta 4,00 kn?

Odgovor: Pomoću Meme prođi sve korake ako je ulaz u prvoj naredbi 4.

IV. ZAKLJUČAK

Metoda početnica Mema nastala je prije četrdeset godina, a danas je još aktualnija nego onda kada je nastala. Tijekom vremena doživjela je neke manje promjene za prvi razred i dograđena je 2014 za drugi, treći i četvrti razred.

Metoda dokazano pomaže [5] u usvajanju znanja iz pravog programiranja u nekom programskom jeziku, ne zahtjeva nikakvu opremu, može je izvoditi učitelj ili učiteljica bez informatičkog predznanja, dostupna je literatura za učitelje na hrvatskom jeziku, za prvi razred postoje i testovi koji mogu olakšati provjeru učenikovog znanja.

Autorice tiskane početnice trenutno rade na prijevodu mrežnih stranica prvog razreda na engleski jezik [2].

Postoje i planovi za daljnji rad. Ako bi među učiteljima postojao interes moguće je materijale za drugi, treći i četvrti razred staviti na Internet, a vrijedilo bi promisliti i da ova metoda doživi izdanje za učenike na računalu.

LITERATURA

[1] Kurikulum za predmet Informatika, dostupno na: http://www.kurikulum.hr/wp-content/uploads/2016/03/ Informatika.pdf, [pristupano: 08.03.2016.]

[2] Početnica Mema, dostupno na: http://www.magistra-nova.hr/dokuwiki_pocetnica/doku.php, [pristupano: 08.03.2016.]

[3] M. Čičin-Šain, "Kompjuterska početnica", Školska knjiga, Zagreb 1990 .

[4] M. Čičin-Šain, S. Babić "Početnica Mema – priručnik za učiteljice i učitelje", Društvo kibernetičara Rijeka, Rujan, 2014.

[5] M. Čičin-Šain, "Utjecaj predznanja na usvajanje novih informatičkih sadržaja u osnovnoj školi", Simpozij "Kompjutor na sveučilištu", Cavtat, 1987.

[6] M. Konecki, N. Kadoić , R. Piltaver, "Intelligent Assistant for Helping Students to Learn Programming ", In Proceedings of Mipro CE conference 2015, Opatija, Croatia, 2015.

Izloženost i navike korištenja medija i računala kod djece u razrednoj nastavi

Tea Pavičić*, Jelena Šurić**
*Prva osnovna škola Varaždin, Varaždin, Republika Hrvatska, tea232pavi@hotmail.com
** Osnovna škola „Braća Radić“, Koprivnica, Koprivnica, Republika Hrvatska, suric.jelena@gmail.com

Sažetak - U ovom se radu istražuju izloženost i navike korištenja računala i medijskih sadržaja učenika razredne nastave u jednoj OŠ u Varaždinu i jednoj u Koprivnici. Rezultati istraživanja prikazuju odnos djece u dobi od 8 do 10 godina prema digitalnoj tehnologiji ali i njihovu empatiju prilikom korištenja računala. Učenici razredne nastave posjeduju računala i tablete, koriste ih, ali vrlo malo u obrazovne svrhe. Većina anketiranih učenika uviđa važnost informatičke pismenosti. Korištenje računala u spomenutoj životnoj dobi donosi mnoge rizike te nužno zahtjeva vodstvo i usmjeravanje. Naši rezultati pokazuju da većina roditelja tek ponekad razgovara o načinima korištenja računala i provjerava što njihova djeca rade na računalu.

I. UVOD

U Hrvatskoj osim nekoliko istraživanja [8] ne postoje longitudinalna istraživanja kojima bismo kroz određeni vremenski rok, mogli pratiti korištenje medija kod djece i mladih. Još je manje istraživanja koja se bave proučavanjem i vrednovanjem informatičke pismenosti učenica i učenika u razrednoj nastavi u osnovnim školama. To je bio glavni poticaj za istraživanje koje ćemo predstaviti u ovome radu. Polazimo od pretpostavke kako navike korištenja medija kod djece i mladih moramo promatrati u kontekstu informatičke pismenosti te slijedom toga, između ostaloga, želimo ispitati tezu prepoznaju li učenici razredne nastave važnost informatičke pismenosti.

2011. i 2013. godine GfK proveo je dva istraživanja o korištenju medija kod djece u Americi. Ispitanici su bili roditelji djece do 8 godina. Rezultati su pokazali da se dječji pristup mobilnim medijima (mobitelima, tabletima) povećao u dvije godine, od 2011. do 2013. godine. 2011. godine 41% djece u dobi do 8 godina imali su pristup pametnom telefonu, a taj se postotak 2013. godine povećao na 63%. Gotovo dvostruko više djece koristi mobilnu tehnologiju 2013. godine, u odnosu na 2011. godinu, a prosječna količina vremena koju djeca provode koristeći mobilnu tehnologiju utrostručila se [27]. Iako su pametni telefoni sve prisutniji, istraživanje (iz 2013. godine) potvrđuje da je televizija i dalje najzastupljeniji medij u djetetovu slobodnom vremenu. 58% djece gleda televiziju barem jedan sat dnevno, 17% njih svakodnevno koristi mobitel,

14% svakodnevno koristi računalo, a 6% igra računalne igrice svaki dan [29].

Gfk[1] je također proveo istraživanje o mladima u Hrvatskoj 2006. godine. Ispitana je skupina djece od 11-13 godina te se pokazalo da 73% djece ima mobitel pri čemu ih djeca koriste najčešće za slanje poruka i igranje igrica. Od toga, 45% djece nema nikakav limit postavljen od strane roditelja vezan za potrošnju.

Mikić i Rukavina 2006. godine proveli su istraživanje o djeci i medijima. Uzorak od 160 ispitanika činila su djeca od 1. do 4. razreda triju osnovnih škola na području grada Zagreba – Druga osnovna škola Luka – Sesvete, Osnovna škola Augusta Šenoe i Osnovna škola Jure Kaštelana. „Predmet istraživanja bilo je značenje masovnih medija u svakodnevnom životu djece“ [19]. Rezultati istraživanja pokazuju da učenici u slobodno vrijeme prakticiraju sljedeće aktivnosti: gledanje televizije (67,50%), igra s prijateljima (66,25%), razgovor s roditeljima (38,75%), pomaganje u kućanskim poslovima (38,15%), igranje računalnih igara (35%), čitanje knjiga ili stripova (30%), gledanje filmova na videu/DVD-u (26,88%). Ista studija pokazuje da 48,13% ispitanika koristi računalo do jedan sat dnevno, 18,13% do 2 sata, 9,38% do 3 sata, 10,63% više od 3 sata dnevno, a 13,75% ispitanika računalo uopće ne koristi. Ispitanici računalo najčešće koriste za igranje igrica.

Poliklinika za zaštitu djece grada Zagreba i Hrabri telefon 2008. godine proveli su istraživanje o navikama i iskustvima djece i mladih prilikom korištenja interneta, mobitela i drugih suvremenih tehnologija. U istraživanju su sudjelovali učenici osnovnih i srednjih škola (od 11 do 18 godina) iz nekoliko hrvatskih gradova – Zagreba, Splita, Dubrovnika, Osijeka i Opatije. Istraživanjem se željelo provjeriti u koje svrhe i za što ispitanici koriste internet. 81% ispitanika internet koristi za komunikaciju s drugima. Jedan od zanimljivijih podataka jest izjava 49% ispitanika koji tvrde kako internet koriste bez nadzora i prisutnosti odraslih. Ipak 46% njih navodi da su roditelji samo ponekad prisutni, a tek 2,5% ispitanika navodi da je jedan roditelj uvijek prisutan dok je na internetu. Mnoga istraživanja istraživala su pozitivan utjecaj računala kao

[1] Gfk je najveća agencija za istraživanje tržišta u Hrvatskoj osnovana 1999 godine. GFK grupu čini više od 150 istraživačkih poduzeća u preko 70 zemalja širom svijeta.

medija na učenikov socijalni, kognitivni i emocionalni razvoj. Zaključci su išli u smjeru kako računala pomažu razvoju učenikove memorije, komunikacijskih sposobnosti i logike rješavanja problema te utječu na poboljšanje literarnih i matematičkih sposobnosti [3]. No, kako bi učenici mogli iskoristiti računalo u svakodnevnom radu za razvoj svojih znanja i vještina presudno je vodstvo učitelja i roditelja u uvođenju računala u njihov život. Kao jedna od temeljnih prednosti računala često se navodi mogućnost za individualizirano učenje jer omogućuje djetetu samostalno učenje vlastitim tempom. „*Djeca vole surađivati jedna s drugom dok rade na računalu i pri tom komuniciraju*." [22].

Kao što smo vidjeli iz pregleda postojećih studija u Hrvatskoj nedostaju istraživanja u kojima se istražuju navike i iskustva konzumiranja medija kod djece u razrednoj nastavi. Stoga smo u ovome radu odlučili upotpuniti tu znanstvenu prazninu. Primjenom znanstvene metode ankete cilj istraživanja bio je ispitati posjeduju li učenici računalo i tablet, koriste li računalo svaki dan, kako se pritom osjećaju, koriste li računalo za pisanje domaćih zadaća, koliko često razgovaraju s roditeljima o pravilnim načinima korištenja računala te kakvo je njihovo mišljenje o važnosti informatičke pismenosti. Istraživanje polazi od sljedećih hipoteza: većina učenika posjeduje računalo i/ili tablet; većina učenika svoje slobodno vrijeme provodi ispred ekrana; učenici računalo koriste svaki dan, ali ga ne koriste u svrhu pisanja domaćih zadaća; učenici se vole služiti računalom; učenici nedovoljno razgovaraju s roditeljima o pravilnim načinima korištenja računala i interneta; učenici već u samim početcima korištenja računala prepoznaju važnost informatičke pismenosti.

II. METODOLOGIJA

Istraživanje je provedeno u dvije osnovne škole, jednoj varaždinskoj [2] i jednoj koprivničkoj[3], u listopadu 2014. godine. Predmet istraživanja je odnos učenika razredne nastave prema digitalnoj tehnologiji. Anketnim upitnikom, kojeg smo nazvali Tehnologija i ja, željelo se odgovoriti na postavljene ciljeve istraživanja. Upitnik se sastojao od 9 pitanja. U zatvorenom tipu pitanja, ispitanici su od nekoliko ponuđenih odgovora morali odabrati i zaokružiti jedan ili više odgovora. Nezavisne varijable istraživanja su razred i spol ispitanika, a zavisne životne navike, informiranost, stavovi i dostupnost informacija.

Ciljevi istraživanja su ispitati i utvrditi:

- posjeduju li učenici razredne nastave računalo i/ili tablet?
- što učenici najčešće rade u svoje slobodno vrijeme?
- koriste li učenici računalo svaki dan?
- vole li se učenici služiti računalom?
- koriste li učenici računalo za pisanje domaćih zadaća?
- razgovaraju li učenici s roditeljima o pravilnim načinima korištenja računala i Interneta?
- kakvo je njihovo mišljenje o važnosti informatičke pismenosti?[4]

Istraživanje polazi od sljedećih hipoteza:

1. Većina učenika posjeduje računalo i/ili tablet.
2. Većina učenika svoje slobodno vrijeme provodi ispred ekrana.
3. Većina učenika računalo koristi svaki dan, ali ga ne koriste u svrhu pisanja domaćih zadaća.
4. Učenici se vole služiti računalom.
5. Učenici nedovoljno razgovaraju s roditeljima o pravilnim načinima korištenja računala i interneta.
6. Učenici već u samim početcima korištenja računala prepoznaju važnost informatičke pismenosti.

Istraživanje je provedeno na uzorku od 194 učenika 2., 3. i 4. razreda. Podaci vrednovani zasebno po školama nisu se značajno razlikovali, stoga nisu prikazani odvojeno. Uzorak je prigodan.

TABLICA 1. PRIKAZ UZORKA - BROJ UČENIKA PREMA SPOLU I RAZREDU

		Kojeg si spola		Ukupno (N)
		muški spol (N)	ženski spol (N)	
Razred	2.	33	18	51
	3.	45	45	90
	4.	18	35	53
Ukupno		96	98	194

A. Prikaz rezultata istraživanja

Sukladno našim očekivanjima i postavljenim pretpostavkama istraživanje je pokazalo kako 182 učenika posjeduje računalo, a čak 105 njih ima i tablet.

Ukupno gledajući, kada je riječ o zadovoljstvu učenika prilikom korištenja računala istraživanje je pokazalo kako se većina učenika (84%) voli služiti računalom.

TABLICA II. PRIKAZ ODGOVORA NA PITANJE VOLE LI SE UČENICI SLUŽITI RAČUNALOM, N = 194

[2] Grad Varaždin, 38.839 stanovnika, 7 osnovnih škola

[3] Grad Koprivnica, 23.955 stanovnika, 3 osnovne škola

[4] Prema Aniću (2006: 441) pridjev informacijski se odnosi na informacije i obavijesti, a pridjev informatički na informatiku. Informacijska pismenost (eng. information literacy) jest sposobnost čovjeka da prepozna kada su mu informacije potrebne, te da locira, ocijeni i učinkovito koristi informacije kojima će zadovoljiti tu svoju potrebu (Panian, 2005: 273). Informatička pismenost (eng. computer literacy) definira se kao sposobnost korištenja računala i računalnih programa (Nadrljanski, 2006: 262).

Voliš li se služiti računalom?					
Razred		frekvencija	postotak	vrijednosni postotak	kumulativni postotak
2.	ne	8	15,7	15,7	15,7
	da	43	84,3	84,3	100,0
	ukupno	51	100,0	100,0	
3.	ne	19	21,1	21,1	21,1
	da	71	78,9	78,9	100,0
	ukupno	90	100,0	100,0	
4.	ne	4	7,5	7,5	7,5
	da	49	92,5	92,5	100,0
	ukupno	53	100,0	100,0	

Zanimalo nas je što u svoje slobodno vrijeme rade učenici od 2. do 4. razreda osnovne škole u 2014. godini. Učenici su u ovom pitanju mogli zaokružiti više ponuđenih odgovora. U Tablici III. vidimo da najveći broj učenika slobodno vrijeme provodi s roditeljima i prijateljima. Od digitalnih uređaja najzastupljeniji su televizor i računalo/tablet. Usporedimo li rezultate po spolu i razredu vidimo da nema značajne razlike u korištenju slobodnog vremena kod dječaka i djevojčica ni u pojedinim razredima. Osim ponuđenih odgovora učenici su imali mogućnost dopisivanja vlastitog odgovora. Najčešće dopisana aktivnost bila je igranje nogometa kod dječaka (N=13) i igranje sa životinjama kod djevojčica (N=5). Osim toga, učenici u svoje slobodno vrijeme uče (N=4), igraju se (N=4), idu u grad (N=3),voze bicikl (N=2)…

TABLICA III. PRIKAZ UČENIČKIH ODGOVORA NA PITANJE O AKTIVNOSTIMA U SLOBODNOM VREMENU, N = 194

Što najčešće radiš u slobodno vrijeme?	Razred			Spol		Ukupno
	2.	3.	4.	muški	ženski	
Provodim vrijeme s roditeljima	35	54	31	52	68	120
Igram se s prijateljima	36	68	34	68	70	138
Čitam knjige	16	30	13	28	31	59
Gledam televiziju	38	51	26	55	60	115
Dopisujem se na mobitelu	8	11	12	16	15	31
Igram igrice na računalu/tabletu	31	37	30	54	44	98
Pretražujem Internet	16	6	16	22	16	38

Učenici posjeduju računalo i vole ga koristiti, no koriste li ga i za ispunjavanje nekih svojih školskih obveza? Na pitanje: „Jesi li ikada koristio/koristila računalo za pisanje domaćih zadaća?“ većina učenika, 57,7%, odgovara negativno na postavljeno pitanje, što potvrđuje hipotezu da većina učenika računalo koristi svaki dan, ali ga ne koriste u svrhu pisanja domaćih zadaća.

TABLICA IV. ODGOVORI UČENIKA O KORIŠTENJU RAČUNALA U SVRHU PISANJA DOMAĆIH ZADAĆA, N=194

Jesi li ikada koristio računalo za pisanje domaćih zadaća?				
	frekvencija	postotak	vrijednosni postotak	kumulativni postotak
ne	112	57,7	57,7	57,7
da	82	42,3	42,3	100,0
ukupno	194	100,0	100,0	

Rezultati pokazuju da najviše učenika u 4. razredu koristi računalo za pisanje domaćih zadaća, 66%. Prethodni podaci pokazali su da svi učenici, bez obzira na dob, koriste i vole koristi računalo. Dakle, učitelji bi bez straha mogli uključiti računalo u izvršenje školskih obveza već od 2. razreda osnovne škole.

TABLICA V. ODGOVORI UČENIKA O KORIŠTENJU RAČUNALA U SVRHU PISANJA DOMAĆIH ZADAĆA, PO RAZREDIMA, N=194

Jesi li ikada koristio računalo za pisanje domaćih zadaća?					
Razred		frekvencija	Postotak	vrijednosni postotak	kumulativni postotak
2.	ne	34	66,7	66,7	66,7
	da	17	33,3	33,3	100,0
	ukupno	51	100,0	100,0	
3.	ne	60	66,7	66,7	66,7
	da	30	33,3	33,3	100,0
	ukupno	90	100,0	100,0	
4.	ne	18	34,0	34,0	34,0
	da	35	66,0	66,0	100,0
	ukupno	53	100,0	100,0	

Na početku rada spomenuta je važnost roditeljskog vodstva u upoznavanju i korištenju računala. Anketirani učenici upitani su stoga razgovaraju li s roditeljima o tome kako se pravilno treba služiti računalom i Internetom te koliko često roditelji sjednu s njima za računalo i pogledaju što oni rade. Čak 63 učenika izjavljuje kako nikada nisu s roditeljima razgovarali o pravilnim načinima korištenja računala i Interneta, 104 učenika razgovaralo je nekoliko puta, a najmanji broj učenika, svega 27, razgovara svakodnevno s roditeljima o tome. Zašto tako veliki broj roditelja nije nikada razgovarao sa svojom djecom o pravilnim načinima korištenja računala i Interneta? Postoji mogućnost da roditelji nisu educirani, da nemaju dovoljno vremena ili ne uviđaju važnost i nužnost takvih razgovora. Koji god odgovor uzeli u obzir jasna je spoznaja da je potrebno educirati roditelje o mogućim negativnim utjecajima digitalne tehnologije na dijete te o važnosti njihova vodstva i nadzora. Provjerili smo i koliko često roditelji sjednu sa svojom djecom za računalo i pogledaju što oni rade na računalu. Budući da govorimo o djeci u dobi od 8 do 10 godina, zabrinjava podatak da samo 21,2% roditelja redovito sjedne sa svojom djecom za računalo.

TABLICA VI. PRIKAZ ODGOVORA NA PITANJE O RODITELJSKOM NADZORU DJETETOVA RADA NA RAČUNALU, N=194

Koliko često roditelji sjednu s tobom za računalo i gledaju što ti radiš?				
	frekvencija	postotak	vrijednosni postotak	kumulativni postotak
nikada	35	18,0	18,0	18,0
ponekad	118	60,8	60,8	78,9
svaki put	41	21,1	21,1	100,0
ukupno	194	100,0	100,0	

III. ZAKLJUČAK

Ovaj je rad prikazao iskustva korištenja medija na koje nailaze učenici od 2. do 4. razreda osnovne škole u dva hrvatska grada. Pozivajući se na postojeća istraživanja u radu smo pokazali kako korištenje računala može učenicima olakšati izvršenje brojnih zadataka, može mu poboljšati uspjeh u školi, potaknuti razvoj jačih i nadogradnju lošijih strana, unaprijediti sposobnost pisanog izražavanja te proširiti njegove poglede na svijet. Preduvjet za to je pravilno korištenje računala te informatička, informacijska i medijska pismenost. Prikazom ključnih inozemnih i domaćih istraživanja pokazali smo da nekontrolirano i prekomjerno korištenje medija i medijskih sadržaja u određenim uvjetima dovodi do zatiranja pozitivnih, a izražavanja negativnih učinaka (prekomjerna tjelesna težina, nesanica, agresivna ponašanja, emocionalna neosjetljivost…).

Dobiveni rezultati istraživanja u potpunosti potvrđuju postavljene hipoteze. Pokazalo se da 93,8% učenika posjeduje računalo, dok 54,1% učenika uz računalo posjeduje i tablet. Dobiveni rezultati vrlo su slični rezultatima istraživanja Poliklinike za zaštitu djece grada Zagreba i Hrabrog telefona (2008) gdje je ustanovljeno da čak 95% djece iz pet većih gradova Hrvatske posjeduje i koristi računalo. Rezultati istraživanja, osim toga pokazuju da učenici većinu slobodnog vremena ipak provode sa svojim roditeljima i prijateljima, dok korištenje računala/tableta i gledanje TV-a zauzima također značajno drugo mjesto. Da postotak korištenja računala i raznih medija s godištem djece raste, pokazuje nam i istraživanje provedeno u svibnju i lipnju 2000. godine u Hrvatskoj iz kojeg saznajemo da učenici od 5. do 8. razreda većinu slobodnog vremena provode koristeći različite medije (televizija, radio, tisak, internet), dok na socijalne kontakte koriste manje slobodnog vremena [9].

Vrlo visok postotak učenika, njih 84% koristi računalo zato što to voli, što bi bila velika prednost kad bi se računalo koristilo više u obrazovne svrhe. Nažalost, većina učenika (57,7%) ne koristi računalo za pisanje domaćih zadaća. Ako dakle gledamo krajnju motivaciju učenika kao korisnika, onda možemo primijetiti kako oni zapravo koriste računala bez pravilne motivacije. Stoga je upravo na ovom području važno definirati nove pristupe i uzeti u obzir motivaciju učenika kao vrlo bitnu varijablu. Samo 21,2% ispitanika redovito razgovara s roditeljima o tome kako se pravilno služiti računalom i Internetom, dok zabrinjavajućih 32,4% ispitanika nikada ne razgovara sa svojim roditeljima o toj temi. Taj je podatak zabrinjavajući obzirom na dob učenika kojima je potrebno roditeljsko vodstvo u usmjeravanju i korištenju računala te zbog brojnih opasnosti kojima se djeca mogu susresti koristeći Internet. Uspoređujući dobivene rezultate s rezultatima Poliklinike za zaštitu djece grada Zagreba i Hrabrog telefona gdje 49% ispitanika izjavljuje kako internet koriste bez nadzora i prisutnosti odraslih, jasno je da je nužno što prije pokrenuti nove aktivnosti kako bi se potaknulo posredovanje odraslih.

Da učenici već u samim početcima korištenja računala uviđaju važnost informatičke pismenosti potvrđuju nam rezultati koji pokazuju da 76,5% ispitanika smatra da je važno znati služiti se računalom i internetom, pri čemu su najosviješteniji učenici četvrtih razreda. Iako se u raspravama kao jedan od ključnih argumenata roditelja može čuti kako je važno da njihova djeca imaju računalo radi učenja, istraživanje Mikića i Rukavine pokazalo je kako 43% djece nikada ne koristi računalo za učenje, a 50% djece ga nikada ne koristi za pristup i pretraživanje interneta. Hoće li se i kada taj podatak promijeniti u korist obrazovanja i pravilnog usmjerenja djece pri korištenju računala, ovisi isključivo o angažmanu roditelja, a kasnije i samih učitelja.

Spoznaje dobivene provedenim istraživanjem na učenicima od 2. do 4. razreda u dvjema hrvatskim osnovnim školama, kao i one iz drugih prikazanih hrvatskih i inozemnih istraživanja, ukazuju na potrebu sigurnog, jasnog pa čak i stručnog vodstva učenika u njihovim početnim medijskim pothvatima. Nužno je stoga informirati i educirati roditelje o temi izloženosti i navika korištenja medija i računala učenika razredne nastave. Upravo će se zato rezultati ovog istraživanja, kao i druga prikazana istraživanja u ovome radu, prezentirati razrednim učiteljicama i roditeljima učenika koji su sudjelovali u istraživanju. Nameće se pitanje je li u današnje vrijeme potrebno kontrolirati izloženost djece i mladih medijima i detaljno istražiti ulogu medija u svakodnevnom životu djece i mladih, u njihovoj socijalizaciji, odgoju i obrazovanju. To je uistinu prostor koji je moguće osvojiti. Roditelji, odgajatelji, učitelji, a naravno i sam vrh hrvatskog obrazovnog sustava moraju osvijestiti važnost medijskog obrazovanja kako bismo, svi zajedno, što bolje koristili pozitivne strane digitalne tehnologije.

LITERATURA

[1] N. Blažević, „Djeca i mediji – odgoj na „televizijski“ način“, Nova prisutnost 10, 2012, pp. 479-493

[2] J. Burić, „ Djeca i mladi kao konzumenti masovnih medija“, Etika i tržišne manipulacije potrebama mladih, Filozofska istraživanja 30, 2010, pp. 629-634.

[3] L.Ciboci, I. Kanižaj i D. Labaš,“ Djeca medija: Od marginalizacije do senzacije“, Zagreb, Matica Hrvatska, 2011.

[4] Deklaracija o medijskoj kulturi, 2012, http://www.djecamedija.org/?p=62 (22. 01. 2016.)

[5] K. Erjavec, „Odgoj za medije: Od koncepta do školske prakse“, Medijska pismenost i civilno društvo, 2006, pp. 78-106.

[6] N. Foretić, S. Rodek i D. Mihaljević,“ Utjecaj medija na fizičku inaktivnost djece“, Školski vjesnik 58 (4), 2009, pp. 381-397.

[7] G. Heiting, „ Kids, Computers and Computer Vision“, All About Vision, 2011., dostupno na: http://www.allaboutvision.com/parents/children-computer-vision-syndrome.htm (22. 01 2016.)

[8] V. Ilišin, „Mediji u slobodnom vremenu djece i komunikacija o medijskim sadržajima“, Medijska istraživanja 9 (2), 2003, pp. 9-34.

[9] V. Ilišin, B. Marinović, A. i R., Furio, „ Djeca i mediji: Uloga medija u svakodnevnom životu djece“, Zagreb, Državni zavod za zaštitu obitelji, materinstva i mladeži i Institut za društvena istraživanja u Zagrebu, 2001.

[10] V. Ilišin, D. Bouillet, A. Gvozdanović, D. Potočnik, „Mladi u vremenu krize: Prvo istraživanje IDIZ-a i Zaklade Friedrich Ebert o mladima“, Zagreb: Institut za društvena istraživanja u Zagrebu Friedrich Ebert Stiftung, 2013.

[11] Istraživanje o korištenju Interneta, mobitela i drugih tehnologija [online], Poliklinika za zaštitu djece grada Zagreba, dostupno na: http://www.poliklinika-djeca.hr/istrazivanja/istrazivanje-o-koristenju-interneta-mobitela-i-drugih-tehnologija/ (22.01.2016.)

[12] Konvencija o pravu djeteta, 1989, http://www.unmikonline.org/regulations/unmikgazette/05bosniak/BConrightsChildren.pdf (20.01.2016.)

[13] T. Legović, „Manipuliranje djecom u televizijskom oglašavanju“, Drugost 2, 2011, pp. 82-87.

[14] G. Livazović, „Teorijsko – metodološke značajke utjecaja medija na adolescente“, Život i škola, časopis za teoriju i praksu odgoja i obrazovanja, 21 (57), 2011, pp. 108-115.

[15] Z. Miliša, M. Tolić, i N. Vertovšek, „Mediji i mladi“, Prevencija ovisnosti o medijskoj manipulaciji, Zagreb, 2009.

[16] Z. Miliša, „Djeca i mladi – ovisnici o ekranima“ [online], 2011, Dnevno.hr., dostupno na: http://www.dnevno.hr/kolumne/zlatko-milisa/42060-djeca-i-mladi-ovisnici-o-ekranima.html (22.01.2016.)

[17] Z. Miliša, M. Tolić, i N. Vertovšek, „ Mladi – odgoj za medije“, Priručnik za stjecanje medijskih kompetencija. Zagreb: M.E.P. Centar menadžerske knjige, 2010.

[18] Z. Mandarić,“ Novi mediji i rizično ponašanje djece i mladih“, Bogoslovska smotra 82 (1), 2012, pp. 131-149.

[19] K. Mikić i A. Rukavina,“ Djeca i mediji “, Hrvatski filmski savez, 2006. http://www.hfs.hr/nakladnistvo_zapis_detail.aspx?sif_clanci=1604#.UiHUYKKQZ4K (22.01.2016.)

[20] Z. Miliša i J. Zloković, „Odgoj i manipulacija djecom u obitelji i medijima“, Zagreb, 2008.

[21] Đ. Nadrljanski, „Informatička pismenost i informatizacija obrazovanja“, Informatologija, 39 (4), 2006, pp. 262-266.

[22] Đ. Nadrljanski, M. Nadrljanski, „Recenzije i prikazi“, Informatologija, 40 (2), 2007, pp. 158- 160.

[23] „Ovisnost o računalu – bolest koja se širi“, Centar za nestalu i zlostavljanu djecu, http://www.cnzd.org/site2/index.php?option=com_content&view=article&id=127:ovisnost-o-raunalu-bolest-koja-se-iri-&catid=35:vijesti (22. 01. 2016.)

[24] Ž. Panian, „Informatički enciklopedijski rječnik“ - L. Zagreb, Europapress holding d.o.o., 2005.

[25] T. Paviči, J. Šurić, „Učinci izloženosti mladih medijskim sadržajima“, 14th Mate Demarin Days: Contemporary Challenges Educational Theory and Practice, Topusko, 2014.

[26] J. Portner, Pavličević, T. Pavičić, „Stav učitelja o korištenju novih tehnologija u odgojno-obrazovnom sustavu – rezultati istraživanja“, CARnet CUC, Zagreb, 2014.

[27] J. Pregrad, L. Tomić, M. Mikulić, N. Šeparović,“ Iskustva i stavovi djece, roditelja i učitelja prema elektroničkim medijima“, Ured UNICEF-a za Hrvatsku, http://www.unicef.hr/upload/file/353/176706/FILENAME/Izvjestaj_-_Iskustva_i_stavovi_djece_roditelja_i_ucitelja_prema_elektronickim_medijima.pdf (22.01.2016.)

[28] V. Rideout, „ Zero to Eight: Children's Media Use in America 2013“, A Common Sense Media Research Study, 2013.

[29] A. Sigman, „Daljinski upravljani: Kako televizija šteti našim životima – i što po tom pitanju možemo učiniti“, Velika Mlaka, Ostvarenje, 2010.

[30] S. Špiranec, Z. Banek, „Inforacijska pismenost: Teorijski okviri i polazišta“, Zavod za informacijske studije Odsjeka za informacijske znanosti Filozofskog fakulteta Sveučilišta u Zagrebu, 2008.

[31] J. Šurić, T. Pavičić, M. Dumančić, „Informatika kao izborni predmet ili izvannastavna aktivnost u razrednoj nastavi“, The 37th International ICT Convention – MIPRO, 2014.

[32] S. Žižek, „O nasilju“, Naklada Ljevak, zagreb, 2008.

[33] Head Start, An Office of the Administration for Children and Families, Early Childhood Learning & Knowledge Centre (ECLKC) Dostupno na: http://eclkc.ohs.acf.hhs.gov/hslc (22.01.2016.)

C T S

I n t e r n a t i o n a l C o n f e r e n c e o n

COMPUTERS IN TECHNICAL SYSTEMS

Steering Committee

Chairs: Vlado Sruk, University of Zagreb, Croatia
Leonardo Jelenković, University of Zagreb, Croatia
Dejan Škvorc, University of Zagreb, Croatia
Jadranko F. Novak, Nomen, Rijeka, Croatia
Nikola Mišković, University of Zagreb, Croatia

Members: Matjaţ Colnarič, University of Maribor, Slovenia
Ţeljko Hocenski, University of Osijek, Croatia
Niko Guid, University of Maribor, Slovenia

INVITED PAPERS

Architecture and Application of Virtual Desk and 3D Process Simulation for Wire Rod Rolling Mills

A.Venuti
Executive Manager Rolling Mill – BU Long Products – at Danieli Automation S.p.A., Buttrio (UD)
a.venuti@dca.it

Abstract – Nowadays the complexity of modern rolling mill plants is steadily increasing and thus the demand placed on automation systems. Therefore the use of virtual process systems can be very beneficial to validate the control system before start-up, as well as a training tool for rolling mill operators. This paper presents the iStand, the Danieli Automation rolling mill simulator platform.

I. INTRODUCTION

The iStand is the Danieli Automation 3D simulation platform for the steel industry. The basic idea behind the iStand is to have a simulation system which can be used across the all stages of the project and throughout the life cycle of the plant.

In the engineering phase, simulator is useful for checking the automation control system. This is an important step for more reliable automation, improving the quality of the delivered PLC program.

At the commissioning stage, the application of this technology becomes more tangible. The use of simulation-based engineering reduces the problems at the startup of the real plant as well as the production ramp-up times and the costs. The simulator enables the possibility of performing virtual commissioning long before plant erection, leading to a sensible reduction of plant start-up time and risks for real equipment.

The simulator is used to enhance the rolling mill operator's skill separate from the real plant. The system is designed to allow operators to be trained in an environment identical to the rolling mill control room, or to the local control station. The iStand implements several mill fault scenarios in which an operator can be trained in a complete risk free environment. This provides safe and efficient operations that benefit the workers and the overall profitability of the plant.

In the past, the customer's engineers have often had to get used to new automation control systems quickly after plant start-up. The iStand allows them to start training earlier. After real plant commissioning the iStand allows engineers to implement and test software modifications in the virtual equipment without taking risks on the real one and assuring a faster production start-up on the plant.

The following section details the iStand architecture.

II. ISTAND ARCHITECTURE

The iStand is a component of the 3Q AUTOMATION, the Danieli Automation state-of-the-art technology for a modern plant operation.

The 3Q system is a complete suite of tools designed with the concept that the system makes suggestions to the operator about what to do, and in situations which present multiple possibilities, it informs the operator intelligently, to help decide which path to choose. The target is to reduce repetitive actions more and more.

As it is shown in Figure 1, the iStand consist of following equipment:

1. Control System Switchboard
2. Main pulpit operator desk replication
    a. Human Machine Interface (HMI)
    b. Operator Assistant (OA)
    c. Technological Work Station (TWS)
3. Local control station replication
4. Giant display for 3D simulation
5. Trainer desk

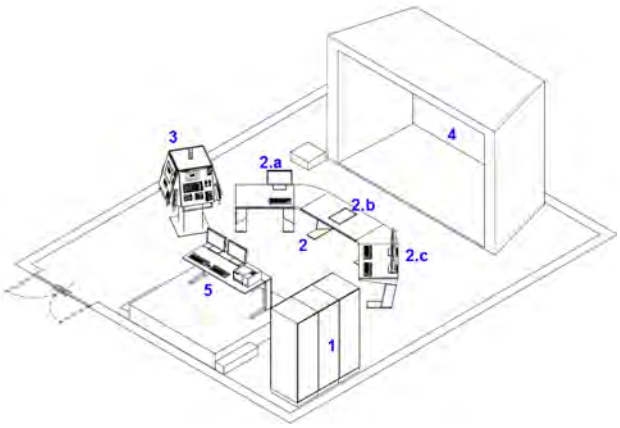


Figure 1. iStand of main control room.

A. *Control system*

The wire rod mill is quite a large and complex plant that demands tight control of many variables to meet production quality targets, especially when product speed is up to 120 m/s. The L1 automation is based on Siemens S7-400 process controllers.

The whole plant automation requires 11 CPUs running different process tasks and grouped as follow:

- 2 CPUs for rolling mill speed control
- 2 CPUs for rolling mill auxiliary services control
- 3 CPUs for wire rod line control
- 1 CPU for mill intermediate shear control
- 2 CPUs for garret line control
- 1 CPU for thermal treatment control

CPUs exchange data with each other through MPI interface (at ~50 ms) and through profinet network for time-critical data (less than 10 ms). The process at wire rod line requires faster CPU tasks running at 3 ms.

The iStand faces the problem of the cost that such amount of automation could have on the project. Nevertheless, a cheaper solution has to run the same S7 software developed for S7-400 controllers.

The initial idea to reduce the HW costs was to run multiple instances of virtualized controllers on a server using VMware workstation and WinAC RTX. WinAC RTX is the software used for PC-based automation from Siemens running on IntervalZero RTOS which permits real-time functions to be carried out on the PC. The tests conducted with virtualized controllers revealed communication latency in data flow between controllers and large controller tasks jitter that caused problems to real-time process simulation. Moreover a few undefined stability problems caused virtual controllers to crash. It was probably an issue related to the compatibility of the virtual machine and WinAC RTX.

Therefore, it was decided to use the Siemens Microbox PC IPC427D. These are compact and powerful embedded PC-based controllers running WinAC RTX which support Ethernet/Profinet communication. The chosen IPC427D processor was the Core i3 1.6GHz which performance allows S7 software to be executed faster than the latest series of S7-400 CPU416. This cycle-gap benefit allows the controllers to execute the L1 software required for the 3D simulation, ensuring real-time and deterministic events response. The data exchange with the supervisors PC is made via Ethernet network. A separate Ethernet network is used for the communication between IPC and for the communication with the 3D simulator.

The size of IPC controllers permits a compact design of the automation switchboard for iStand which includes the supervisors PCs and the 3D simulator PC as shown in Figure 2.

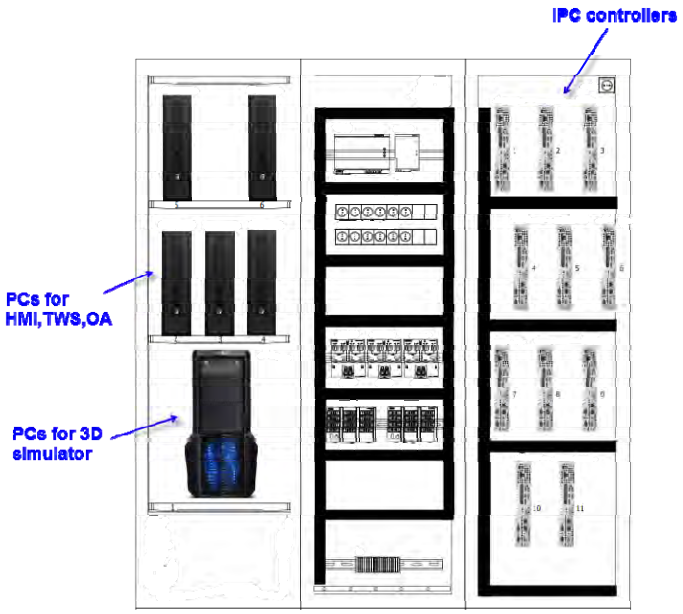


Figure 2. iStand control switchboard, internal view.

B. *Main pulpit operator desk and supervisors systems*

One of the key points of 3Q system by Danieli Automation is the provision of a full soft-desk pulpit, totally based on computer screens, through which the operator can both monitor the plant and operate it at the same time. Danieli Automation puts its know-how into this new system architecture. This knowledge-based approach ensures the operator is shown only the useful information required for that precise process step. Moreover the process information is shown in an innovative and intelligent interface, allowing operators to work more efficiently and easily. The iStand is designed to let operators be trained using the same systems they will have in the rolling mill control room together with a risk free virtual environment. These are the HMI, the OA and the TWS.

The Human Machine Interface (HMI) is the main interface which allows operators to view details of the process and mill equipment. The current situation of the rolling mill is presented to the operator on set of graphical overview pages in which all the devices are represented with icons and color describing their current status. Any malfunctioning equipment, or equipment which is not ready to roll, is immediately visible on the overview.



Figure 3. Human Machine Inteface (HMI), overview example.

All rolling mill equipment has its own detail screen on HMI showing all of its statuses/alarms and possible commands to control automation. Commands and ready-to-roll signals have interlock details which advise the operator of any missing device conditions.

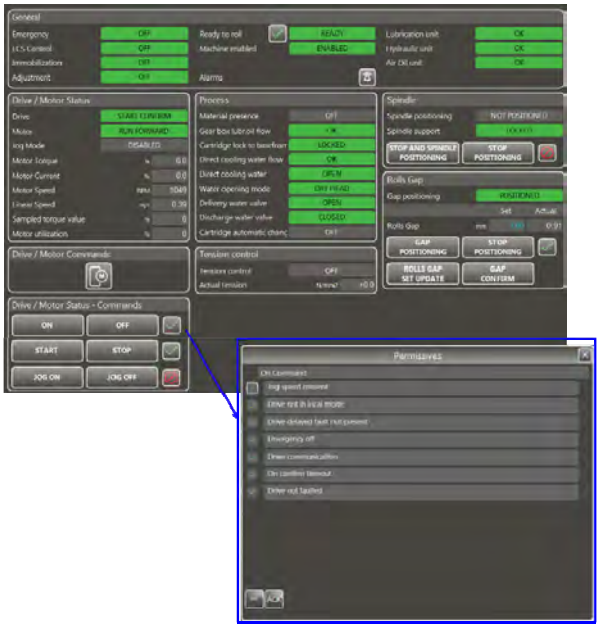


Figure 4. Human Machine Inteface (HMI), machine detail screen and interlocks.

The Operator Assistant (OA) is the new and most innovative component of the desk. It not only replaces all the manual controls (such as pushbuttons and lamps) present on traditional pulpit, but with its multi-touch display that needs more than a simple brush-by to activate, the OA has become the fulcrum of the main pulpit. Its innovative interface allows the operator to organize the six available desks in the OA as he requires. The OA concentrates much of the intelligence needed within the automation system and reduces the number of commands that operators have to consider. It's an interface with a knowledge-based approach that advises about the situation of the plant and suggests the right actions to obtain the expected result, hence saving time. This feature is achieved thanks to a coded list of messages triggered by the L1 and L2 systems. The messages concern both process and production and suggest to the operator the action to be taken. Other helpful OA features are the checklists which help the operator to ensure consistency and completeness in carrying out a task. For example, the start-up checklist of the rolling mill requires different actions, such as starting the air/oil, hydraulic and lubrication units, sending a production schedule to L1, starting all mill devices required for the scheduled production. These actions can be sent to L1 directly from the OA interface by pushing the action-buttons on the checklist. As soon as the checklist is completed, the rolling mill is ready to roll. The interlocked controls on OA are displayed as disabled, in gray. The list of interlocks appears by pressing on disabled control showing the causes for which the command is disabled.



Figure 5. Operator Assistant (OA), desks screens.

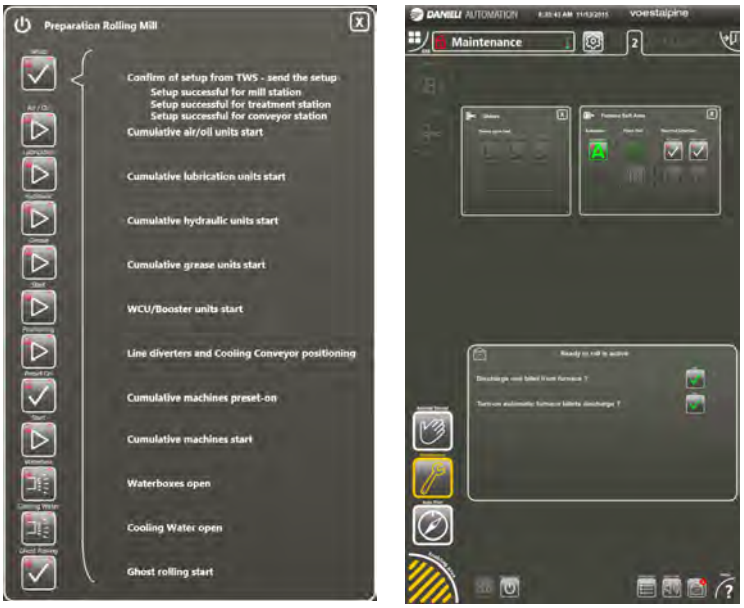


Figure 6. Operator Assistant (OA), mill preparation checklist detail and example of L1 driven message.

The Technological Work Station (TWS) is the client of L2 located on the main control desk in order to perform the rolling mill setup functions. The iStand PC running the TWS includes the L2 server-side components required for communications with L1 automation. Each process setup parameters for all the rolling mill mechanical equipments can be edited independently from the others. All of them are stored and retrieved together as a whole in the form of set-up-tables. The set-up-tables are related to the production of a particular product and when the operator needs to prepare the equipment for a specific production to run, he only needs to recall the appropriate setup table from the technical database and activate it. Even though the operator can handle the setup-tables manually, the

TWS is designed to link them to the schedule of production, in order to achieve a fully automatic setup management, maximizing the productivity of steel plants having short batch productions.



Figure 7. Technological Work Station (TWS).

C. *Local control station*

The aim of the local control station pulpit is to improve the operational skill of rolling mill field operators. In normal operation, their main assignment is the rolling mill dressing due to production change, in which one of the major tasks is the change of stand configuration. Even though stand changes run completely autonomously under the control of automation system, when a failure situation occurs, the automatic sequences stop, leaving the control to the operators. During a long period of trouble-free automated production, they have no opportunity to train their skills. The demand of uninterrupted production is in contrast with the training opportunities on the real plant. iStand overcomes this issue giving even to field operators the training possibilities that meets the same operating conditions as in reality, thanks to the replication of the stands change LCS in combination with real-time 3D simulation.

The iStand local control station is an innovative design which allows a combination of the possible stand change scenarios to be simulated: the GCC stand change, the Twin Module Block (TMB) cartridge change and the Sizing Mill (CSU) stand change.

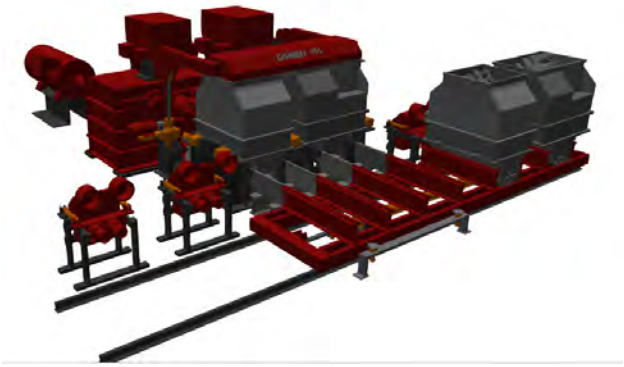


Figure 8. TMB cartridge change simualtion.

D. *Giant dispay for 3D simulation*

Danieli Automation chooses IndustrialPhysics from Machineering GmbH as CAE simulation software. The design of 3D simulation models start from the 3D-CAD mechanical drawings of the plant. To achieve a smooth rendering of the simulated scenarios, a crucial phase is the simplification CAD models. The simplified geometries have been imported into IndustrialPhysics via the STEP files interface. Even though most of the models are based on the physical behavior of machines, and have been developed using built-in functions and parameters at component level available in IndustrialPhysics, the simulation of the material flow process along a rolling mill required the development of new tracking objects and rendering technology inside IndustrialPhysics. Around one thousand input and output signals are communicated between all IPCs and the simulator PC via Ethernet network. The overall performance of the architecture allows the data roundtrip time to reach 20 ms which result in a smooth, real-time simulation.



Figure 9. 3D simulation, plant model.



Figure 10. 3D simulation, running material simulation on loopers.

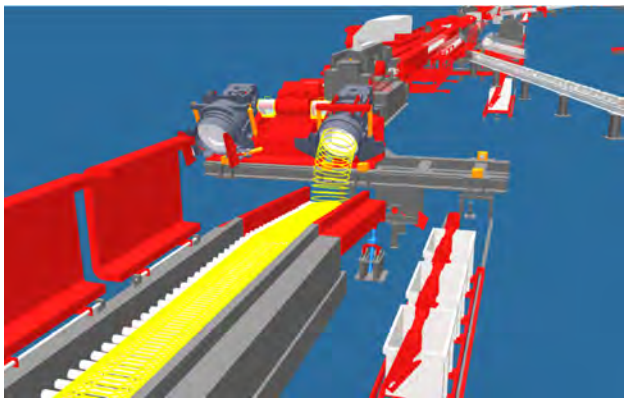


Figure 11. 3D simulation, running material simulation on turning forming head and cooling conveyor.

E. *Trainer desk*

The iStand is designed to train the operators under the supervision of a tutor who controls the simulation scenarios from the trainer desk, located behind the main pulpit. The trainer loads and runs the simulation in IndustrialPhysics which is automatically available on the giant monitor in front of the operator. Based on the running simulation, the trainer can trigger several fault conditions through dedicated screens on the HMI to test the trainee capabilities and verify the operational procedures.



Figure 12. Fault condition simulation screen.

III. CONCLUSION

iStand proves to be a useful platform to check-out automation at the engineering stage improving the quality of the control programs and saving plant start-up time. iStand is a valuable step for know-how transfer to operator in a safe and risk-free virtual environment which allows regular training sessions to be carried out to prevent operators to forget their abilities. iStand would be a valid test bench for all automation control improvements during the whole life of the plant.

REFERENCES

[1] Danieli Automation: 3Q Automation – One step ahead in Process Control, p12-15.

[2] Danieli Automation: iStand 3Q Simulation System – Seamless Pipe Mill.

[3] Machineering GmbH: IndustrialPhysics and Hardware in the loop simulation. http://www.machineering.de

Use of offline computational tools for plant data analysis and setup model calibration: a perspective in the industry of flat metal production

C. Aurora*, F.A. Cuzzola*
* Danieli Automation SpA, Flat Products Dept., 33042, Buttrio (UD), Italy
Email: {c.aurora, f.cuzzola}(at)dca.it

Abstract – The exploration and analysis of large databases represents a subject of research efforts for several industries. In this paper the case of production of flat metals, where advanced optimization technologies based on mathematical modeling are integrated in the automation architecture together with databases, is discussed: in particular the use of MATLAB as an offline tool for model analysis and calibration is presented. Finally, a case of study taken from a real model calibration activity is discussed.

I. INTRODUCTION

In the last years the market of flat metal products required continuous evolving and demanding quality standards: in particular, the economical downturn and the corresponding decrease of demand for metallurgical products have moved the competition on the quality field. Optimization of production in terms of lower costs, reduced maintenance and higher productivity has become a primary target as well. Moreover, some market fields, like the aluminum one, are naturally characterized by very high standards serving several industries from food to avionic applications. From the automation point of view, this implies the need for more sophisticated control solutions. For this reason, while up to few years ago inclusion of mathematical models of technological processes into automation systems was considered an option, nowadays this feature has become a standard and it is called to meet strict accuracy and reliability requirements.

Mathematical models are based on physical and empirical equations describing the behavior of production processes [1]. In the field of flat metal products, many kind of manufacturing processes take advantage from the adoption of optimization technologies based on physical models integrated in the automation system: hot and cold rolling mills (hot plate mills, cold reversing mills and tandem mills, [2][3]); skinpass and temper mills; pickling lines [4]; plate leveler machines; hot dip galvanizing lines (air knives and furnaces). In some other cases, sophisticated algorithms - which may be considered as "non-technological" models - are applied to a number of optimization problems, like plate cut-schedule [5].

All the previously mentioned plants must be properly preset before processing each piece. In some cases the final target is reached through a repeated number of operations (e.g. multiple reduction steps for a rolling mill). The whole processing schedule, including the set-points of the main control variables, is called *setup*. Moreover, the advantages introduced by the use of optimization tools based on mathematical models for setup calculation is widely recognized.

The efficiency of a setup model is mainly represented by its ability of matching a process behavior with precision, i.e. the ability of providing good predictions of set-points for product processing, requiring minimal adjustments to the low-level process regulators. On the other hand, from the implementation and operational point of view, accuracy is not the only feature which makes a model effective for online applications. Indeed, reliability and reduced computational time are important requirements as well. Finally, the range of validity of a model is another key point to be carefully evaluated.

Hence, design of mathematical models for online applications is a delicate task which must aim at the optimal compromise between accuracy, limited requirement on computational complexity and easiness of configuration (i.e. the minimization of the enumerability of the set of model parameters). In general, this optimal solution consists of implementing setup models with suitable tuning parameters easily adaptable to different but affine installations.

Despite online setup models are usually provided with self-learning adaptation routines, a good model calibration activity is always essential to properly configure a model for a new installation and the corresponding set of products. This is a typical commissioning activity which requires deeply inspecting the behavior of a plant from the recorded feedback data so as to calibrate the model accordingly.

Different plant types may require different kind of interventions, not only towards the adjustment of some specific parameters, but also through the calibration of possible *ad-hoc* corrective functions into physical models in front of un-modeled or partially modeled phenomena detectable from the collected feedback. Additionally, the need of a fast commissioning execution, in turn, requires that such model calibration activity relies on standard tools, allowing fast analysis of large databases of plant feedback and a systematic approach that excludes possible human mistakes.

In the following a consolidated procedure for setup model calibration will be presented. It involves both online applications and a set of offline tools running in a MATLAB environment. It is based on:

- techniques for unsupervised online feedback data collection and selection;
- suitable routines for offline data analysis, easing the task of plant feedback inspection through the graphical tools;
- automatic identification of the most suitable model corrective parameters/functions based on medium-large feedback data sets.

The paper is organized as follows: Section II provides a description of an automation system for flat metal production that includes a MATLAB-interfaceable setup model; Section III focuses on feedback data collection for model tuning purposes and describes the advantages of MATLAB-based data analysis and model tuning; in Section IV a real example of model calibration activity is presented; Section V provides some conclusions.

II. Architecture of an Automation System for Metallurgical Flat-Products Processing Plant

Automation systems for hot and cold flat metal manufacturing plants are typically based on hierarchical structure including Level 3, Level 2 and Level 1 PCS (Process Control Systems, Fig. 1).

Among the three levels of automation, Level 2 may be considered the heart of technological functions while Level 3 is focused on production scheduling and Level 1 is intended for real-time process control. Level 2, instead, being provided with long-term databases where both production and technological information are stored, is the place where mathematical models are implemented and feedback data collection for both online model adaptation and offline model calibration takes place.

More precisely, the Level 2 system consists of a number of tasks (Fig. 2), each one devoted to a specific function (e.g. production management; data sampling; statistical data analysis; logging of plant stoppages and accounting of consumables) running as services on a non-real time operative system. Between these tasks some of them represent the implementation of one or more setup models. Level 2 tasks communicate among them through a TCP message *dispatcher* and, at the same time, each task performs read and write operations on a relational

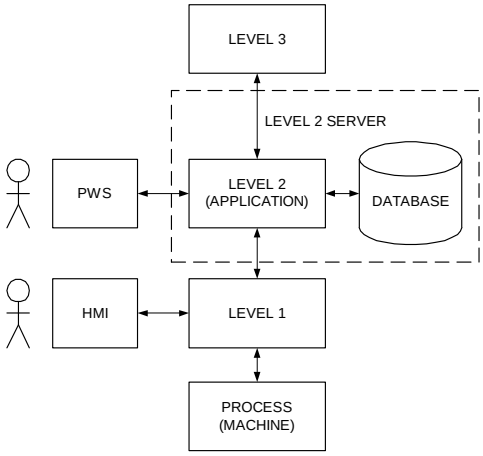


Figure 1. Hierarchical structure of automation system for flat products.

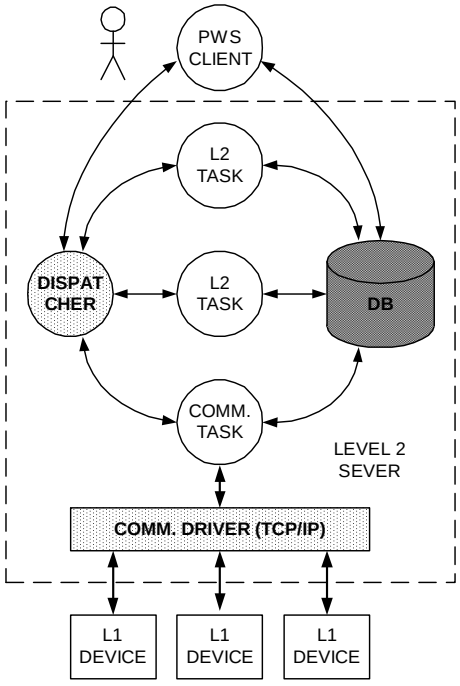


Figure 2. Structure of Level 2 PCS.

database. The Level 2 PCS is completed by one or more clients, where the so-called Process Workstation (PWS) is installed. The PWS is the Level 2 graphical interface, which allows the operator interacting with the production and the control system.

From a software point of view, a setup model consists of an executable file and by one or more dynamically-linked libraries that will be referred to as the "TASK" and the "DLLs", respectively. The model TASK is in charge of managing the communication with the other Level 2 tasks, with the Level 2 clients, with the Level 1 and managing the interface to the database. The mathematical optimization engines are instead implemented in the DLLs that can be called by the model TASK at the occurrence (Fig. 3). This structure has the advantage of separating the customizable part of the model (i.e. the TASK) from the standard one having a technological content (i.e. the DLLs). The DLLs are developed as standard components which, at the same time, can be interfaced to the MATLAB environment for model tuning purposes, as

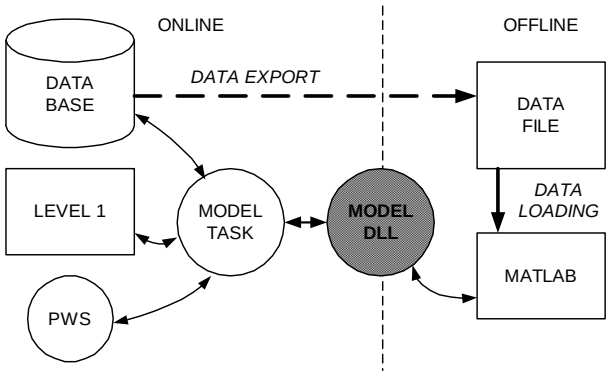


Figure 3. Model DLL for both online and offline utilization.

discussed in the sequel.

III. Feedback Data Collection and MATLAB-Based Analysis for Model Calibration Purposes

The basic concept behind the calibration activity for setup models is that possible deviations between *model predictions* are the *plant process measurements* collected in a database are subject of analysis. Of course this is a preliminary activity for then properly proceeding with tuning activities i.e. for adjusting some tuning parameters and/or for introducing additional corrective functions into the physical models so as to compensate partially modeled phenomena observed from data.

During material processing a feedback is collected both in *transient* and in *steady-state* conditions: for instance, when the processing speed is subject to accelerations or deceleration, then the corresponding data collection is discarded since it is considered not suitable for model calibration purposes. During steady-states, instead, all the main actuation variables are subject to minimal variability in time: in these conditions a reference working point can be extracted from the recorded signals by suitable *signal processing* techniques.

Typically, Level 1 transmits feedbacks with a sampling cycle time between 0.05 and 1.0 s, depending on the process dynamics and on the need for more or less detailed quality evaluation of final products. Level 2 is provided with a task devoted to data collection and analysis. Indeed, such a task automatically recognizes the steady-states in plant feedback and extracts the useful information. The extracted data from each steady-state is then stored into the historical database (Fig. 4).

These data can be then exported from the database (Fig. 4) and then loaded in the MATLAB environment (Fig. 3), where the model can be re-executed though the direct link realized between the model DLL and the MATLAB environment.

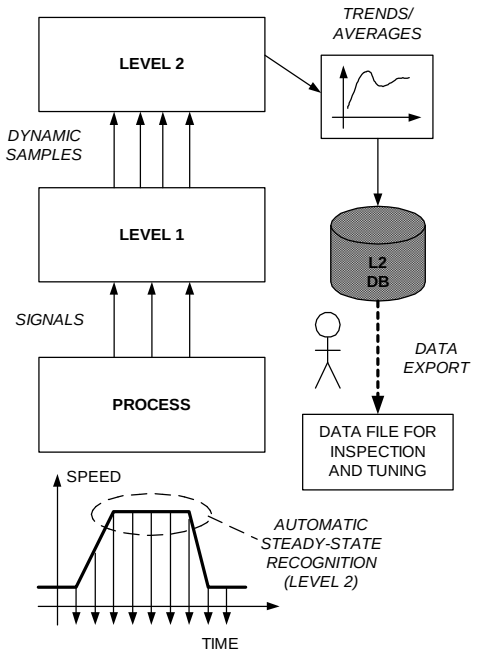


Figure 4. Online data collection and steady-state recognition.

In such a way the model outputs can be compared with the corresponding feedback signals. Indeed, the possibility of using the same DLL for both online and offline application gives the following key advantages.

- This approach guarantees consistency between online and offline calculations. The alternative solution of re-implementing the model code directly in the MATLAB environment involves the risk of misalignment between the two different software implementations. Numerical problems and implementation errors can compromise the consistency of calculations. The maintenance of two nominally equivalent implementations of the same model is of course time wasting.
- The compiled model code in the DLL format is definitively much more efficient than other possible alternative. The need for large data sets analysis requires short computational times to be repeated and verified in several consecutive phases.
- The verification of the model calibration result is to be executed by re-running the tuned mathematical model (which will be available in the online task) on a large amount of data that can be different from the data-set used for calibration (reliability check).

Moreover, the MATLAB environment offers powerful tools for graphical and statistical analysis of data, together with the possibility of creating personalized libraries of functions (toolboxes).

IV. MATLAB-Based Offline Model Calibration: A Case Of Study

The calibration activity here presented can be applied to an aluminum hot mill model. In this plant, hot plates coming from a furnace with a thickness of 450÷600 mm and a maximum temperature of 550 °C are rolled up to a final thickness of 2.5 mm. When the piece thickness drops below a certain threshold, rolling process switches from plate to coil mode, while the piece temperature may drop down up to 300 °C or less; a whole rolling schedule from plate to coil may require more than 30 rolling passes, thus giving rise to a fairly wide range of working conditions.

Accuracy of rolling mill models is, for instance, evaluated through the prediction error of the rolling force required to reduce the material thickness. The resistance of rolled material to plastic deformation is defined Yield Stress (YS) typically expressed in *Pascal*s. For the aluminum hot rolling case the Yield Stress is a quite complex function of chemical composition, temperature, strain and strain rate [6]. Hot Yield Stress is commonly modeled through the so-called *Constitutive Equation*, here expressed in the following form:

$$YS = \alpha\left(\left[chem\right]\right) \times f\left(T,\varepsilon\right) \times g\left(T,\dot{\varepsilon}\right) \qquad (1)$$

where [*chem*] is an array of percentages of chemical additives characterizing an aluminum alloy; T (°*C*) is the

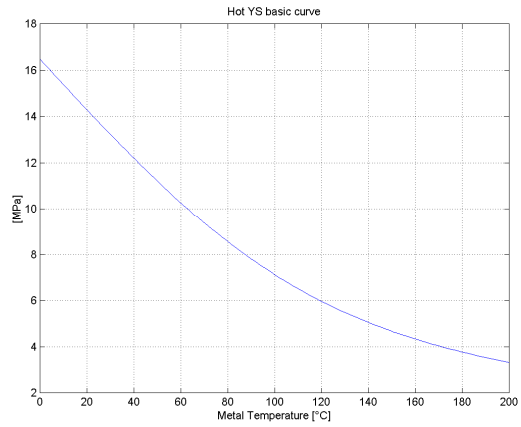


Figure 5. Basic Yield Stress curve.

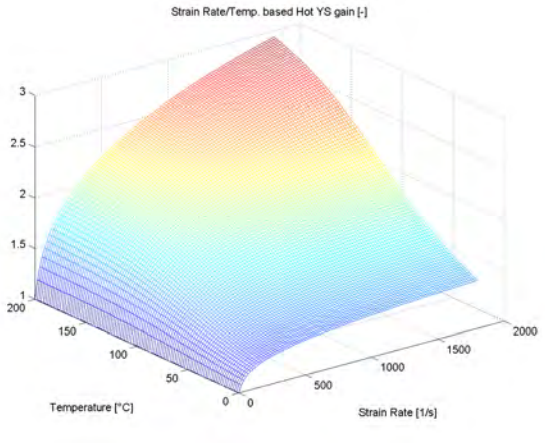


Figure 6. Strain rate/temperature-based YS corrective term.

metal temperature; ε (adimensional) is the *strain*, i.e. the amount of strip deformation; $\dot{\varepsilon}$ (*1/s*) is the *strain rate*, i.e. the deformation rate. The Yield Stress model can be rearranged in the following way:

$$YS = \alpha \times f(T,\varepsilon_0) \times g(T,\dot{\varepsilon}_0) \times \frac{f(T,\varepsilon)}{f(T,\varepsilon_0)} \times \frac{g(T,\dot{\varepsilon})}{g(T,\dot{\varepsilon}_0)} \qquad (2)$$

where ε_0 and $\dot{\varepsilon}_0$ are small (constant) quantities used as normalization references. By properly grouping terms of Yield Stress equation (2), the following equation can be obtained:

$$YS = YS_0([chem], T) \times f'(T,\varepsilon) \times g'(T,\dot{\varepsilon}) \qquad (3)$$

The proposed re-formulation of the original Yield Stress Model allows decoupling the effects of temperature-strain and temperature-strain rate (Fig. 6) from a "basic" Yield Stress curve (Fig. 5) which depends on chemical composition and temperature only.

The calibration activity here discussed is devoted at tuning the Hot Yield Stress model. Along a period of three months several thousands of steady-state samples can be unsupervisingly collected by the Level 2 PCS for different alloys and for different processing regimes. Data are then automatically transported in the MATLAB environment where a preliminary pre-filtering is executed for data redundancy check and for discarding the less reliable data samples (for instance, those corresponding to too short steady-states). Several thousands of piece rolling schedules including both low and full speed steady-state samples are finally selected for data analysis. This final data-set includes some tens of different alloy families, each one characterized by a specific chemical composition.

The first phase of calibration job consists of running the un-calibrated model and comparing the model output (predicted forces) with the corresponding feedback. This phase mainly aims at looking for possible correlations between model errors and some other process variables. Indeed, one of the most interesting features of MATLAB-based model inspection is the possibility of plotting the model prediction error(s) versus a selected set of technological variables (see Fig. 7) to find out possible model "polarizations" which require corrective actions.

In the discussed case the un-calibrated model exhibits a fairly good behavior for plate rolling case, while coil mode passes are subject to some uncertainty (see Fig. 8).

Coil rolling, with respect to plate rolling, is characterized

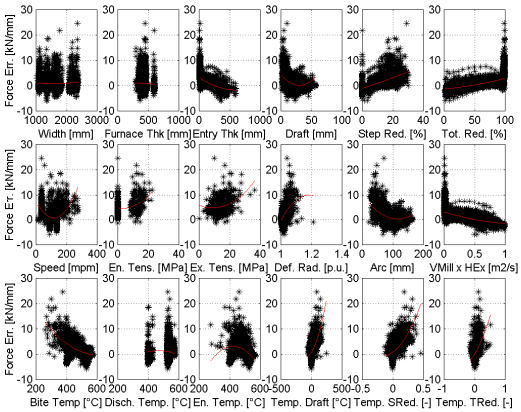


Figure 7. Force prediction error "polarizations" (before of calibration).

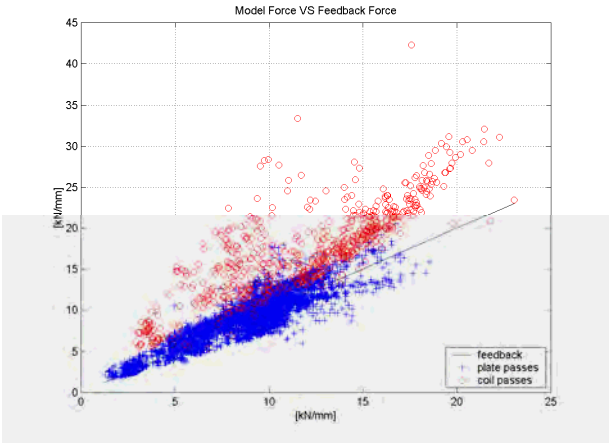


Figure 8. Force prediction error for plates and coils (uncalibrated).

by lower thickness, lower temperature and higher strain rate. Inspection of un-calibrated model output suggested focusing the tuning efforts on function $g(T,\dot{\varepsilon})$. This term comes from the well-known *Zener-Hollomon equation* [6], which is suitable to be rearranged in a number of ways making it tunable for different kind of materials. In the discussed case the tuning parameter is an exponent, defined as *alloy constant*, applied to the *Zener-Hollomon parameter*:

$$Zener' = Zener^{\frac{1}{N_{alloy}}} \tag{4}$$

First of all, an iterative algorithm has been used to identify, for each alloy, the value for N_{alloy} which minimizes in a statistical sense the error between model prediction and feedback force. Secondly, the so identified values for N_{alloy} are correlated with the chemical compositions of the considered alloys (Fig. 9) in order to derive an analytical expression for N_{alloy} as a function of the chemistry [*chem*]. In such a way the new function $N_{alloy}([\text{chem}])$ is expected to be valid also for the alloys not included in the data-set used for the calibration procedure.

In other words, N_{alloy} will be not anymore an input

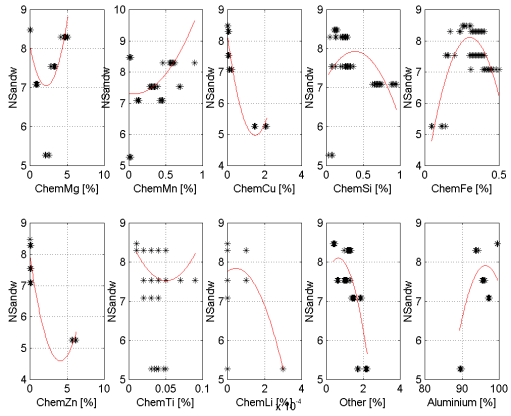


Figure 9. Correlation: alloy constant and chemical elements.

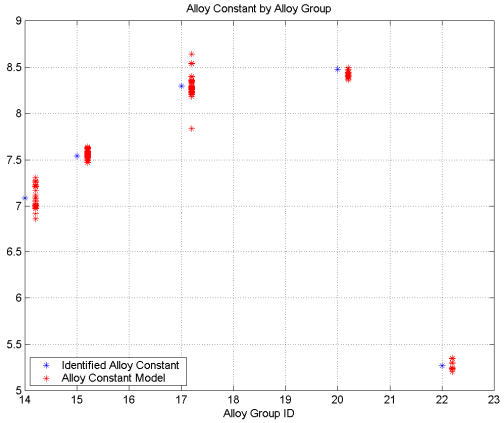


Figure 10. Alloy constant matching force and identified model.

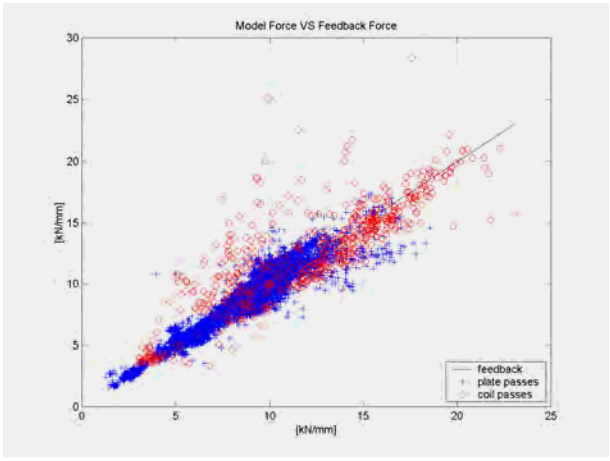


Figure 11. Force prediction error after alloy constant tuning.

parameter of the model, but the result of a calculation that exploits other model inputs like the chemistry [*chem*] (Fig. 10) This first model calibration step greatly improves the rolling force prediction in coil mode (Fig. 11).

The second step of model calibration job aims at identifying a *Yield Stress Offset* through a regression-based method, where a corrective term is introduced into the physical model (see eq. (5)), as a function of some key technological parameters and/or chemical species. More precisely, the YS Offset is introduced in the *Constitutive Equation* in the following way:

$$\begin{cases} YS &= \left(YS_0(T) + YS_{OFFSET}\right) \times f'(T,\varepsilon) \times g'(T,\dot{\varepsilon}) \\ YS_{OFFSET} &= h\left(p_1, p_2, \ldots, [chem]\right) \end{cases} \tag{5}$$

where [p_1, p_2,…] represent the selected technological parameters (e.g. thickness, speed, …). The aim of this further calibration step is removing the residual model polarizations of the prediction errors with respect to other measured information (graphically shown in Fig. 12).

The final result of model calibration can be appreciated from Fig. 13 since the dispersion on force prediction error has been further reduced. Fig. 14Figure 14. highlights that all the model polarizations with respect

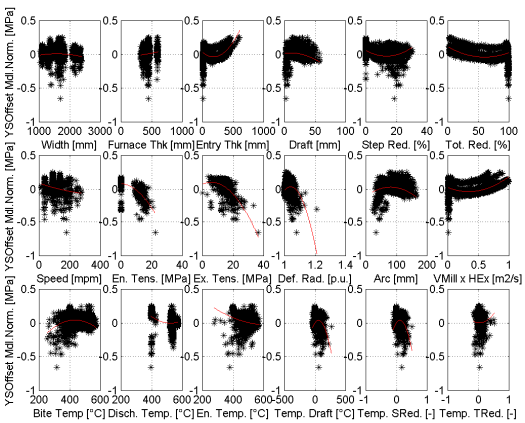


Figure 12. Yield Stress Offset correlations.

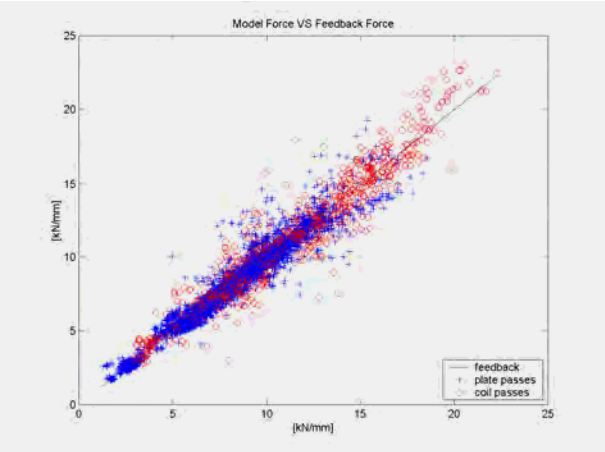


Figure 13. Calibrated Force Model.

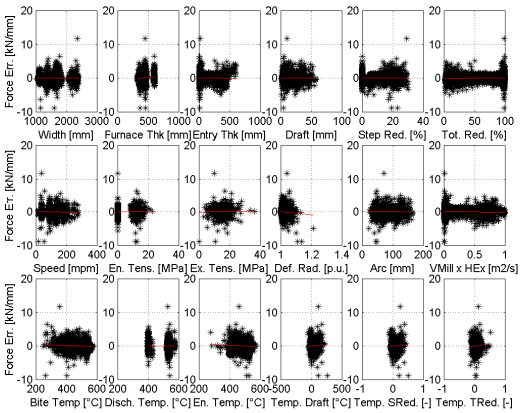


Figure 14. Force prediction error "polarizations" (after calibration).

to the considered technological variables have been considerably attenuated. The final result in terms of model force error is remarkable: the percentage of analyzed records having a prediction error lower than 5% changes from 67% to 97% (Fig. 15).

Calling the precompiled model DLL from MATLAB for all the selected records requires about 30 seconds on a computer with Intel i5 CPU. The calibration routines are incorporated inside the model DLL with a procedure that allows incorporating *automatically generated* code starting from the data-analysis procedure so as to avoid any possible error or inaccuracy associated to manual code programming.

V. CONLUSION

In this paper a systematic approach to feedback-based calibration of setup models for flat metallurgical applications has been discussed. In the last years, adoption of mathematical models into the automation control systems applied to flat products has become essential to cope with the high quality standards required by the

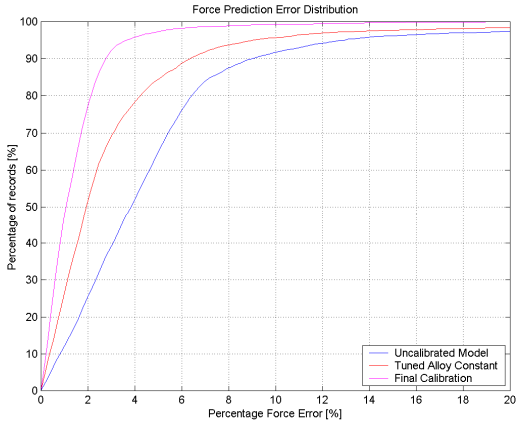


Figure 15. Distribution of force prediction errors.

competitive market. Calibration of models – aimed at compensating un-modeled or partially modeled phenomena – is a key-point of commissioning activity to make their use really effective. Tuning job requires deeply analyzing the deviations between model predictions and process measurements for large database dimensions. A possible procedure for model calibration has been presented: online tools provide automatic collection and pre-selection of plant feedback data and offline MATLAB-based routines allow for fast analysis of such large data sets and effective model tuning. Implementation of mathematical models in the form of standard DLLs – which can be both integrated in the online applications and linked to the MATLAB environment – guarantees the maximum reliability of model tuning results and short computational times.

REFERENCES

[1] W.L. Roberts, "Flat Processing of Steel", Marcel Dekker, New York, 1988.

[2] E. Garber, A. Traino, and I. Kozhevnikova, "Novel mathematical models for cold-rolling process", Flat-Rolled Steel Processes Advanced Technologies, pp. 179-190, CRC Press, V.B. Ginzburg eds., 2009.

[3] F.A. Cuzzola and T. Parisini, "Automation and control solutions for flat strip metal processing", The Control Handbook, CRC Press, W. Levine eds, 2010.

[4] C. Aurora, D. Sclauzero, F.A. Cuzzola, "Pickling Line Modeling for Advanced Process Monitoring and Automation", L. Ferrarini, C. Veber (editors), Modeling, Control, Simulation and Diagnosis of Complex Industrial and Energy Systems, Chapter 2 International Society of Automation (ISA/O3NEIDA), N.Y., 2009.

[5] C. Aurora, D. Cettolo, F.A. Cuzzola, "Cut scheduling optimization in plate mill finishing area through mixed-integer linear programming" IEEE Transactions on Control System Technology, Vol. 18, No. 1, pp. 118 – 127, 2010.

[6] T. Sheppard, A. Jackson, "Constitutive equations for use in prediction of flow stress during extrusion of aluminium alloys", Materials Science and Technology, Vol 13, No. 3, pp. 203–209, 1997.

Architecture and implementation of a MES system in a large scale steel plant: Severstal Cherepovets success story.

Gianpiero Brunetti*,
* Vice President Information Technology - BU Steelmaking - at Danieli Automation S.p.A., Buttrio (UD)
g.brunetti@dca.it

Abstract – Implementation of an MES (Manufacturing Execution System) in a large scale steel plant raises several technical and organizational challenges. Danieli Automation summarizes in this article its experience starting from the initial stages of the project when the overall approach was defined: from dividing the implementation in steps down to analyze the overall architecture and functionalities to be provided.

I. INTRODUCTION

Implementing an MES in a large scale steel plant poses a series of problems of a technical and organizational nature that must be addressed and solved in tight collaboration with the customer. When we talk about large scale steel plant, we refer to an integrated steel plant (i.e. a plant that starts its production cycle from coke and iron ore) producing cold and hot rolled products and with a production capacity per year which exceed the 2 Mt. The number of users of an MES system in such implementations exceeds usually the 500 and can reach several thousand in biggest plants.

Severstal is one of the world's leading vertically integrated steel and steel related mining companies, with assets in Russia, Latvia, Ukraine, Poland, Italy and Liberia with approximately 52,000 employees (2014).

Cherepovets Integrated Iron & Steel Mill, the major steel asset of the Division, was founded in 1955. The Iron & Steel Mill is an integrated steelworks of full production cycle integrating more than 100 major process units, from iron ores and coals processing to finishing operations where high-value products are produced. The Cherepovets Mill has a production capacity of 11.6 Mt (2013).

Danieli Automation has developed its MES solution since the early '90, its latest version is named Q3MET and is the result of a large number of projects completed worldwide.

Project Complexity, in this scenario, is not only linked with the dimension of the site but also with the architecture of the overall solution. In Cherepovets, and the same situation can be found in similar plants, a plethora of existing system is usually available and there are well known, although not always well documented, operating practice which increase the complexity of the analysis and architectural phase.

II. APPROACH

Therefore the first goal of the project was the definition of the business process and of the overall system architecture. The results of this initial stage are the following deliverables:

- Business Process Analysis and Design
- Architecture Landscape As Is and To Be

The above results are then used as a blueprint for all the steps of the project which proceeds then with subsequent implementation of sub-portion of the entire plant. The reason for this strategy is that the MES system of a brownfield project of this size, with the plant in operation, cannot be replaced in a single shot simply for organizational reason, in fact neither the supplier nor the customer have the resources to test, train users and support the Go Live for the whole site. Therefore the implementation must be done in steps according to organizational structure of the site and decoupling point of the production process.

III. BUSINESS PROCESS DESIGN

The first step was to analyze and design the business processes where a business process is a set of related business activities that are combined to provide something having a value (e.g. goods, services or information) to a customer. Despite traditional organizational chart which organize activities in functional units a business process is a horizontal flow that crosses organizational department.

The defined business processes are then the starting point for the analysis of the MES functionalities and a reference for the future organization of the business.

The introduction of a new MES system in a large organization is not only an IT project but it is also a change management process aiming to implement new software solution and business changes. The role of the customer in managing the internal change is than crucial for the successful results of the project.

Considering the scope of the MES, focus was placed on the process of delivery goods against a customer order (MTO Make to Order). Other internal processes, which have no relation with current implementation for example processes for maintaining customer account, to check customer credit, receive payables and others which are not

related with MES implementation, were not analyzed. The result of the Business Process Design is a set of documents each of them describing a defined process. On top of everything there is an overall business process which represent the value chain of the customer and that provides an overview of all identified business process and their interaction. A portion of such overall process is represented in below figure 1 where it is possible to identify two types of processes the one drawn using a dashed rectangle that is named external process because implemented in other system than the MES and those which are part of the MES implementation and are represented by a continuous line rectangle.

For each Business Process the design document provides a description of inputs, outputs, involved organizations; a graphical representation of the process; a description of each individual step including its inputs and outputs; the involved business actors and the proposed KPI (Key Performance Indicators) used to measure performances.

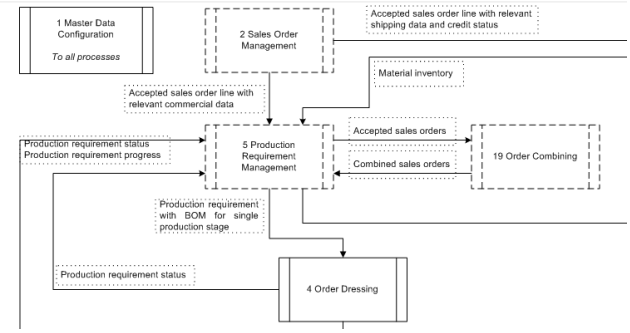


Figure 1. Portion of the overall Business Process Design

IV. ARCHITECTURE

The high level customer requirement for architecture design can be summarized as follows:

- Define a unique solution for all department;
- Having a centralize database for each production site;
- Manage Master Data in a centralize way.

On the basis of the above high level requirements three main architectural aspects have been considered:

- Overall System Landscape: describing the involved systems, their interaction and the mapping with the Business Process Design;
- MES Database Architecture: defining the MES database in terms of a centralize or distributed database structure;
- MES Hardware Architecture: defining system requirement for hardware and network devices;

The overall system landscape includes the definition of the As Is architecture, which describe the systems in use in the plant and their actual interaction. To define the As Is architecture an initial list of systems in use including their main functionalities has been defined. The list has then been converted in a block diagram depicting systems and their interactions including definition of main interface messages.

Considering the functionality of the supplied MES system a To Be schema has been designed, where legacy systems, which functionalities are part of the MES scope, have been replaced.

Considering the complete system is implemented in steps, department by department, intermediate To Be schemas have been also defined.

The business process design has been then mapped on the To Be architecture in order to identify possible gaps and to define a preliminary list of interfaces between systems. The focus was in understanding the processes that are going to be automated; the mapping of the activity to selected systems assuring that a solution is selected, configured and implemented to support them.

Another step was to define a database architecture to support the system implementation consider its further extension to remote site (production site part of Severstal group but located in different geographical areas). Analyzing business processes it was understood that there are operations which are performed by centralize structures and other that are distributed to individual department. Other aspects to be considered are reliability of network connection and requirement for independency of specific department/site.

The result is database architecture, with

- Centralized schemas: which are used for all departments and also for the remote site;
- Centralized schemas per site: which are used for those processes that shares data with all department but that are too crucial that connectivity issues requires to have a local copy in each site;
- Department schema: a database schema for each department to manage production, inventory, quality and shipping data.

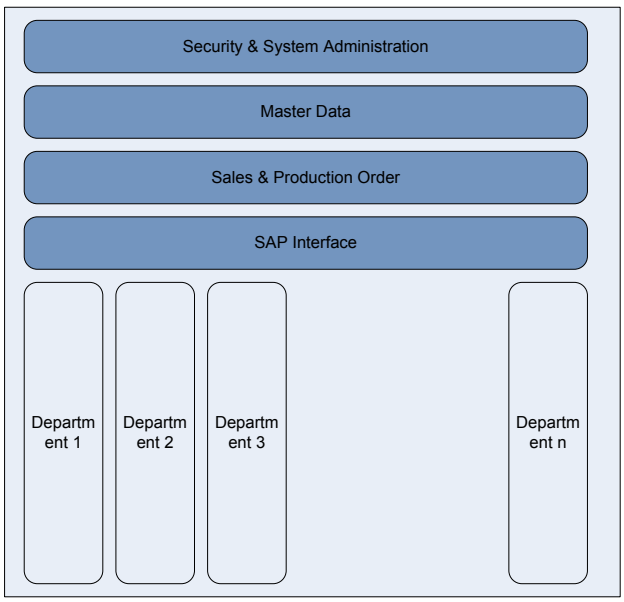


Figure 2. Database Organization

Last step to complete was the definition of the hardware and network requirement; this was prepared considering the database architecture, the number of users, the estimated transactions and the data volume.

V. INTEGRATION WITH OTHER SYSTEMS

One of the results of the architectural analysis is a consistent reduction of the number of systems used to implement the selected Business Processes, with a consequent reduction of system interfaces and interface messages. To further simplify the maintenance activities all interfaces has been designed using a single interface technique which is based on interface tables in database.

The MES system has a central role in all data flow between the shop floor and the ERP (Enterprise Resource Planning) system. The MES system receives Customer and Production order from SAP, convert them in a production program that is then transferred to the MTS (Metal Tracking System), a system dispatching and collecting data from downstream automation that was introduced to provide a standard interface on top of existing automation architecture.

The MES system is also interfaced with some external systems which support logistic operation in the plant, and some legacy systems providing specific quality functionalities. The first generates shipping documents and manage railcar tracking the latter collects NDT test (Non Destructive Test) and perform calculation to estimate quality property of the produced material.

VI. MODULE FUNCTIONALITIES

The System consists from a functional point of view of

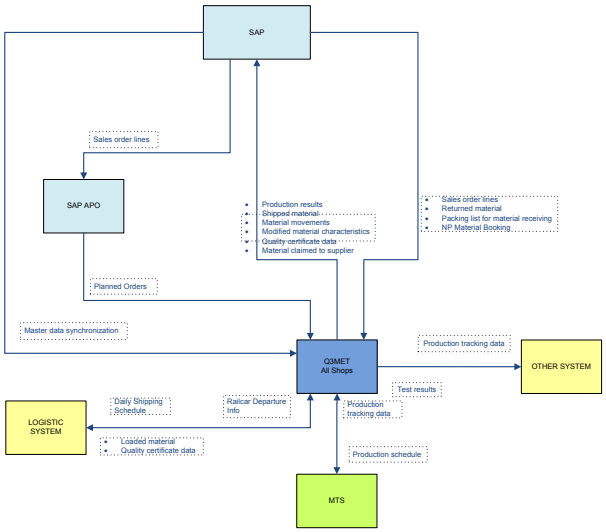


Figure 3. Interfaced System

the following modules:

- Master Data Management;
- Order Dressing;
- Production Progress Control;
- Stockyard Management;
- Shipping Management;
- Quality Management;
- Business Intelligence.

A. *Master Data Management*

Master Data Management module consists of two main sections. The first one is used to collect and maintain all those technological parameters and plant configuration for supporting the activities of the system. The second one is used to configure the other modules.

The technological parameters section includes catalogues of commercial and internal grades, product, profile types and international standard that are used to describe the steel products. Plant configuration data includes production unit and production routing which are used to describe the sequence of operation and equipment required to produce material required by a specific order.

In addition to the static technological catalogue Q3MET provides a fully dynamic, flexible system for configuring commercial and technological parameters, known as Attribute Catalogue. Attribute are then used to describe Customer Order, Production Order and produced material characteristics. Each attribute identifies a specific characteristic of the material such as the width or the length or the tolerance. This function is in charge of storing all the plant business rules required by each Plant Unit involved in the actual routing of production. Any rules which may change over time will be adapted in this catalogue directly by the final key users, without any programmers' assistance.

Master Data module is also used to maintain data catalogues of main modules and to configure them, more in details the system provides following functionalities:

- Stockyard Configuration: it includes configuration of storing location and management of "on hold" (a catalogue of reasons to quarantine the material) and material defects;
- Shipping Configuration: this includes a catalogue of vehicle classes and a catalogue of loading instructions. The latter defines the instruction for vehicle loading based on type of dispatch and material dimensions;
- Quality Configuration: including property subject of destructive testing such as yield, elongation, charpy and so on and a set of configuration tables that defines test limits depending on a set of property of the order (standard, grade, size and so on). It also includes configuration for Quality Certificate printout;
- Production Progress Control Configuration: including a catalogue of reasons to describe production stoppages.

B. *Order Dressing Management*

The Order Dressing Management module supports the organization by identifying the Customer Sales Order and the Production Order information required to provide other subsystems with the information they need to schedule, produce, store, certify and ship finished products.

Main task of the module is the *dressing* of the order i.e. the operation required to complete the order with all technical information required for its production including selection of the production routing. Authorized users are entitled to view the overall situation of all order types, all attributes of the selected order are available in an intuitive layout, and any inconsistencies with business rules will be clearly shown highlighted.

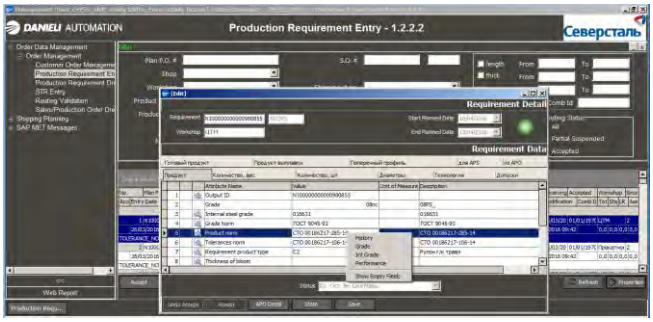


Figure 4. Order Dressing Management Form

C. *Production Progress Control*

The module receive production schedule from the interfaced APS (Advanced Planning System) system and transfer it to the concerned MTS with all required information to implement the production program. The module includes background functionalities which collects production results from the MTS interface in automatic mode. As a result of received production report the module update the production program in order to reflect actual progress of the production, it consumes input material and it generates information related to the produced material.

In the event that all or part of the concerned MTS is out of operation for any reason, such as planned maintenance or damage, and the Plant is enabled to continue operating by manually introducing operational instructions through automation systems or the concerned MTS attempts to pass data to MES that rejects them due to inconsistencies identified in the message content, a set of functions allows the user to enter production results directly into the Q3MET system.

D. *Stockyard Management*

The main purpose of the module is to manage the material movement within the plant and to maintain traceability data about produced material.

The module manages all kind of material movements except those related to the dispatch to final customer which are in charge of the Shipping module. There are two main types of internal movement: the one within the same department which can require or not acceptance by authorize destination yard user; and the one between different departments which is managed using internal Load Request i.e. a formal documents containing the list of goods, the destination and the vehicle used.

The module provides a graphical representation of the yard, each location is represented by a corresponding rectangle and the percentage of occupation is represented using a set of colors. User is allowed to move pieces between different locations by simply using drag and drop functionalities which results in a useful tool to correct inventory status.

The module manages the inventory in terms of information of individual production entities (a slab, a coil, a plate a bundle of bar and so on). For each of this entity the system displays material characteristics, order data, quality characteristics, and traceability information including: history of piece movement, history of the piece and material genealogy.

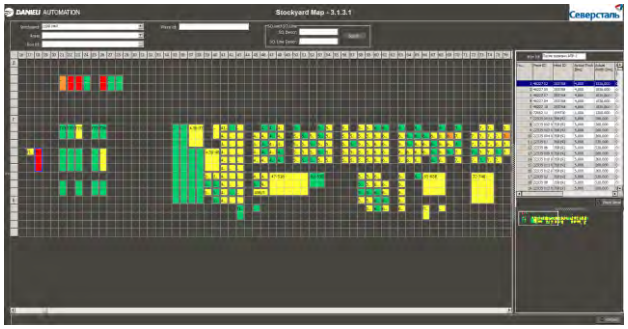


Figure 5. Yard Map

E. *Shipping Management*

The module manages dispatch of finished goods to customers. It includes a set of inquiries that allows the user to get a forecast of shipment in a selected period of time including customer order data, required dimensions, order progress (i.e. quantity ordered, quantity produced, quantity already shipped and so on), destination railway station and an estimation of the number of vehicles required.

Operating practice requires preparation of a Daily Shipping Plan, this is a daily plan of shipment for a specific stockyard, a transportation mean (railcar, truck) which defines for each shift the orders and the quantities to be dispatched. The system based on configured rules provides a preliminary estimation of the vehicle required.

Starting from the prepared plan, a set of Load Request are then generated by user using system functionalities. During this process individual pieces belonging to the selected customer order and ready to be shipped are selected. The Load Request is then updated when material is physically loaded and vehicle assigned after completion of this operation the Quality Certificate for the load is generated.

F. *Quality Management*

Quality Management is the module used to manage the certification and release of the material.

The module manages the verification of produced material against technological and dimensional parameters; user can decide whether produced material is within acceptable requirements or otherwise decide to quarantine it or to downgrade it. Next step is related to the quality certification of the goods. Material is grouped in so called "certification group" while representative samples are addressed to the laboratory to be tested. The test results are then compared to Customer Order requirement and used to certify material lot. Certified material, which is a finished product and has not suspension (on hold) of any type become "ready to ship". Upon issuing the Bill of Lading the system generates the corresponding Quality Certificate in compliance with configuration and rules defined in the Master Data.



Figure 3. Sample data management

G. *Business Intelligence*

Q3Intelligence is the Business Intelligence and Analytics module which provides advanced tools to analyze, through metrics and KPIs (Key Performance Indicators), the performances of each Business Process inside the company, using the data from the Q3MET system.

The module can help plant managers and business unit leaders to keep under control each single step of Sales Order processing, from order entry to shipping, thanks to a continuous monitoring of the performances. Each single user, among production areas or business units, has the opportunity to browse production and business information thanks to a centralized data repository together with fully customizable dashboard templates, based on a designed KPI Library.

As a strategic decision support tool, Q3Intelligence is based on the following key features:

- Deep global industry experience;
- Centralized data repository, as the source for multidimensional data analysis and reporting;
- KPI Library designed on the specific needs and requirements of the defined Business Processes.

Monitoring of the performance allow a continuous improvement of the designed business processes: reducing the total production cost, getting lower direct costs, and improving product quality and customer service level.

VII. CONLUSION

There are important achievements obtained with the above described solution:

- Centralization of Master Data and Business rules in a unique repository allows a centralize management of this important assets of the company, that become less dependent on individual user decisions and which simplify managing unified business processes among different departments;
- The definition of a common solution based on a set of well-defined Business Processes allows to deploy the same solution in all department providing a uniform and rationale management of the business, which result in simplifying training process and a reduction of the TCO (Total Cost of Ownership) of the MES;
- The architecture defined is used for all shops and this reorganization of the IT architecture reduces the number of involved systems with consequent positive effects on maintenance of the IT infrastructure;
- The project is driven by a set of qualitative and quantitative goals; the latter ones are measured by means of specific KPIs detecting business performances.

ACKNOWLEDGMENT

Thanks to all member of the Danieli Automation team for their dedication and professionalism; the work was carried out with great intensity in a climate of commitment and enthusiasm allowing the first shop (Converter Shop) to Go Live after only 13 months from the project "kick off" meeting.

REFERENCES

[1] ANSI/ISA-95.00.01-2010 (IEC 62264-1 Mod) Enterprise-Control System Integration - Part 1: Models and Terminology

[2] Severstal Web Site https://www.severstal.com/eng/about/index.phtml. .

PAPERS

Anfis as a Method for Determinating MPPT in the Photovoltaic System Simulated in Matlab/Simulink

Dragan Mlakić, MSEE[1], Full Professor Srete Nikolovski, PhD[2]
[1]Electric power company HZ-HB Inc. Mostar
[2]Faculty of Electrical Engineering Osijek, Croatia

Abstract - Photovoltaic systems as part of larger distribution systems for the production of electricity have non-linear current-voltage characteristics. These characteristics depend on the temperature of the solar cells and insolation of panels. The method for the search of the maximum power or MPPT is widely known, and there are many offered algorithm solutions. The aim of this paper is to present ANFIS (Adaptive Neural Fuzzy System Inference) as a method of computer intelligence in the search function of maximum power or maximum utilization of the PV system, without prior knowledge of the system. MPPT ANFIS will be located between the PV panels and devices for voltage regulation. Inputs used to go inside the fuzzy regulators and outputs of the fuzzy regulators are going into power changing device. The advantage of this method is the low cost of implementation because it uses embedded computer systems and does not require additional sensors for measuring temperature and insolation. It is very easy to upgrade in the case of increasing the knowledge base used in FIS membership function. Everything is simulated by MATLAB/Simulink software on the one-phase IBC system.

I. INTRODUCTION

The renewable energy systems are well known due to their advantages of natural fuel and eco-friendly nature. Among them, solar Power Voltage based electric power generation technology has remarkable developments in the last few decades. The power generated by the PV cells is affected by the temperature and solar radiation. The changes in temperature and/or solar radiation results change in output power of the PV module instantly. The maximum power point tacking techniques are used to ensure the maximum power delivery from the PV source to load. This technique continuously tracks the output of the PV and adjusts the operating point of the power conditioning unit whenever required, and generate the maximum power at all instances [2]. Most of MPPT algorithms were presented in the literature [3-13], but many of them are limitated due to non-linear nature of solar PV module characteristics. Artificial intelligence (AI) based MPPT methods presented in literature [6-10] can track the MPP more efficiently under changing environmental conditions: working temperature and irradiance level. AI-based methods (in this case Fuzzy logic and artificial neural networks) are the most suitable for improvement of the dynamic performance of the maximum power point tracking. The fuzzy logic and artificial neural networks (ANN) are two main AI methods used for many automation computing not just in MPPT problem but far more. ANFIS combines the advantages of ANN and fuzzy logic, hence it deals efficiently with non-linear behavior of the solar PV modules. Considering the non-linear characteristics of the solar PV module, the AI method ANFIS offer fast, flexible and computing challenging solution for the MPPT problematic. This report, design and implementation of ANFIS-based MPPT model, who is interfaced with an open loop boost converter, is presented. Designing of the open loop DC-DC boost converter is also carried out. It is used for impedance matching and maximum power transfer between the load and solar PV module [2].

II. MAXIMUM POWER POINT TRACKING

The typical voltage of a PV cell is approximately 0.7V. PV cells are connected in series-parallel combination to achieve the desired output power. Figure 1 shows the voltage-current and voltage-power characteristics of a PV cell. P_{mp} is the maximum output power of the cell [5].

The maximum power point changes whenever there is a change in temperature and/or solar radiation. The PV module model based on mathematical equations is developed using MATLAB/Simulink [4-6] and the model solar cell located inside MATLAB/Simulink. The module model presented in [10] is the generalized model that can be used to analyze the input/output characteristics of any PV modules under constant or different temperature and solar radiation levels. The specifications of PTL10 module is given in Table 1 [3]. The model used in this paper is illustrated in Table 1 under different operating temperature and the irradiance level [5] [7].

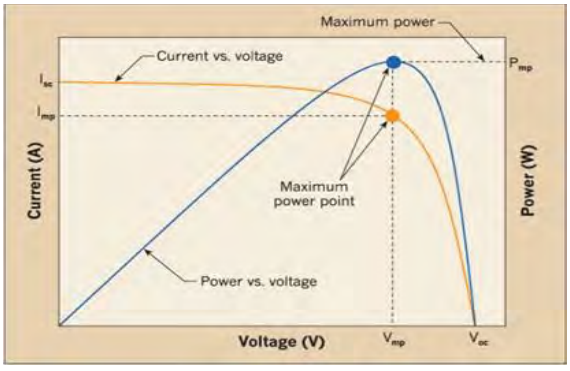


Figure 1. I–V and P-V characteristics of a PV cell

Table 1. PTL10 solar PV module specifications

Description	Value
Maximum Power (Pm)	10 W
Maximum Power Voltage (Vmpp)	16.8 V
Maximum Power Current (Impp)	0.595 A
Short Circuit current (Isc)	0.67 A
Open Circuit voltage (Voc)	21.0 V
Temperature co-eff of Voc	-0.38/C%

III. DESIGNING THE ANFIS CONTROLLER

Fuzzy systems, artificial neural networks and neuro-fuzzy systems are examples of the artificial intelligence systems. AI systems are systems that make decisions like humans beings by adapting to the situations and coming with adequate decisions automatically for future similar situations [4-5]. They do that by learning rules from examples. The primary learning mechanism of fuzzy logic (FL) is based on conditional if-then rules, called fuzzy rules, who use fuzzy sets as linguistic variables in antecedent and conclusion parts of decision making. A set of these fuzzy if-then rules can be issued by human experts or can be generated from observed and analyzed data. The main advantage of fuzzy systems is the lightness to interpret knowledge in the rule base. Neural networks get inspiration from biological neuron systems found in brain, and mathematical learning theories [5]. They are determinated by their learning ability with a parallel-distributed structure and can be considered as black box designing. [2]

The adaptive neuro-fuzzy inference system (ANFIS), proposed by Jang [8], is one of the examples of neuro-fuzzy systems in which a FL system is embedded in the framework of adaptive networks [5-7]. ANFIS constructs an input-output mapping based on human knowledge (in the form of fuzzy if-then rules) and generated input-output data pairs using a hybrid algorithm who is a part the least-squares and part back propagation gradient descent method. In this paper, the ANFIS reference model is developed by using the ANFIS editor of Matlab/Simulink software package [1].

Implementation of ANFIS-based MPPT is presented in Figure 2.

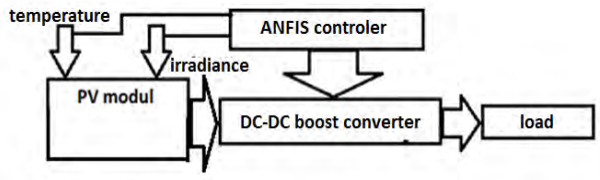


Figure 2. Implementation of ANFIS-based MPPT controller inside the PV system

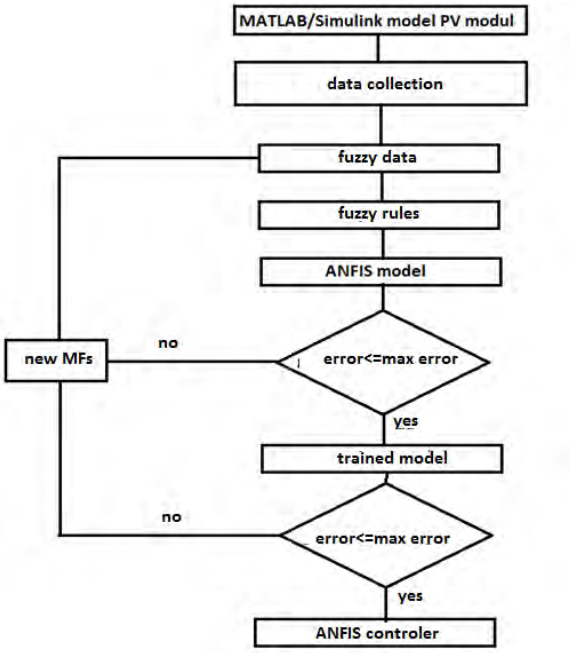


Figure 3. Workflow diagram for the ANFIS controller

ANFIS is able of coming with the input-output mapping of training data sets when trained with sufficient number of epochs [2]. ANFIS generates the set of fuzzy rules in order to produce the appropriate output for different values of inputs. Parameters of membership functions are adjusted or changed till the error is reduced to a minimum value. Once all the parameters of the membership function are adjusted, the ANFIS model becomes a learning model which is ready to be used in the MPPT control scheme [5]. However, before using the ANFIS learning model for the MPPT control, its results are checked by using the checking data which is different from the training data. If a produced error is more than a desired value, parameters of membership functions are adjusted to bring down the error. The DC-DC boost converter is designed to be placed between a solar PV module and a load in order to transfer the maximum power to load by changing the duty cycle of the DC-DC boost converter [2-4].

From the Matlab/Simulink model of the PV module is generated data set used for training of ANFIS by varying the operating temperature in steps of 5°C from -10°C to 35°C and level of solar irradiance in a step of 50 W/m^2 from 100 W/m^2 to 1000 W/m^2. For each pair of the input data (solar irradiance and operating temperature), maximum available power of the PV module is recorded in form of table of crisp data [2-6]. Together, 190 training data sets with 100 epochs used to train the ANFIS reference model and training error is reduced to about 5%. As result, ANFIS constructs a FIS by input/output data sets and membership function parameters of FIS with 50 MFs. MFs are tuned using the hybrid optimization method who is combination of the least-squares type of method and back propagation algorithm. The ANFIS structure developed by the Matlab code is shown in Figure 4 [5-6].

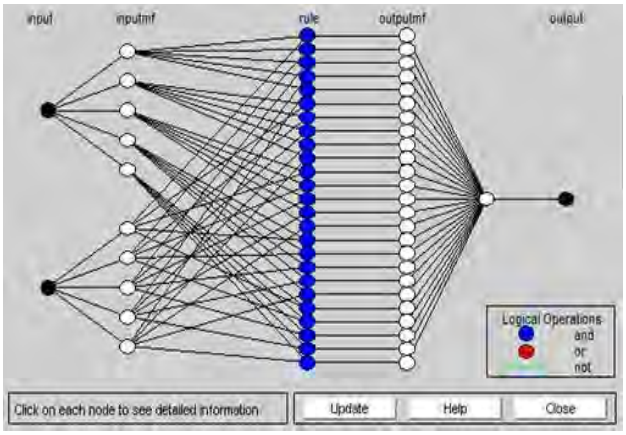


Figure 4. ANFIS-generated levels [2].

Each input parameter has three membership functions that are generated by the ANFIS method. According to the output mapping of data sets, nine fuzzy rules are derived to produce the maximum output power for each value of the input temperature and irradiance level. The ANFIS generated surface is shown in Figure 5, which is a 3D plot between temperature, irradiance and maximum power. The ANFIS surface depicts that the maximum available power of the solar PV module increases with an increase in the irradiance level and moderate temperature that verifies the non-linear behavior of the PV module.

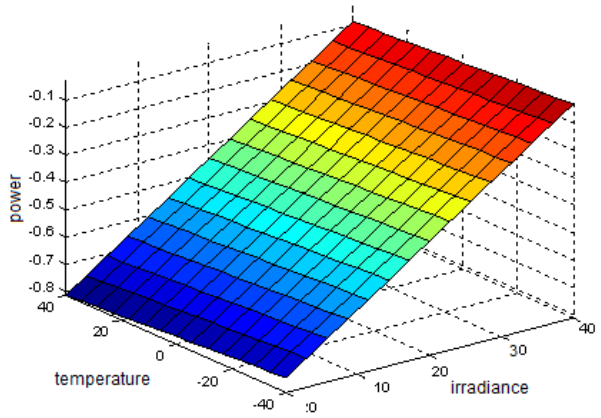


Figure 5. Generated surface for 2 inputs and 1 output

ANFIS-based MPPT is designed by using user friendly icons from Simulink library of Matlab/Simulink. Pure resistance from 0 to 100 Ω is connected on the output of the PV module where it contains MPP of modeled solar PV module on 1000 W/m2 irradiance. Duty cycle of the converter is varied with the help of an MOSFET switch. Simulated adjusting signals for DC-DC boost converter are depicted in Figure 6. Crisp value of the maximum available output power at a specific irradiance level and temperature without load, is inserted from the ANFIS reference model and it is compared with the actual output power taken from the simulated model of the PV module in form of voltage and current.

The divergence between these two powers is a signal who goes into on the DC-DC boost converter which has installed a pulse DC voltage source connected to high frequency P-type MOSFET. The pulse DC voltage source is about 50 kHz and serves as a signal for ON or OFF for MOSFET. From that signal is generated simulation for ON and OFF time according to the irradiance level and working temperature, and by that, the maximum power of the PV module is achieved.

IV. SIMULATION

Block schematics of MPPT is constituted of a solar PV module, DC-DC boost converter and ANFIS model. In this scheme can be added more blocks if we change DC-DC converter structure. The inputs (irradiance level and operating temperature) are taken for the ANFIS reference model and for PV module. The ANFIS reference model outs the crisp value of the maximum available power from the PV module (reference value) at a specific temperature and irradiance level, just like learned from data generation just without load. The actual output power from the PV module, at the same temperature and irradiance level, is calculated by using the multiplication block on sensed operating voltage and current. Two powers are combined and the signal for voltage repair is forwarded to MOSFET inside the DC-DC boost converter.

The DC-DC boost converter turns ON or OFF until the power from the PV module gets higher than the suggested power from the ANFIS module. When the signal for voltage repair becomes “negative”, MOSFET goes OFF and voltage goes down on the output of the DC-DC boost converter. By that, the voltage gets lower until the power from the PV module gets lower compared to the power suggested from the ANFIS module. The block schematic of the maximum power point tracking scheme is shown in Figure 6 [12].

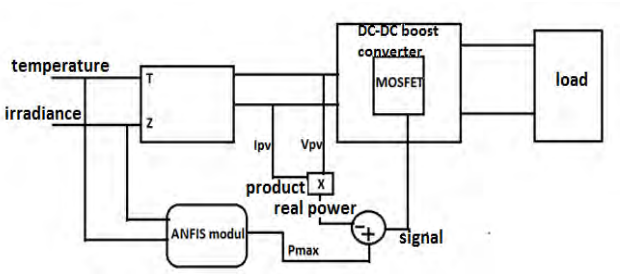


Figure 6. ANFIS-based MPPT control scheme

The DC-DC boost converter adjust the unregulated DC input voltage, supplied by solar PV modules, to a controlled DC voltage output at a voltage level required by the network loads [5-7]. They mostly perform the conversion by applying a DC voltage across an inductor for some period of time (in the range of 20 kHz to 5 MHz). That makes current to flow through inductor and store energy magnetically into capacitor. Then switches voltage off and makes the stored energy to be transferred to the voltage output in a controlled fashion [4-7]. The output voltage is adjusted by the ratio of on/off switching time. This is done by using fast switching power components like IGBT or MOSFET who disperse negligible power. This is called Pulse-width modulation (PWM). It allows control and adjustment of the total output voltage. It is discussed to be the core of the power supply, and it will affect the overall performance of the power supply system inflicting voltage. The ideal

converter, just in theory, exhibits 100% efficiency, in practice, efficiencies of 70% to 95% are typical.[8].

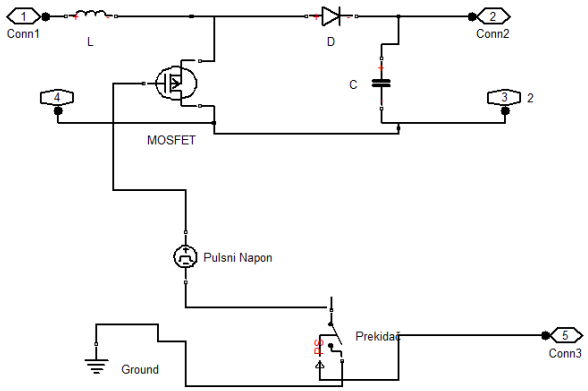


Figure 7. Scheme for the DC-DC boost converter

The DC-DC boost converter composition, as shown in Figure 7, is consisted of boost inductor L, controlled switch MOSFET, diode D and filter capacitor C. Figure 7 does not contain the PV module which comes between point 1 and 4 and the load from the network which comes between comes points 2 and 3.

DC voltage gain of the boost converter is given as:

$$M=V_0/V_S=1/(1-D) \quad (1)$$

Where V_s is the input voltage, which is in this case voltage from PV module, Vo is the output voltage, in this case that's voltage that goes to load, and D is the duty cycle of the pulse width modulation (PWM) signal used to control the ON and OFF states of MOSFET. The selection of different components of a converter is done as follows.

Large inductance tend to increase the start-up time slightly. Small inductance allows the coil current to build up to higher levels earlier than the switch turns off. Boost inductance is selected based on the maximum allowed undulation current at the minimum duty cycle D and at the maximum input voltage V_s. The DC-DC boost converter operates in the continuous conduction mode for the value of inductance L >L_b [12, 13].

$$L=\frac{\left(1-D^2\right)*D*R}{2*f_s} \quad (2)$$

Here, L_b is the critical inductance as the inductance at the border between continuous and discontinuous modes of the DC-DC converter. R is the equivalent load and f_s is the switching frequency of MOSFET.

The current that is supplied to the output RC circuit is discontinuous as result of switching [8]. Thus, a larger filter capacitor (C) is needed to prevent the output voltage from going above allowed undulation. The first and main criterion for selection of output filter capacitor is its capacity (magnetic energy storage) and equivalent series resistance (ESR). While the ESR of the capacitor acts on efficiency, low ESR capacitors is used for best performance of voltage output. The output filter capacitor is selected to meet a maximal output voltage undulation level. Maximum undulation on the output voltage can't exceed a few percentages of the output voltage [10].The minimum calculated value of the filter capacitor that provides the output current from DC-DC to the load when the diode is off is given in [10].

$$C_{min}=\frac{V_o*D}{f_s*\Delta V_o*R} \quad (3)$$

Where is ΔVo ripple on the output voltage which is selected in the volume of less than 5% of the output voltage from DC-DC converter.

Minimum values of the capacitor and inductor are calculated by equations (2) and (3). Actual values of the inductor and capacitor, used in the simulation, are kept higher than these values [7-8]. This makes big difference when simulating model. The values of installed component of the DC-DC boost converter used for simulation are given in Table 2.

Table 2. Adjusted values inside the DC-DC boost converter

PV modul		**DC-DC boost converter**	
Description	Value	Description	Value
Maximum Power (Pm)	10 W	Inductivity (L)	33μH
Maximum Power Voltage (Vmpp)	16.8 V	Capacity (C)	33μF
Maximum Power Current (Impp)	0.595 A	Load (R)	0-100 Ω
Short Circuit current (Isc)	0.67 A	Diode voltage	0.6 V
Open Circuit voltage (Voc)	21.0 V	MOSFET voltage	1 V
Temperature co-eff of Voc	-0.38/C%		

ANFIS MPPT based standalone PV system with one phase DC-DC boost converter, simulated in MATLAB/Simulink, is presented in Figure 8.

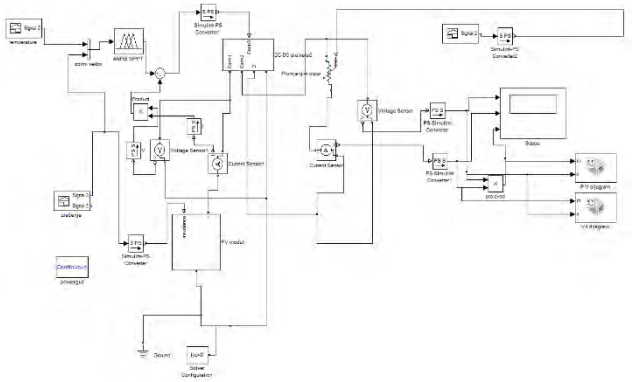


Figure 8. Modeled ANFIS MPPT based standalone PV system with one phase DC-DC boost converter

The output from the PV module is given as an input for the DC-DC boost converter which, by the navigation from ANFIS MPPT, adds or subtracts voltage on the output toward load. Output characteristics according to time changing load are collected in the period of 20s and presented in Figure 9. According to the graphs, ANFIS MPPT works toward increasing voltage because the points are grouped around the largest values of the voltage. It is necessary to review L, C and MOSFET to bring points on

graphs to the peak of the P-V graph and the knee of the I-V graph. Retraining of ANFIS is not necessary because the PV module has not changed.

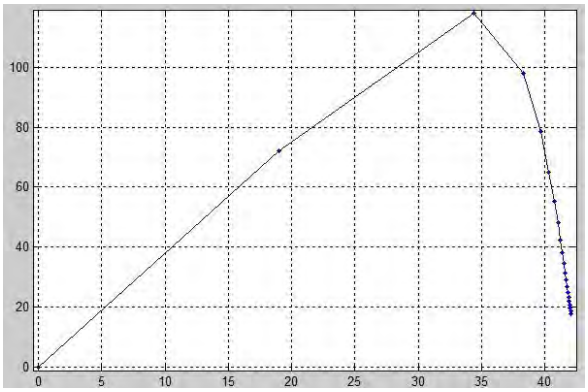


Figure 9. P-V graph after testing simulation

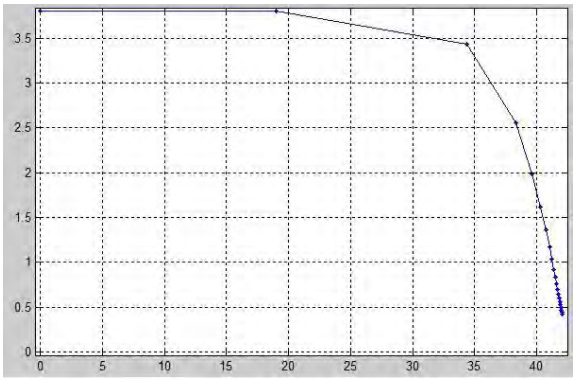


Figure 10. I-V graph after testing simulation

V. CONCLUSION

In this paper, the design and implementation of the ANFIS-based MPPT control scheme with an open loop boost converter is presented. The scheme of the DC-DC boost converter is presented and according to the calculated values, the components are parameterized. The simulation of the proposed control scheme is done using the MATLAB/Simulink software package. The simulation results show that the ANFIS-based MPPT control scheme is efficient to track the maximum available power from the PV module under varying weather conditions. The reaction of ANFIS MPPT control is fast with benefits on all radiation levels and working temperature. In addition to this simulation, the parameters of DC-DC components are to be reviewed in order to gain better results.

REFERENCES

[1] R. Kharb1, M. Ansari, S.. Shimi, "Design and Implementation of ANFIS based MPPT Scheme with Open Loop Boost Converter for Solar PV Module", IJAREEIE, Vol. 3, Issue 1, January 2014.

[2] S. M: Sulthan, „Fuzzy based Maximum Power Point Tracking Controller for Stand alone Photovoltaic system using MATLAB/Simulink", ResearchGate, ARTICLE, Vol. 10 No.55, May 2015.

[3] M. F. Ansari, R. K. Kharb, S. Luthra, S. L. Shimmi and S. Chatterji, "Analysis of barriers to implement solar power installations in India using interpretive structural modeling technique", Renewable and Sustainable Energy Reviews, Vol. 27, pp.163–174, 2013.

[4] V. Lingareddy, G. Ravichandra and P. K. Maddukrui, "Effective Strategy for MPPT in PV/Wind Hybrid Electric Power System Interconnected with Electrical Utility Grid," International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 3, Issue 7. July 2013.

[5] S. D. Saravana, "Modeling and Simulation of Incremental Conductance MPPT Algorithm for Photovoltaic Applications," International Journal of Scientific Engineering and Technology, ISSN: 2277-1581, Vol. No.2, Issue No.7, pp: 681-685. 1 July 2013.

[6] M. Lokanadham and K. V. Bhaskar, "Incremental Conductance Based Maximum Power Point Tracking (MPPT) for Photovoltaic System," International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622, Vol. 2, Issue 2, pp.1420-1424. March-April 2012.

[7] H. Can, "Model of a photovoltaic panel emulator in Matlab-Simulink," Turkish Journal of Electrical Engineering & Computer Sciences, March 2013.

[8] Jang, J.-S.R. (1993). "ANFIS: adaptive-network-based fuzzy inference system". IEEE Transactions on Systems, Man and Cybernetics 23 (3). doi:10.1109/21.256541, 1993.

[9] R. Singh and M. Pandit, "Controlling Output Voltage of Photovoltaic Cells using ANFIS and Interfacing it with Closed Loop Boost Converter," International Journal of Current Engineering and Technology. Vol. 3, No.2. June 2013.

[10] D. Radianto, D. A. Asfani, T. Hiyama and Syafaruddin, "Partial Shading Detection and MPPT Controller for Total Cross Tied Photovoltaic using ANFIS," ACEEE International Journal on Electrical and Power Engineering, Vol. 03, No. 02. May, 2012.

[11] A. Durgadevi and S. Arulselvi, "ANFIS Modeling and Experimental Study of Standalone Photovoltaic Battery Charging System," International Journal of Modern Engineering Research (IJMER), Vol.2, Issue.4, pp-2516-2520. July-August, 2012.

[12] R. R. Maziar, R. R. Mani, S. Akbari and S. A. Taher, "Using ANFIS, PSO, FCN in Cooperation with Fuzzy Controller for MPPT of Photovoltaic Arrays," Advance in Electrical Engineering Systems, Vol. 1, No. 1. March 2012.

[13] A. M. Zaki, S. I. Amer and M. Mostafa, "Maximum Power Point Tracking for PV System Using Advanced Neural Networks Technique," International Journal of Emerging Technology and Advanced Engineering, Vol. 2, Issue 12. December 2012.

[14] M. F. Ansari, B.C Sharma and P. Saini, "Maximum Power Point Tracking of A Solar PV Module using ANFIS," The 3rd IEEE International Conference on Sustainable Energy Technologies, Nepal, 24 – 27 September, 2012.

[15] P. Pradipkumar S, H. Ajaykumar V and Prof. B. N. Vaidya, "Design of DC/DC Boost Converter for the Application of Photovoltaic System," Global Research Analysis, Vol. 2, Issue 4, April 2013.

Linear Motion Calculation of the High Voltage Circuit Breaker Contacts Using Rotary Motion Measurement with Nonlinear Transfer Function

mr.sci. Kerim Obarčanin dipl.ing.el.
Software Engineering Dept.
DV Power
Stockholm, Sweden
kerim@dv-power.com

Radenko Ostojić dipl.ing.el.
Application Engineer
DV Power
Stockholm, Sweden
radenko@dv-power.com

***Abstract*—Motion measurement of the contact system of high voltage circuit breakers is of crucial importance for assessing the condition and health of the test object. To monitor motion of the contact system, depending on the construction, motion transducer is usually mounted on the moving parts of the operating mechanism linkage. That introduces a measurement error into the result. This paper presents the method and the software implementation of the measurement-error mathematical-correction.**

Keywords—Circuit breaker, Motion measurement, Motion analysis, Rotary motion

I. INTRODUCTION

Life expectancy of a newly installed high voltage circuit breaker (HVCB) is approximately 40 years. Throughout its life, under normal conditions, the circuit breaker will operate less than ten minutes total. During abnormal conditions, the breaker will operate less than one minute in its lifetime. [1]

The primary goal of diagnostic breaker testing is to determine the condition of the circuit breaker. Breakers are mechanical devices that have moving parts, components that allow main and arcing contacts to open or to close. If the motion of the contact system is in accordance with manufacturer specifications, than we can reasonably conclude that all mechanical parts, from the mechanism that provides the movement to the contacts themselves, are in good condition.

The accurate motion measurement itself is very important but parameters such as speed, acceleration, arcing contact overlapping distance, overtravel, rebound, and damping time, which are calculated from the motion curve, get more exact and accurate values.

Inaccurate motion measurement can affect conclusion regarding mechanical defects of the kinematic chain, overall mechanical performance, slow operation due to jammed mechanism, deterioration of mechanical damping, contact wear, arcing contact length etc. [2]

From the motion curve or its derivatives - a velocity or acceleration curve can be calculated in order to reveal even marginal changes that may have taken place in the breaker mechanics. [3]

Furthermore, for the dynamic resistance measurement, quality of the arcing contacts is determined correlating motion and voltage drop curves. If the motion measurement is not accurate, calculation of the overlapping time and overlapping distance would be inaccurate and the overall insight into the quality of the arcing contacts would be compromised. [4]

Most HVCB contacts have linear, straight line motion. However, depending on the construction of the HVCB, placement of the motion transducer on the contacts is usually not possible and the measurement of the contacts motion has to be performed on the lever between the contacts themselves and the operating mechanism.

In such cases good option is the rotary motion transducer connected at the point of the lever rotation (which transmits a movement from the mechanism to the main contacts) or at the main shaft in the mechanism.[5] Results obtained at that point, can be converted into a linear motion of the main contacts using a simple transducer scaling ratio (rotation/ distance).

However, for most circuit breaker mechanisms a constant transducer scaling ratio of rotation to distance does not accurately relate the measured rotation to the contact motion. Because of that incorrect calculations for motion parameters and velocity may result.

Alternative solution for contact motion calculation is using a nonlinear transfer function based on the trigonometric functions, since it causes considerably smaller error compared to linear conversion.

II. MATHEMATICAL APPROACH

Consider the mechanical scheme shown in the Figure 1.

In this example a rotation at the lever's point of rotation "A" drives a lever arm connected by a linkage to the contact drive shaft which drives the moving contact inside the circuit breaker chamber.

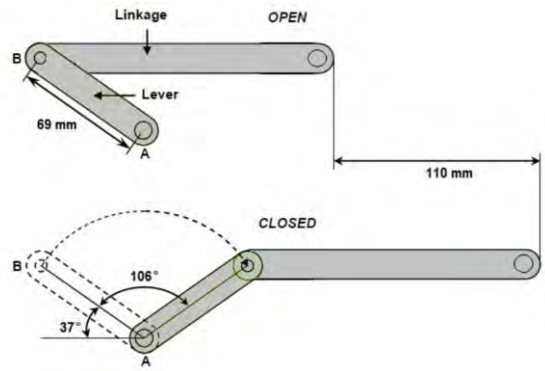


Figure 1. Rotary to linear mechanism [6]

As the lever rotates, the motion at point "B" has both horizontal and vertical components (Figure 2). Because the mechanism shaft (which drives the main contacts) is confined to horizontal travel by the shaft guide, it is only affected by the horizontal component of motion. The vertical component of motion at point "B" does not result in any motion of the contact drive shaft itself.

One approach to obtain contact motion curve would be relating the contact travel to the measured rotation using contact transducer scaling ratio of 110mm / 106° which can be simplified to 1.04 mm/°.This approach brings a deficiency assuming the relationship between the contact travel and the rotation of the lever point "A" is linear. In the reality this assumption is incorrect, as illustrated in the Figure 2.

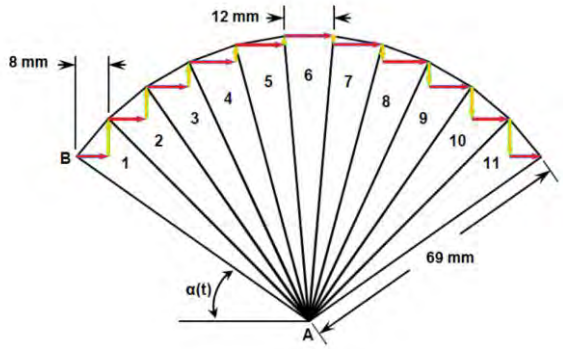


Figure 2. Motion of the mechanism's lever [6]

Relationship between the rotation of the lever and the contact motion is trigonometric function and it is described by the following equation:

$$L(\alpha) = R \cdot [\cos(\alpha_0) - \cos(\alpha_0 - \alpha)] \quad (1)$$

where is:

R – Lever length
α_0 – Initial lever angle
α – Rotation measured by rotary transducer
$L(\alpha)$ – Linkage displacement as a function of rotation

The $L(\alpha)$ obtained in the Equation 1. is the horizontal component of the linkage motion which represents the linear contacts motion.

Therefore a constant mm/degrees ratio cannot be used to accurately calculate the motion at the contacts at each angle during the circuit breakers stroke. For example, using nonlinear conversion, ratio contact motion/ transducer rotation will be 0.83 mm/deg (8 mm/ 9.6˚) at the end positions and 1.25 mm/degree (12 mm/ 9.6˚) at the middle position of the motion, while this ratio will be always 1.04 mm/deg using linear conversion. This means, some motion parameters calculated at the end positions of motion (Overtravel, Rebound, Contact wipe, Velocity etc.) will be higher than real values when use simple linear conversion.[6] This will be shown in the following examples.

III. IMPLEMENTATION

According to the mathematical approach described in the Section II, software solution for setting the parameters, performing the calculation algorithm and result presentation has been developed. The user interface for the settings adjustment is shown in Figure 3. The software allows user to input following parameters:

- Alpha total
- Alpha zero (in relation to the X axis, Figure 4)
- Stroke
- Reference position
- Axis of the linkage motion (Figure 4)

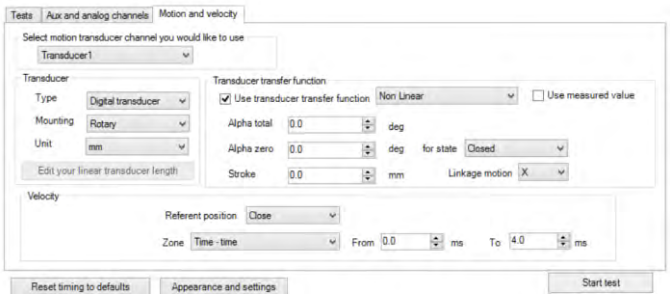


Figure 3. Software interface for the transfer function parameters input

To perform testing smoothly and faster, algorithm has been improved in the way that it takes for the Alpha total parameter value that is measured and calculated from the raw signal (Stroke parameter). This function can be switched on and off using "Use measured value" checkbox.

As mentioned above there are 5 different parameters that should be set (Figure 3.). The stroke parameter is a contact motion from fully open to fully closed position of the HVCB or vice versa. This parameter is provided by the manufacturer and can be found in the HVCB specification.

The Alpha total parameter is value of rotation measured at the place of the rotary transducer mounting and expressed in degrees during single HVCB operation. Linkage motion is the direction of movement of the linkage connected to the lever at the point of rotation where the rotary transducer is mounted.

The direction of movement along X or Y axes (Figure 4) should be set depending on the direction of linkage movement at the tested HVCB. Alpha zero parameter is initial angle of the lever in relation to X axis. In addition to this parameter, the circuit breaker state (open or closed, parameter Referent position) for which this angle is measured should be provided as an input parameter.

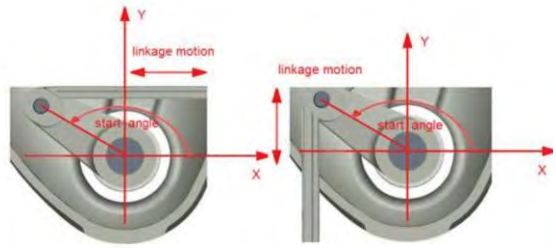


Figure 4. Mechanism linkage motion along X axis (left) and along Y axis (right)

IV. EXAMPLE I

For the purpose of mathematical approach verification (Section II) and the implementation (Section III) the experiment has been conducted on ten (10) HVCB in the field. Two chosen examples are presented in this paper.

In the first example motion measurements are performed using both, linear and rotary transducer simultaneously, and the results are compared. Also, comparison of the simple linear conversion (ratio) and the nonlinear trigonometry based conversion is performed.

The specifications of the HVCB used as a test object for the Experiment I are given in the Table 1.

The rotary transducer has been mounted on the center of the rotation of the lever between mechanism linkage and the main contacts shaft. The linear transducer has been mounted on the same lever at the point where the connected mechanism linkage is (point B, Figure 1). Transducers placement is shown in the Figure 5.

TABLE 1. The specification of the Energoinvest HG 6/8C

Manufacturer	Energoinvest
Type	HG 6/8C
Breaks per phase	1
Rated voltage	24 kV
Medium	Oil
Operating mechanism	Motor - spring

For this circuit breaker the manufacturer provided a relationship of 110 mm linear travel at the contacts corresponding to 106° measured rotation at location of the rotary transducer. In the Experiment 1, motion results from the rotary transducer are converted to linear contact motion using simple linear conversion and compared with the linear transducer measurement. Comparison between a contact motion measurement obtained with the linear transducer and converted motion measurement obtained from the rotary transducer using linear correlation is shown in Figure 6.

As can be noticed in the Figure 6, the motion curve obtained from the rotary transducer (full line) does not overlap with motion curve from the linear transducer (dashed line). These curves overlap only at the beginning, middle and the end of the contact motion. At the cursor position in the Figure 6. (where the value of motion measured with linear transducer is 85.8 mm), the measurement error with rotary transducer is more than 5 mm. As a result, the linear conversion of the rotary transducer reading will cause an error in motion parameters extracted from the motion curve such as average velocity, contact wipe, overtravel and rebound.



Figure 5. Digital rotary and analog linear transducer mounted and connected to the circuit breaker HG 6/8C

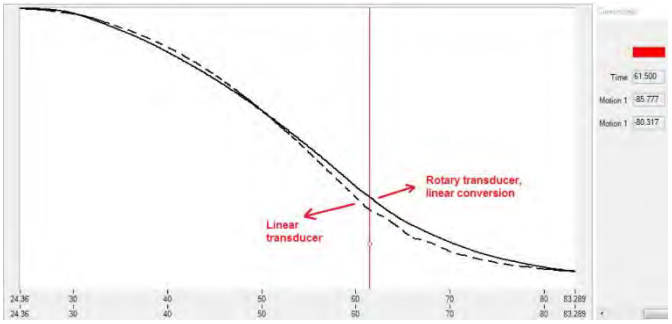


Figure 6. Comparison of motion curves obtained from linear transducer and rotary transducer with linear conversion of the measurement

Furthermore, simultaneous measurements were performed with linear transducer and rotary transducer using nonlinear conversion. Overlaid results are shown in the Figure 7. This comparison indicates that nonlinear conversion of the rotary transducer measurement causes considerably smaller error compared to linear conversion.

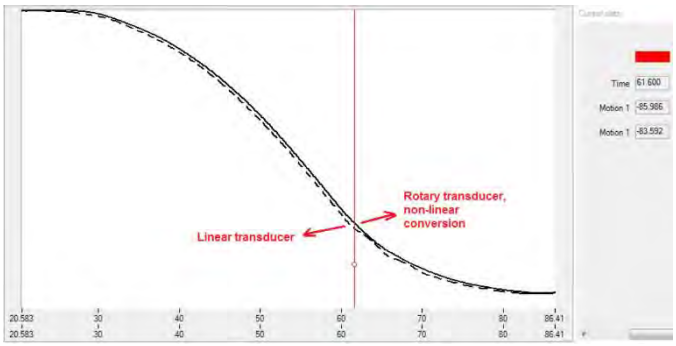


Figure 7. Comparison of motion curves obtained from linear transducer and rotary transducer with nonlinear conversion of the measurement

As can be seen in the Figure 7. motion curve obtained from the rotary transducer using nonlinear conversion almost overlaps with the referent curve obtained from linear transducer. Maximum error is only 2 mm (red cursor point) comparing the two motion curves at the same moment in time.

However, absolute error of 2 mm may be the consequence of the fact that the motion curves are obtained in two successive HVCB operations and it is not always likely that the contact motion will have the identical value at the same moment on two different operations.

Previous experiment results and deviations are shown in the Table 2.

TABLE 2. Comparison table of the measurement results for the Example 1

Experiment	Motion Ref. value [mm]	Error [mm]	Error [%]
1	85.8	5	5.29
2	85.8	2	2.33

V. EXAMPLE II

The Example II shows how Ovetravel parameter can be miscalculated using the simple linear conversion. We have shown the correction done using the non-linear trigonometric function. As a benchmark Ovetravel parameter, the data obtained with linear transducer at the previous testing of the circuit breaker is used.

Contact motion measurements are performed using digital rotary transducer which readings are automatically (by software) converted to contact motion using both linear conversion and non-linear trigonometric function. These converted results are then compared and analysis of the Ovetravel parameter is performed.

Testing was done on the SIEMENS 3AP1 FI circuit breaker equipped with the digital rotary transducer which was mounted at the lever in the bottom of circuit breaker pole (Figure 8). The HVCB specifications are given in the Table 3.

TABLE 3. The specification of the Siemens 3AP1 FI

Manufacturer	Siemens
Type	3AP1 FI
Breaks per phase	1
Rated voltage	245 kV
Medium	SF6 gas
Operating mechanism	Motor - spring



Figure 8. Digital rotary transducer mounted at the 3AP1 FI circuit breaker

According to manufacturer's specifications for this circuit breaker, a relationship of 150 mm linear travel at the contacts corresponding to 60° measured rotation at the lever where rotary transducer is mounted.

Contact motion measurement on the closing operation is analyzed. Comparison between the contact motion measurement obtained by conversion of digital rotary transducer reading using linear conversion and non-linear trigonometric function is shown in the Figure 9.

According to the linear conversion, motion of contacts is calculated using the measured rotation (degrees) multiplied by a simple, constant mm/deg transducer scaling ratio which is 2.5 mm/ deg (150 mm/ 60˚). This method of linear motion calculation from the rotary reading will result in the nonlinear error discussed in the section II.

As can be noticed in the Figure 9, the motion curve obtained with the linear conversion (dashed line) does not overlap with the motion obtained with the nonlinear trigonometric function (full line). As a result, the linear conversion of the rotary

transducer reading will cause an error in motion parameters extracted from the motion curve, especially at the end of the motion (Overtravel, Rebound etc.). For example, Overtravel parameter calculated based on the linear conversion is 21 mm, while this parameter based on the nonlinear conversion is 18 mm. Since reference value of Ovetravel parameter obtained using linear transducer is 17.5 mm, it is clear that error is smaller with application of nonlinear conversion (Table 4).

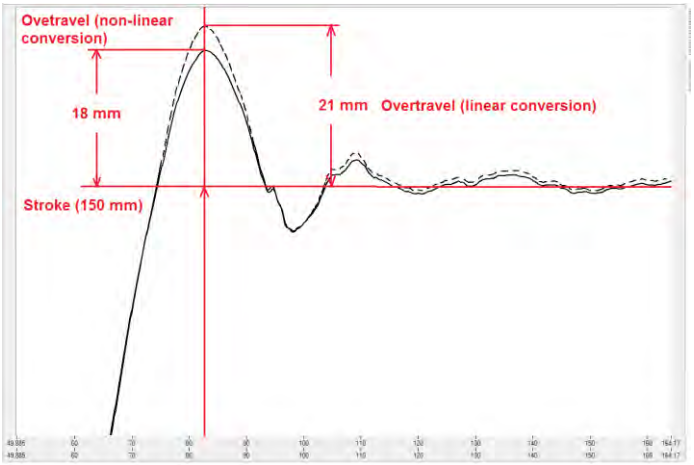


Figure 9. Comparison of motion curves obtained with rotary transducer measurement on the 3AP1 FI circuit breaker with application of linear conversion and nonlinear function

TABLE 4. Comparison table of the measurement results for the Example 2

Conversion	Overtravel Ref. value [mm]	Error [mm]	Error [%]
Linear	17.5	3.5	20
Nonlinear	17.5	0.5	2.8

These results confirm that parameters calculated in the end positions of motion curve based on linear conversion are higher than real values, i.e. higher than parameters calculated based on the nonlinear trigonometric function.

Although in this case, Overtravel parameters for the closing operation obtained with both conversions are according to manufacturer specifications (10-25 mm), problem can arise with use of linear conversion in the case of circuit breaker when real value of Overtravel parameter is near the upper limit (25 mm), but still within limits. In that case linear conversion will get Overtravel parameter higher than upper limit, which may falsely indicate a failure in the mechanism.

VI. EXAMPLE III

The Example III describes how Average velocity parameter during the opening operation can be miscalculated using the simple linear conversion and how the incorrect conclusion can be avoided using proposed non-linear trigonometric function.

Testing has been performed on the SIEMENS 3AP1 FG circuit breaker equipped with the digital rotary transducer which was mounted at the lever in the bottom of circuit breaker pole (similar like for example in the Figure 8). Specifications of this CB are given in the Table 4.

TABLE 4. The specification of the Siemens 3AP1 FG

Manufacturer	Siemens
Type	3AP1 FG
Breaks per phase	1
Rated voltage	145 kV
Medium	SF6 gas
Operating mechanism	Motor - spring

According to manufacturer's specification, a relationship of 110 mm linear travel at the contacts correspond to 60° measured rotation at the lever where rotary transducer is mounted.

Further, specification limits for opening time are from 26 ms to 34 ms and for the Average velocity from 4.2 m/s to 5.3 m/s (Table 5). Measured opening time for this circuit breaker is 33.2 ms, that is close to upper limit of operating time. This should mean the circuit breaker probably operates close to lower velocity limit. Since operating times are within limits, average velocity must be within limits as well. Using the linear conversion, calculated Average velocity is 4 m/s, which is below physically possible velocity (since it is out of specified limits). According to non-linear conversion, calculated Average velocity is 4.2 m/s. This result is meaningful since it is expected that ciruit breaker operates close to lower limit velocity. The comparison of obtained and calculated results is given in the Table 5.

TABLE 5. The specifications and results for the Example 3

Parameters		Limits	Results	Comparison
Opening time [ms]		26-34	33.2	Pass
Average velocity [m/s]	Linear	4.2 -5.3	4.0	Fail
	Non-linear		4.2	Pass

Using non-linear conversion we eliminated suspicion of wrong detection of the circuit breaker condition.

Average velocity is measured in the arcing zone during the interval of 10 ms, starting from the moment of contact separation (Figure 10).

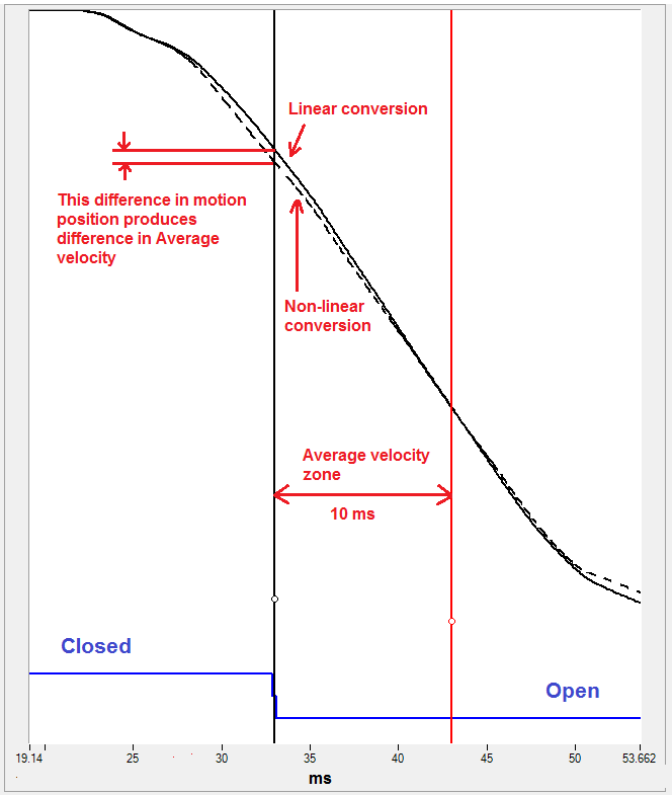


Figure 10. Calculation of Average velocity for the 3AP1 FG circuit breaker with application of linear conversion and non-linear function

VII. CONCLUSION

Since the measurement of the contacts motion often has to be performed on the linkage between the contacts themselves and the operating mechanism using rotary transducer, depending on the HVCB construction, results obtained using circuit breaker analyzer and timer devices are incorrect in the certain ranges of the motion path while applying simple linear conversion.

The conversion algorithm based on trigonometry functions and its implementation provides more accurate contacts motion measurement, which is specially emphasized in the Section IV and Section V.

Based on above considerations conclusion is that nonlinear conversion of the rotary transducer readings should be preferably used, since it causes considerably smaller error compared to linear conversion.

REFERENCES

[1] J.Levi "Timing and Motion Testing", Electricity Today, September/October 2014, Volume 27, No7

[2] T.Renaudin, "Circuit Breaker Testing – New Approach", Electricity 2013, Jerusalem, Israel

[3] "Technical Guide - Circuit Breaker Diagnosis", Programma Electric AB, Sweden,

[4] K.Obarcanin, A.Secic, N.Hadzimjelic, "Design and Development of the Software Solution for Analysis and Acquisition of the Hight Voltage Circuit Breakers Dynamic Resistance Measurement Results", MIPRO 2015, Opatija, Croatia

[5] J.R.Brown, "Circuit Breaker Linear Motion Measurement using Linear and Rotary Transducer", 79th Annual International Doble Client Conference, 2012

[6] "Use of the Transfer Function for Rotary to Linear Motion Conversion", Application Note, DV-Power, Stockholm, 2015-01 -22

Robot arm teleoperation via RGBD sensor palm tracking

Filip Marić, Ivan Jurin, Ivan Marković, Zoran Kalafatić, Ivan Petrović

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{filip.maric, ivan.jurin, ivan.markovic, zoran.kalafatic, ivan.petrovic}@fer.hr

***Abstract*—In this paper we present a simple and intuitive approach to teleoperating a 3 or higher degrees-of-freedom (DOF) robotic arm in Cartesian space. Using an RGBD camera, we retrieve the position of the user's palm. This position is then translated into the desired robotic arm position, which is then used as an input to a control loop. The entire system is implemented in the Robotic Operating System, enabling simple functionality transfer to any compatible robotic arm. The system was tested on the Kinova Jaco 6DOF robotic arm with the aim of using it for object manipulation. We use the inverse kinematics for calculating the joint rotation velocities required for following the Cartesian path of the human hand. The resulting joint velocities are then sent to the robotic arm control interface which then passes commands to the pertaining API. Results corroborate the validity of the proposed approach for robotic arm teleoperation, opening the possibility for many potential applications.**

***Index Terms*—Robotic arm, Palm detection, Teleoperation, Kinect**

I. Introduction

Robotic arms have been used for decades as means of manipulating objects in dangerous environments, performing arduous, repetitive tasks, and assisting the disabled. In cases where robotic arms are not autonomously performing these tasks, they are usually teleoperated by humans. More often than not, these methods of teleoperation offer a somewhat steep learning curve, requiring the user to be familiar with various modes of control and settings. Some newly developed methods [1]–[3] are more intuitive, but require the user to wear certain equipment which is not always practical and requires a sizeable amount of set up time. On another note, a common problem with machine learning systems like [1], [2] is that they require training for each new user since joint mapping does not scale.

Successful replication of human arm movement with robotic arms is barred by the complexity of the shoulder and finger joints. In practice most currently available robotic arms cannot match human dexterity. For this reason, direct mapping of human arm joint positions to a robotic arm is still not a viable option, but is a promising one for the future.

Such considerations have caused researchers to explore various modes of teleoperation that reduce the dimensionality of control required by the operator, giving some control up to various algorithms in exchange for a more practical system. Studies in human motor-control have suggested that a low dimensional representation is feasible at the kinematic level [4]. The work presented by [1], [2] follows this result, using electromyography signals corresponding to the activation of the shoulder muscles to control a robotic arm in 3D space. We find another interesting approach in [3], where a rig with a force feedback mechanism is used to map the joints of a human arm to an anthropomorphic robotic arm.

In this paper we propose a system which produces a simple and intuitive way of manipulating objects that, once set up, requires little to none prior knowledge of the system. The hardware used for detection is easily obtainable and low-cost. We maintain a focus on making the system generalizable to a variety of robotic arms, exposing all the important set up parameters through the Robot Operating System (ROS) parameter interface. All of the libraries and packages used in this project are well documented and maintained, hopefully making our code compatible with future versions of ROS. Unlike the majority of complex and slow algorithms used in hand-pose estimation and tracking problems, the main advantage of the proposed approach is its simplicity and speed which offers a large bandwidth for possible upgrades.

The paper is organized as follows. In Section II we cover the theoretical framework for such a system under the Robot Operating System (ROS) architecture. Section III contains an overview of the hand-tracking algorithm developed for the Microsoft Kinect RGBD camera. In Section IV we describe the proposed system for controlling the Kinova Jaco 6DOF robotic arm. The system was tested in a variety of situations, with an accent on prehensile and non-prehensile manipulation of everyday objects. It is our opinion that these results show potential for this kind of system in a number of different applications. In the end, Section V concludes the paper.

II. System Overview

Conceptually, our proposed system takes the form of a high level control loop (Fig. 1). In this analogy, the detected human hand and robotic arm positions serve as input and feedback data, respectively. Before subtraction, both the input and feedback data are placed in the same reference frame. The motion planner, serving as the regulator, subtracts the data and

This research has been partly supported by the Ministry of Science, Education and Sports of the Republic of Croatia under the grant "Centre of Research Excellence for Data Science and Cooperative Systems".

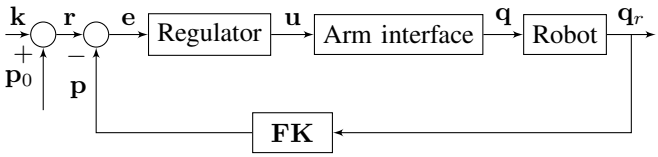


Fig. 1: Block diagram representing the system. Note that **FK** represents the forward kinematics of the robot. The control and interface blocks compose the motion planner.

converts the resulting difference into joint speeds of the robotic arm. In this section, we focus on the motion planner and the various components it utilizes to control the motion of the robotic arm.

A. Input data

The input data takes the form of a vector in $\mathbb{R}^6$ and a separate scalar in $\mathbb{R}$. The scalar c represents the amount of end-effector closure, while the values $x,\ y,\ z,\ \psi,\ \theta,\ \rho$ represent the positions and rotations in Cartesian space:

$$\mathbf{k} = \begin{bmatrix} x & y & z & \psi & \theta & \rho \end{bmatrix}^T; \quad c\ . \tag{1}$$

We denote the current pose of the end effector in the base coordinate frame as $\mathbf{p}$ and the end-effector closure as c_p:

$$\mathbf{p} = \begin{bmatrix} x & y & z & \psi & \theta & \rho \end{bmatrix}^T; \quad c_p\ . \tag{2}$$

To make the input data usable, we need to transfer it to a coordinate frame of the robotic arm. Another thing which we also have to account for is the difference in range between the human and robotic arms.

One possible solution for this would be to find a transfer matrix between the Kinect and robot arm base coordinate frames, directly mapping $\mathbf{k}$ and c to the desired pose of the end-effector. The downside to this method is that it depends on the position of the Kinect relative to the robotic arm. This makes it impractical to set up the system and makes it prone to outside disturbances. We circumvent this problem by making the system start only when the human hand is near a certain position in the Kinect coordinate frame. When the hand is first detected, its position it then matched with the starting position of the end effector $\mathbf{p}_0$. In this manner we create a new coordinate system at the center of the human hand. All movement in this coordinate system is interpreted as the displacement of the end-effector from the position $\mathbf{p}_0$. Now we can define the modified reference data $\mathbf{r}$:

$$\mathbf{r} = \mathbf{p}_0 + \mathbf{k}\ . \tag{3}$$

Now we can also define $\mathbf{e}$ as the difference vector between the current hand position and the starting position displaced by detection data:

$$\mathbf{e} = \mathbf{r} - \mathbf{p} = \mathbf{p}_0 + \mathbf{k} - \mathbf{p}\ . \tag{4}$$

Notice that the end-effector closure variable c is coordinate system independent, thus allowing us direct mapping:

$$\mathbf{c}_p = \mathbf{c}\ . \tag{5}$$

B. Regulator

The regulator uses the value $\mathbf{e}$ to control the robotic arm path. The path is realized and regulated using the Cartesian velocity interface of the robot. We have chosen this option as opposed to a position based approach since it avoids time-based parametrization and is inherently smooth in most cases. One of the advantages this gives is the ability to use PID regulation on the path itself, which determines the inertia with which the robotic arm follows the human hand.

If we interpret $\mathbf{e}$ as the unregulated desired velocity of the arm then we can define $\mathbf{u}$ as the regulated variant. The D and I functions represent the integral and derivative components, respectively [5]:

$$\mathbf{u}(k) = K_p\mathbf{e}(k) + K_i\mathrm{I}(\mathbf{e}(k), T) + K_d\mathrm{D}(\mathbf{e}(k), T), \tag{6}$$

where T represents the rate at which the system operates. It is practical to adjust T to values where the distance the robot hand can move in one interval is strictly lower than $\mathbf{u}$. This adjustment keeps the movement smooth since it is less likely that the end-effector will overshoot the target position.

In cases where a Cartesian velocity interface is not available, one can be constructed given that the robotic arm supports joint velocity commands. This requires an exposed inverse kinematics function for the particular robot which is used for transforming Cartesian positions to joint values. The inverse kinematics problem is not trivial and can be approached in a variety of ways. Given the inverse kinematics function IK and the values of current joint angles $\mathbf{q}$, we can calculate the required joint velocities:

$$\mathbf{u}_q = \mathrm{IK}(\mathbf{r}) - \mathbf{q}\ . \tag{7}$$

Another way to transform Cartesian into joint velocity would be using the inverse Jacobian method [6]:

$$\mathbf{u}_q = \mathbf{J}^{-1}\mathbf{u}\ . \tag{8}$$

This method applies to velocities, but instead we use $\mathbf{u}$ without changing any parameters. We use the fact that $\mathbf{u}$ is relatively small because of the high operating frequency, allowing for this kind of approximation. This method was tested in a simulated environment using *Gazebo*, a 3D interactive simulation system with an in-built `ros_control` interface.

The inverse kinematics of the Jaco robotic arm is efficiently calculated by the official API. Moreover, since the official API also provides Cartesian velocity commands, it was used for performing the tests described in the results section.

C. ROS implementation

ROS is an open-source, meta-operating system that handles hardware abstraction, low-level device control, implementation of commonly-used functionality, message-passing between processes, and package management. It also provides tools



Fig. 2: Graphical user interface for hand tracking

and libraries for obtaining, building, writing, and running code supporting distributed computing. At the file system level, the main organizational component of a ROS system is the package. A package may contain executables (nodes), libraries, data-sets and configuration files. In a robot control system, nodes process data and communicate with each other through the Computation Graph.

The Computation Graph is the peer-to-peer network of ROS processes that are processing data together. Communication is done by nodes subscribing to and publishing standardized data structures called messages by way of topics. Topics can be seen as a location for a certain type of message to be subscribed and published to. Possibly the greatest advantage ROS has on other similar projects is the community. Most of the packages are community maintained by a large and active user base, which makes the process of learning the system considerably easier.

The system is implemented in ROS Indigo-Igloo, compatible with Ubuntu 14.04 OS. It consists of 2 nodes: the `hand_detection` node interfaces with Kinect and runs the hand detection algorithms on the recieved data, the `arm_control` node generates the control data $\mathbf{u}$, which is then sent to the `ros_control` interface. The `hand_detection` node, described in more detail in Section III, publishes the processed Kinect data $\mathbf{u}$ in the form of a `geometry_msgs\Twist` ROS message[] to the `detection_data` topic.

The `detection_data` topic is subscribed to by the `arm_control` node, which runs the control algorithm every time a new message arrives to the topic. From this follows that we can control the refresh rate T by adjusting the publishing rate of the `hand_detection` node.

The `arm_control` node then sends arm and gripper commands to their respective controllers, spawned and managed by the `ros_control` interface. It is important to note that the gripper is also controlled using joint velocities, so a value c of 0 would correspond to a maximum velocity closing of the gripper. This is managed by configuring the controller joint angle and velocity limits in the configuration files.

III. Hand-tracking Algorithm

The problem of hand pose estimation is already solved with some degree of success in classic RGB images using complex algorithms and large training sets [7], [8], [9]. The main disadvantages of using that approach lie in its dependency on the light parameters of the image as well as its complexity. These approaches have problems with real-time detection which is a requirement for our system. As depth cameras became more and more available to general public, an opportunity is provided to use depth data as a more reliable and robust data source which works well in any provided light condition.

A. Hand localization

Regardless of the camera being used, the first step in precise detection of a hand pose is to determine its position in the image. Various methods have been proposed for solving this problem, but most of them use some kind of a hint where the hand might be in the image. Moreover, many of the methods require users to wear a special non-reflective bracelet or some sort of marker object that can help estimate the hand location. This can vastly speed up hand isolation and offer more time to conquer hand pose estimation.

Qian et al. in their work [10] initially assume that hand is located closest to camera in relation to other objects in scene. Moreover, it requires human to wear non-reflective bracelet and uses flood-filling algorithm to extract part of image that is closer than the bracelet. Shotton et al. propose a different initial method, which is used to detect human pose in depth image [11]. It uses depth invariant features and trained random forest for binary per-pixel classification whether it is part of hand or not. Thompson et al. [12] use this method to extract one of more hands from the image. The method is extremely efficient and fast, especially on GPU but within this work there are no possibilities to use cluster of 1000 cores to train random forest for a few days so we were trying to conquer this problem using a simpler approach with still satisfying results.

In our work the background subtraction method assumes that hand is the closest object to the camera (Fig. 2 shows the user interface of the hand tracking algorithm). Unlike [10] the bracelet is replaced with simple and fast depth filter at an empirically determined depth offset of 110 mm. Figure 3 shows successful hand isolation invariant to provided palm incline even when palm is horizontally placed or closed. This simple but efficient and above all fast filter leaves more processing time to be spent in the upcoming detection stages.

If the closest point of the palm is on the middle finger, the carpal part of the palm could be omitted and also if the closest point is on the wrist, a part of forearm can be included (Fig. 4). Palm location is considered as the center of a contour that separates the palm from the background. Depth coordinate of the palm location is calculated in the proximity of that point. Although some hand poses can lead to the center point being outside of the palm itself, in practice this scenario is relatively rare and will be disposed as invalid (depth cameras have issues with detecting that kind of poses in the first place).



Fig. 3: Successful palm isolation

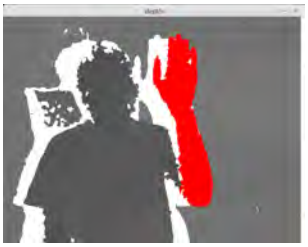


Fig. 4: Poor palm isolation

B. Hand pose estimation

Qian et al. have modeled an average hand using 48 spheres and 26 DOF [10]. Extreme points in extracted hand are candidates for tips of fingers and wrist. Using inverse kinematics they try to fit constructed model and best fit is considered as the current hand pose. Thompson et al. [12] use convolutional neural network cascade trained on set of 70 000 marked images. As well as [10] this paper also uses inverse kinematics to determine current hand pose from the neural network output. To model the hand this paper uses an open source library *libhand* [13]. It is notable to mention that both methods are able to recognize hand pose from single depth image only. Once the palm is extracted from the background, further processing is much simpler. Considering application of this tracker, precise hand pose estimation using many DOF is unnecessary. A focus has been made on palm openness as a binary indicator and some attempts to extract palm orientation have been made.

Hand openness is relatively abstract concept. We define a closed palm as a pose where every interphalangeal joint is flexed to a high degree possible. Everything else is considered an open palm. This concept emerges from the application of the tracker. As robot grasping mechanism differs from robot to robot, a direct correlation between per finger flexion and extension is not possible so the classifier determines whether robot will try to grasp something or not. This ensures the controlling does not depend on grasping type.



Fig. 5: Palm contour and convex hull

Binary classifier is a trained function of only two partly correlated features, depth invariant palm area and ratio of palm area and area of its convex hull (Fig. 5). Whilst constructing the classifier special attention has been made to prefer quality of classifying palm as closed at the expense of quality of open palm classification. This is also application specific improvement that assures once the robot has grasped something it will not release it unintentionally.

Training has been made on a database of 2600 palm images with the same amount of opened and closed examples. Using information from sequential images, a major improvement has been made by canceling noise from the depth camera. Empirically the best noise cancellation has been made using 7 sequential images and applying a mean filter on the detection outputs.

C. Results

Using an unoptimized Python script and average personal computer this method can generate up to 125 images per second (and even more if optimized and transfered to C++) and leaves space for possible upgrades. Determining the quality of detection has been accomplished in two ways. As the detector uses a single image for detection and sequential images for noise cancellation, a test has been developed for both single and sequential images. In both cases, the test consisted of classifying the palm correctly as opened and classifying it as closed.

Test has been made on 3108 sequential images of which 1323 with open and 1785 with closed palm. Unlike the learning process which was developed using a single persons palm in the testing process multiple persons have been asked to provide data. As expected, sequential images have proven to be significantly bettera with over 96% of success rate (Tables I, II, III, and IV). Closed palm detection is better than open palm detection which was the goal. The reason why the closed palm detection should be especially taken care of is the process of grasping which should be as reliable as possible. For a grasp to fail, using 30 fps data feed and having in mind robot speed, approximately 10 sequential detection fails are required. This makes it unlikely for such an event to occur.

	Hit	Miss	Sum
Example count	1203	120	1323
Percentage	90.9%	9.1%	

TABLE I: Classification of open palm state in single image

	Hit	Miss	Sum
Example count	1728	57	1785
Percentage	96.8%	3.2%	

TABLE II: Classification of closed palm state in single image

	Hit	Miss	Sum
Example count	1256	67	1323
Percentage	94.9%	5.1%	

TABLE III: Classification of open palm in sequential images

	Hit	Miss	Sum
Example count	1740	45	1785
Percentage	97.5%	2.5%	

TABLE IV: Classification of closed palm in sequential images

IV. Experimental Results

The system was developed and tested using the 6 DOF Kinova Jaco robotic arm. The official arm API offers both velocity and position commands in cartesian and joint space. This allowed us to construct a cartesian velocity interface for the robot which can handle position and orientation inputs.

A. Practical Considerations

It is useful to limit the tracking space to a sphere about 1.5 meters wide, centered at a point 1.5 meters away from the lens. This way the system is only first triggered by the opperator stepping into the sphere. Such a set up allowed for maximum movement space of the operators hand and simple cessation of operation. Leaving the sphere causes the hand to return to its starting position.

At the time of writing, an effective way of detecting human hand orientation using the Kinect is still to be developed, so the orientation of the arm remains constant in our testing. To implement a hand orientation detecting algorithm, one only has to publish `geometry_msgs\Twist` messages to the topic subscribed to by the control node. Swapping detection algorithms is made easy by use of the roslaunch system.

B. Simulation

The system proposed in Section II (Fig. 1) was simulated using the *Gazebo* robot simulation environment. As mentioned earlier, *Gazebo* allows for controlling the robot model using the `ros_control` interface. This enables the developed control algorithms to be easily transferable to a real robot, provided it also has an `ros_control` interface.

The inverse velocity kinematics for the robotic arm are calculated using *The Kinematics and Dynamics Library* (KDL) [14]. KDL uses the same robot description file as *Gazebo* to construct a kinematic chain describing the robotic arm it is then possible to solve forward kinematics, construct a Jacobian matrix and calculate inverse kinematics using iterative methods. System precision was tested by defining a

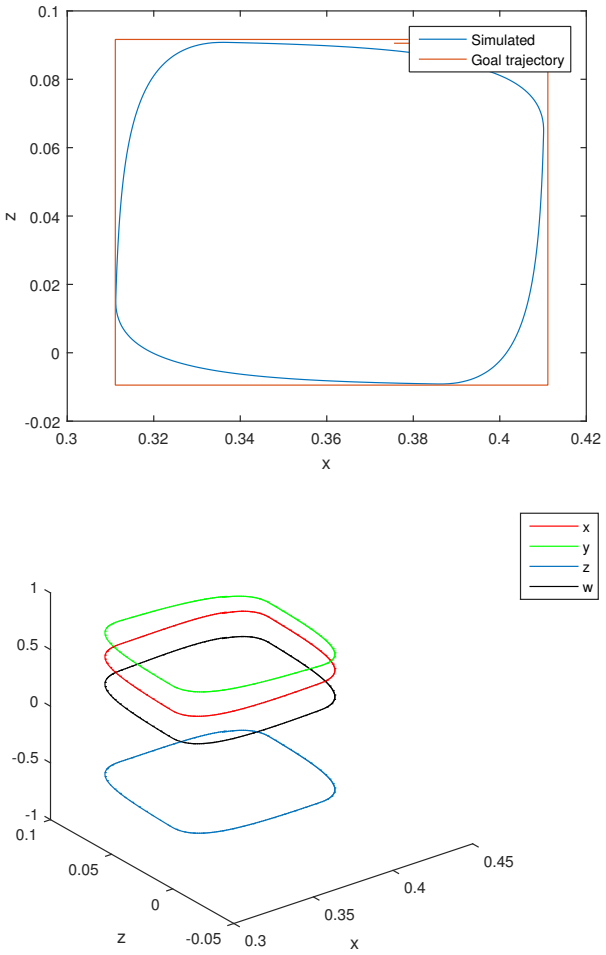


Fig. 6: **Top**: Desired Cartesian trajectory and the trajectory performed by the simulated Jaco arm. **Bottom**: Orientation of the arm during trajectory execution. In both the upper and lower images we see that the orientation is near constant.

square trajectory for the arm in the $x{-}z$ plane with constant end-effector orientation. The distance between goal trajectory points is around 0.001 cm, which gives a presumed arm speed of around 3 cm/s given the Kinects refresh rate. The results show a relatively low margin of error even with the most basic form of Jacobian-based inverse velocity kinematics (Fig. 7).

C. Testing

Testing the system consisted of picking up various objects from the table and placing them in a designated area. In the first scenario, a bottle was manipulated using the arm and consecutively moved from one side of the table to the other. The second scenario consisted of placing a series of objects inside of a plastic box.

We found that a using only the proportional term of (6) gives satisfying results, following the human hand effectively in 3D-space. This simplification gives a small absolute error when the hand is static, but error build up is avoided since the

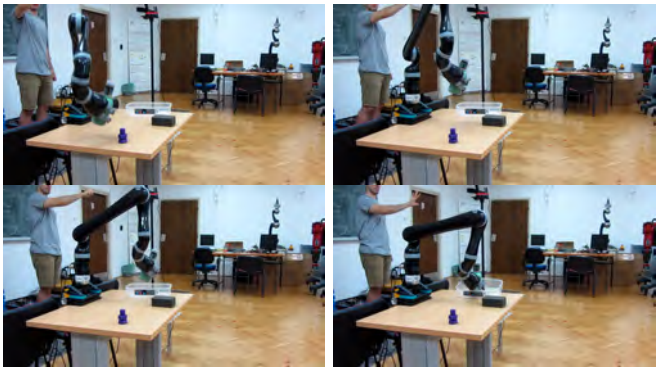


Fig. 7: Successful grasp for the first scenario

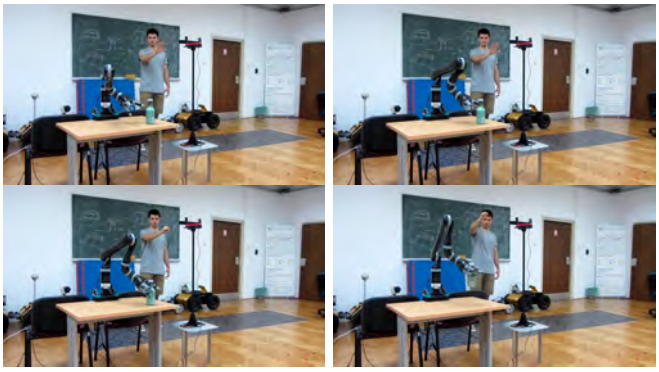


Fig. 8: Successful grasp for the second scenario

system constructs each new command using the initial position of the robot.

The first scenario, shown in Fig. 7 resulted in a successful grasp every time. We noticed that the system is suitable for grasping bottle-like objects, since the hand tracking algorithm has better hand detection capabilities when the hand is oriented towards the camera. It was also found that it is useful to limit the operation space of the manipulator in a way such that it corresponds to the free space in front of the user in order to maintain controllability. When the users hand leaves the operation space, the manipulator will return to it's initial position as defined when the system was started.

In the second scenario, shown in Fig. 8, we attempted sorting various objects into a plastic box. The lack of end-effector orientation control made it somewhat difficult to grasp smaller objects like screws, but grasping and placing non-rigid objects like sponges proved to be successful. The high refresh rate makes the system immune to noise when detecting hand closure, resulting in a very low percentage of failed grasps.

V. Conclusion

Visual teleoperation is proven to be a useful tool, but to make it usable outside of a laboratory setting, it needs to be made simple and affordable. In our research we have shown that this is possible by using relatively low-priced detection hardware and open-source software. The open-source ROS infrastructure enabled us to develop high abstraction algorithms, making the system reusable and adaptable to a variety of hardware. The hand detection algorithm was developed using the Microsoft Kinect platform with a focus on performance rather than precision, since the control loop design of the system enables efficient error correction. Finally, the tests performed on the system showed great promise, as the intuitive control mechanism made it easy to manipulate objects without much prior knowledge of the system.

A lot of improvement can still be made going towards a ubiquitous visual teleoperation system. Improving hand detection on low-priced hardware is inevitable, since it is important to correctly capture the users commands. With more advanced tracking systems for human joint positions, a general purpose algorithm transforming human to robotic arm joint positions could be developed, thus improving movement mimicking capabilities.

References

[1] P. K. Artemiadis and K. J. Kyriakopoulos, "Emg-based control of a robot arm using low-dimensional embeddings," *Robotics, IEEE Transactions on*, vol. 26, no. 2, pp. 393–398, 2010.

[2] J. Leitner, M. Luciw, A. Förster, and J. Schmidhuber, "Teleoperation of a 7 dof humanoid robot arm using human arm accelerations and emg signals," in *12th International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS), Montreal, Canada*, 2014.

[3] G. S. Gupta, S. C. Mukhopadhyay, C. H. Messom, and S. N. Demidenko, "Master–slave control of a teleoperated anthropomorphic robotic arm with gripping force sensing," *Instrumentation and Measurement, IEEE Transactions on*, vol. 55, no. 6, pp. 2136–2145, 2006.

[4] B. Lim, S. Ra, and F. C. Park, "Movement primitives, principal component analysis, and the efficient generation of natural motions," in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*. IEEE, 2005, pp. 4630–4635.

[5] B. C. Kuo, *Automatic control systems*. Prentice Hall PTR, 1981.

[6] S. R. Buss, "Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods," *IEEE Journal of Robotics and Automation*, vol. 17, no. 1-19, p. 16, 2004.

[7] M. de La Gorce, D. J. Fleet, and N. Paragios, "Model-based 3d hand pose estimation from monocular video," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 9, pp. 1793–1805, 2011.

[8] B. Stenger, P. R. Mendonça, and R. Cipolla, "Model-based 3d tracking of an articulated hand," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2. IEEE, 2001, pp. II–310.

[9] J. M. Rehg and T. Kanade, "Visual tracking of high dof articulated structures: an application to human hand tracking," in *Computer VisionECCV'94*. Springer, 1994, pp. 35–46.

[10] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and robust hand tracking from depth," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 1106–1113.

[11] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[12] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 5, p. 169, 2014.

[13] M. Šarić, "Libhand: A library for hand articulation," http://www.libhand.org/, 2011, version 0.9.

[14] R. Smits, "Kdl: Kinematics and dynamics library (2001)," 2013.

A Proposal for a Fully Distributed Flight Control System Design

M. Šegvić, K. Krajček Nikolić and E. Ivanjko
University of Zagreb/Faculty of Transport and Traffic Sciences, Zagreb, Croatia
miroslav.segvic@computer2cockpit.com, karolina.krajcek@fpz.hr, edouard.ivanjko@fpz.hr

Abstract - Since the delivery of the first A320 airliner with a Fly-by-Wire Flight Control System (FCS) in 1988, aircraft avionics architecture evolved significantly. Federated Architecture applied in the A320 family of aircraft presumed one computer per function. Limits regarding weight and space availability were reached and new generation of aircraft designed in early 2000s were equipped with Distributed Integrated Modular Avionics Architecture consisting of shared hardware resources running separate software modules according to aircraft priorities. Flight Control Computer functions were assumed by the Flight Control Module. While reducing the cost, weight and number of computers on board the aircraft, problems with troubleshooting and system modifications emerged. Proving that critical Systems perform within certain certification safety requirements became infeasible due to unpredictable dependencies between software modules. Abovementioned problems are addressed in this paper with a proposal of a Fully Distributed Flight Control System (FDFCS) design. Main contribution is that aircraft stability and trajectory control logic is distributed to a network of independent Control Units (CU) collocated on actuators collaborating to control the aircraft with respect to common goal. This paper outlines design for FDFCS and its CUs. Problems that distributed FCS implies and solves are identified. Finally, requirements for planned FDFCS Hardware in the Loop Simulator are set.

I. INTRODUCTION

The aircraft avionics architecture is mainly composed of processing modules and communication buses. It supports flight control and management, navigation, communication, weather systems, collision-avoidance systems, and aircraft health monitoring systems. Avionic systems represent about 40% [1] of aircraft costs for civil aircrafts, and more than 50% [2] in military aircraft. It is evident that an avionics architecture is central part of the modern aircraft.

Flight control system (FCS) consists of flight control surfaces, cockpit controls and connecting linkages. Fly-by-wire (FBW) Flight Control System replaces mechanical linkages with transducers, wires and actuators. A reliable communication network provides the backbone of every FBW system. Electrical components comprising the FBW system are integral part of the avionics architecture. FCS performs critical applications as flight stability augmentation, flight guidance and envelope protection. Therefore, the FCS system must satisfy safety and dependability requirements, and meet performance specifications for certification. Possible implication of FCS failure are severe.

Real-time performance of FCS control software and communication network must be assured. Aircraft FCS must deliver the right commands to the control surfaces at correct time in coordinated fashion. Even in the event of failure of one part of the system, the rest of the system must continue to work within certain specification called degraded mode. Communication network latencies that are defined by wiring and network infrastructure layout have to be accounted for during system design to assure real-time performance. During normal aircraft operation, aircraft equipment may fail. Dependability of FCS is dictated by the avionics architecture. Unpredicted events such as loss of a part of the communication system or failure of any of the flight control computers must not have catastrophic impact on the performance of the system. Performance during such events must be validated for system certification.

Safety analysis must be performed in accordance with Advisory Curricular AC25.1309, for System Design and Analysis [3]. The part 25 airworthiness standards of the FAA Federal Aviation Regulations are based on the fail safe concept, which considers the effects of failures and combinations of failures, both detected and latent, in defining a safe design. Requirement of AC25.1309 is that airplane systems must be designed so that the occurrence of any failure condition, which would prevent continued safe flight and landing of the airplane is extremely improbable [3].

Special care while designing the software for the FCS computers should be taken, to assure that similar software is not used on different computers. Using dissimilar software minimizes the risk of common error that could cause all computers to underperform or fail at certain unexpected condition. The concept is referred to as “Similar Redundancy” and it is also applied to hardware design of FCS computers.

Dependability, Safety and Real-time requirements define main constraints while designing the FCS. The designer must also consider other parameters as ease of maintenance, troubleshooting, upgrades and modifications, and account for future growth in required system functionality [4]. Focus of this paper is to provide concept of Fully Distributed FCS and to set requirements for the system simulator. The first stage of research and requirements to build the simulator to test the proposed

control concept are described. The concept block scheme of the needed simulator will be described in more detail.

II. AVIONICS ARCHITECTURES AND FLIGHT CONTROL SYSTEM

We can separate two different approaches for the design of the avionics architecture. Federated avionics architecture (Fig. 1) was used on commercial aircraft until 1990's. It is characterized by the one function - one computer concept. Each aircraft system is run on separate dedicated hardware resources connected to its sensors and actuators. Major advantage of such an approach is that dependencies between systems are easily identified. Troubleshooting of federated systems is simplified. Real-time behaviour for FCS can be proven and time delays can be measured. Certification requirements for the federated system are easily tested.

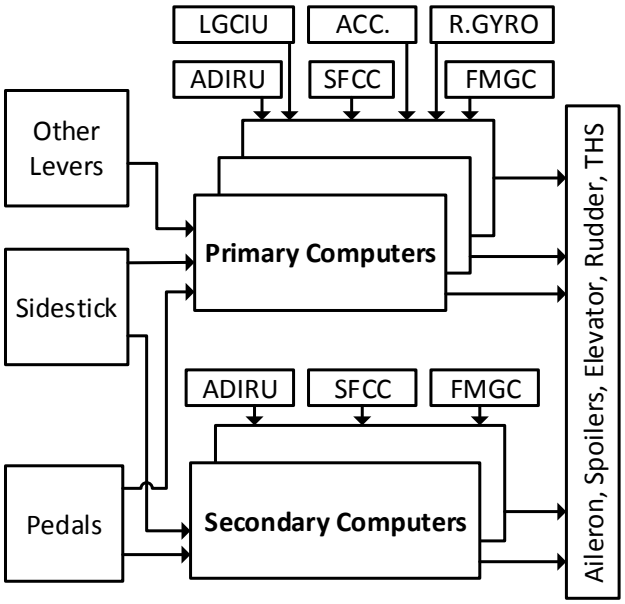


Figure 1. A320 Flight Control System (partially displayed), Federated Architecture

Major disadvantage of a federated system is a large number of different line replaceable units (LRU) comprising the system. From the operator standpoint, this causes difficulties since large number of components need to be stored on stock for replacement purposes. From the manufacturer standpoint, each new subsystem adds weight to the aircraft, reducing passengers and cargo capacity.

With the growing needs for avionics computing resources, the Federated Architecture reached its weight and complexity limits, and distributed integrated modular avionics architecture (IMA) was developed (given in Fig. 2). The IMA concept is based on sharing of hardware resources called IMA processors that are running separate software modules for each system. One IMA processor can run several modules with different priorities. Architecture development is accompanied by introduction of high-speed common communication channels. Common input-output interfaces are introduced for sensors and actuators.

Major advantage of the IMA concept is the reduced number of computers and communication data cabling on-board the aircraft [5]. From the manufacturer perspective, this advantage allows for reduced aircraft weight and required engine power. Operators need to store less LRU's for repairs. However, many disadvantages arise from sharing resources. Determining dependencies between systems running on common hardware resources and common data channels is difficult. Troubleshooting is also proving to be complicated. Testing real-time behaviour and measuring time-delays for FCS is almost infeasible. All said, demonstrating fulfilment of certification requirements for the IMA system is a major issue.

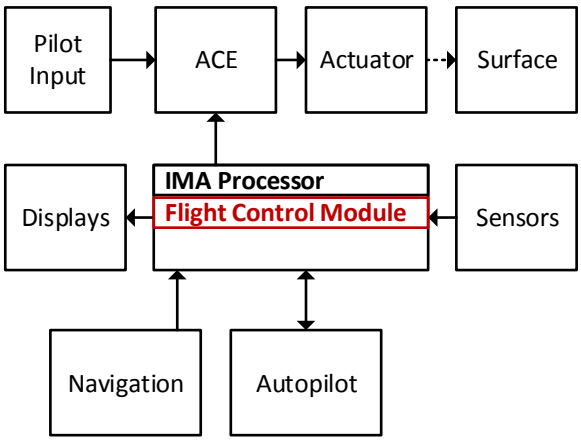


Figure 2. Distributed IMA (full lines represent network channels, dashed line presents mechanical linkage).

FCS, within the Federated architecture, is controlled by five separate computers [6]. Each computer has a control and monitor channel. Upon disagreement between channels, the affected computer is passivated while the next highest priority computer takes control. It is possible to fly the aircraft with only one functioning computer.

FCS within IMA is a software module. Upon failure of one IMA processor, the other one or a backup IMA processor loads the necessary software module and takes over the functions of the system. Meanwhile the actuator control system (ACE) provides a direct link from pilot controls to control surfaces [7].

III. MOTIVATION FOR A FULLY DISTRIBUTED FLIGHT CONTROL SYSTEM

A proposal for fully distributed flight control system (FDFCS) design is given in this article. The authors believe that FCS functions can be distributed down to the actuator control level. FDFCS would be a network comprised of actuator collocated control units, collaborating to control the aircraft with respect to a common goal. Major advantage of this approach is moving the requirements for certification of FCS from IMA to separate subsystem. Although increased number of embedded devices is required on-board the aircraft, the overall number of high power, high price computers can be reduced. The embedded devices for FDFCS should be collocated or installed near the control surface actuators, and have relatively low processing power thus low unit price. Furthermore, control units (CUs) can be of same type for all positions thus reducing the number of

replacement units necessary for operator to store. The reduction in data and control wiring by distributing the FCS functions to actuator level is significant. Integrating micro electro-mechanical sensors (MEMS) into each CU can reduce communication loads and the amount of data wiring for sensors. Redundancy of the overall system is exponentially higher. The proposed system is described in detail in the following chapter.

IV. FULLY DISTRIBUTED FLIGHT CONTROL SYSTEM

A fully distributed flight control system is defined as one where all the FCS roles and functions are distributed to the network of embedded CUs located on, or near the control surface actuators. Control units have to be networked in a secure and reliable way for the FDFCS design to operate safely. The choice of the connection standard is not proposed for the system; however, the controller area network (CAN) will be used as an example to demonstrate how safety and certification standards can be assured. Two separate CAN networks are assumed for redundancy. Terminating the two networks at different parts of aircraft assures that no part of the system is left unconnected for the case when the communication lines break at one point as shown in Fig. 3

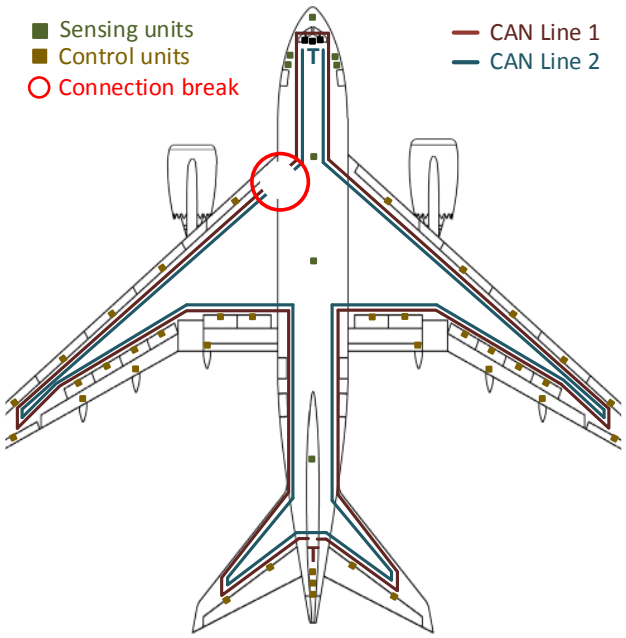


Figure 3. FDFCS CAN network routing with separate termination points for case of connection break near left wing root

Additionally provisions within the CUs have to be made to make actions in case of connection loss to ensure minimal intermission of the disconnected unit to the operation of the rest of the system and the controllability of the aircraft. This can be achieved by implementing automatic passivation of the affected control surface in a neutral position, on total connection loss.

The proposed design of the CUs consists of three embedded systems integrated into a single case called simply units. The term units will be used in this paper to avoid confusion with the term module used to describe software modules run on IMA. The primary embedded unit within the CU performs FCS functions and roles, and will be referenced from here as flight control unit (FCU). The secondary embedded system, referenced as external override unit (EOU), has the sole purpose of overriding the FCU outputs on a certain event, and allowing the remote control of the corresponding actuator.

The power regulation and communication level translators are doubled and not shared amongst the units, removing any chance for communication loss on both devices within the CU caused by translator or rectifier failure. The third embedded system is the actuator control unit (ACU) or the executive unit. The role of this unit is to manage actuator(s) connected in a way ordered by FCU and when overridden, the EOU. The unit uses both power provided by FCU and EOU as a redundancy to assure that when at least one unit is operating, the ACU has power available. Figure 4 shows the proposed design for the CU.

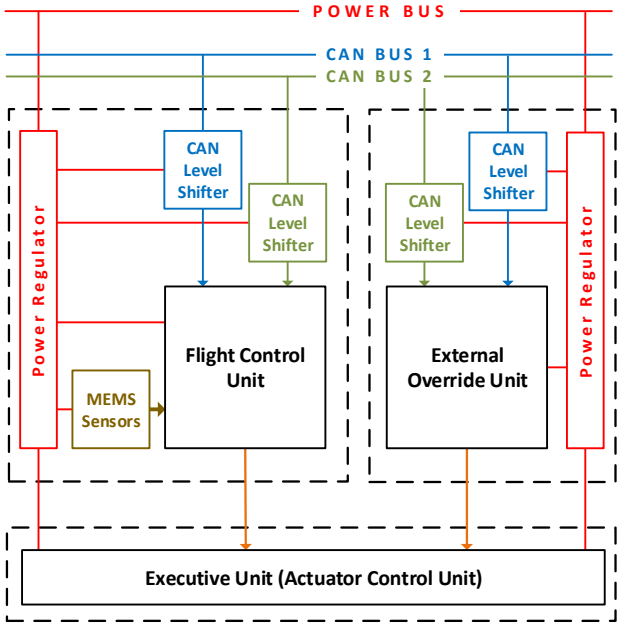


Figure 4. Proposed CU design for FDFCS

MEMS sensors are not yet precise enough for aviation purposes [8] to serve as the only input of positioning data instead of the sophisticated and expensive inertial reference units (IRU). However, it can be expected that they will reach the required specifications in future [9]. Considering the low price of MEMS sensors, it is reasonable to propose the integration of MEMS sensors within each CU. Low cost and low precision MEMS sensors within every CU can be used to estimate positioning data for short periods of time. Higher precision IRU should be used to correct MEMS sensors positioning on regular intervals. This approach helps to reduce the traffic on the communication network that would be caused by constant positioning data transfer from the IRU, GPS receivers and other sensory units to the CUs. In addition, the number of IRUs on board can be reduced once the precision of MEMS sensors rises to the required level. Theoretically, once the satisfactory precision can be maintained for the time aircraft requires to complete the precision approach, the aircraft should be fully capable to

continue the approach to the airport (runway) in case of IRU failure at the most critical moment, or at the beginning of the approach to the airport (runway).

Control system should not be solely time triggered or event triggered [10]. It would be beneficial that sensory units as GPS receivers and IRUs broadcast data on the network on regular intervals. That would assure that all the units have the positioning data corrected at certain regular interval. CUs should communicate between each other on a specific event, only when communication is required to perform FCS functions. However, provisions in each node have to be made to protect the buses by limiting the data bandwidth consumption [11].

All units need to transmit two kinds of data. The first kind would be data request, and the second data send. Units should be able to request and send data from and to other units such as control surface position or positioning data. Such a request allows that the monitoring function of one unit can be assigned to any other unit on the system, facilitating fault detection. When a certain number of units on the system detect a malfunction in operation of the monitored unit, the EOU should be activated and take control of the control surface.

In the emergency event of loss of many systems necessary for the normal or the automated operation of FCS, the degraded mode of operation should be available. The degraded control mode must allow direct control of the units. Direct control should transfer pilot commands to the control surfaces without any interference. Under no circumstances like other systems failure or corruption, should direct mode of operation be affected. These dependencies have to be designed in the system and validated.

The normal control mode should improve the stability of the aircraft and protect the flight envelope, independent of weather the aircraft is operated by a pilot or guided by the flight management system (FMS) and controlled by the autopilot. When normal control mode is active, CUs must cooperate to control the aircraft. For instance, the port and starboard ailerons and spoilers should differentially deflect in a way to prevent unwanted yaw. Communication is required for whichever motion the surfaces are coupled and produce total effect. To assure coordinated outputs, cooperation between units should be organized. Massive voting can be applied to achieve the desired result. However, for the system to be fully distributed there should be no central unit assigned to decide on the control surface positions. The network and its units should be self-sufficient to provide for all roles of FCS.

V. Requirements for the FDFCS HIL simulator

In order to validate the concept of the proposed FDFCS system a simulation has to be done. To improve the quality of the simulation a real time hardware in the loop simulation (HIL) is always recommendable. In order to make such a simulation certain requirements have to be fulfilled. An accurate aircraft and flight model including atmosphere and aircraft engines have to be simulated in real-time using appropriate simulation software like Matlab/Simulink, and the communication network with the CUs has to be implemented as a real system is this case. In such a simulation setup the control hardware will receive accurate inputs and computed control outputs will be forwarded to corresponding actuators and act as a feedback to the simulated aircraft, as shown in Fig. 4

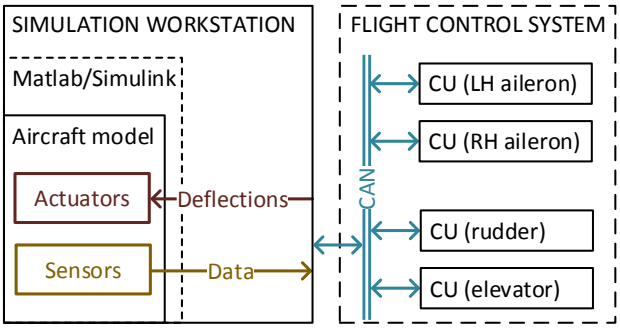


Figure 4. HIL simulator for the FDFCS showing the simulation workstation and hardware side of the system embedding the FCS

Envisaged HIL simulator design will use the CAN network to connect the needed components. A simulation workstation equipped with an aircraft model running in Matlab/Simulink will be used to simulate the aircraft in flight ensuring realistic sensory data. The simulation workstation will be equipped with a CAN interface to establish communication with several CUs. Each CU will be assigned to control one aircraft control surface.

VI. Conclusion

The embedded CUs will be designed so they will be able to run the FCS. CUs will be running control laws necessary to control the simulated aircraft control surface actuators. Needed aircraft data sensors like air data unit, inertial reference unit, and GPS receiver will be emulated in Matlab/Simulink to ensure needed realistic sensor measurements. The dynamics of the actuators controlled by CUs, will also be emulated in Matlab/Simulink as part of the aircraft model. Finally, the equal processing power centralised control FCS will be developed alongside, to serve as a reference for comparison of the centralised and fully distributed system. Another important requirement is detailed logging and analysis of data traffic in order to create procedures to enable certification tests of the proposed architecture. Additionally, these data logs will be used for in-depth performance analysis of the distributed control architecture. Conclusion

Distributing the FCS logic to actuators removes the need for any sophisticated computers performing FCS function in the avionics bay. Routing of sensory and data wiring is unnecessary, resulting in reduction of aircraft weight. Designing the units to be universal by modifying their control purpose depending on the installation functional item number, allows for storage of minimum number of LRU's for system repairs. However, the issue of unwanted similarity arises and needs to be addressed within the software. The complexity of the system reduces significantly especially because of integrating MEMS sensors into the CUs. Real-time behaviour is defined by the network but should be analysed for the proposed concept of control that mandates voting and data sharing.

However, fitting CUs directly on or near the actuators surely reduces latencies caused by usual over the network control of actuators. The ease of troubleshooting of the system is expected by introduction of fault detection and self-testing functions, and will be further explored.

The proposed distributed FCS system rises many questions about the choice of the appropriate control concept, dependability, implementation of fault detection and dependencies between control units. The next step in our work is to build a dedicated HIL simulator using the Matlab/Simulink environment to answer these open questions and estimate the benefits and disadvantages of the proposed architecture.

LITERATURE

[1] P. Bieber, F. Boniol, M. Boyer, E. Noulard, and C. Pagetti, "New Challenges for Future Avionic Architectures", Aerospace Lab Journal, Issue 4, May 2012.

[2] R.P.G. Collinson, "Introduction to Avionics Systems", Springer; 2nd edition July 17, 2006.

[3] US Department of transportation, Federal Aviation Administration, "25.1309-1A - System Design and Analysis", Advisory Circulars, June 21, 1988.

[4] Jim Moore, "The Avionics Handbook – Ch. 33. Advanced Distributed Architectures", CRC Press LLC, 2001.

[5] M. Halle, F. Thielecke, "Next Generation IMA Configuration Engineering – from Architecture to Application", IEEE 34st Digital Avionics Systems Conference, September 13-17, 2015.

[6] D. Briere, C. Favre, P. Traverse," Electrical Flight Controls, From Airbus A320/330/340 to Future Military Transport Aircraft: A Family of Fault-Tolerant Systems", CRC Press LLC, 2001.

[7] United States Patent, "US6,443,399B1", September 3, 2002.

[8] A. G. Kuznetsova, Z. S. Abutidzeb, B. I. Portnova, V. I. Galkina, and A. A. Kalika, "Development of MEMS Sensors for Aircraft Control Systems," Gyroscopy and Navigation, vol. 2, no. 1, pp. 59–62, 2011.

[9] D. Arch. (2012, December) MEMS Journal, Inc. [Online]. http://www.memsjournal.com/2010/12/high-performance-mems-gyroscopes-current-status-and-emerging-trends.html

[10] Amos Albert, "Comparison of Event-Triggered and Time-Triggered Concepts with Regard to Distributed Control Systems", Embedded World 2004.

[11] K. Ahlstrom, J. Torin, K. Fersan, P. Nobrant, "Redundancy management in distributed flight control systems: experience & simulations", IEEE 21st Digital Avionics Systems Conference 27-31, Oct. 2002.

Control of thermal process with Simulink and NI USB-6211 in real time

Ivan Tikvić, Goran Vujisić, Mato Fruk

Elektrotehnički odjel
Tehničko veleučilište u Zagrebu
Konavoska 2, Zagreb, Hrvatska
itikvic@tvz.hr, gvujisic@tvz.hr , mato.fruk@tvz.hr

Summary– This paper describes method for control and regulation of small chamber heating thermal process using the acquisition card NI USB-6211 and MATLAB Simulink in real time. It will also point out the problems that have arisen in the process of testing the card, a description of the necessary steps which allow the card to work with Simulink and results obtained by testing.

I. INTRODUCTION

By the development of acquisition device it became much easier to collect and process data from many physical processes. Therefore, the idea has appeared that, by using Matlab's Simulink® package, the facilities could be simulated and the controller could be adjusted and optimized in the laboratory. It would save a lot of time required for starting plant production. In this paper physical model was made and the controller was designed in Simulink®. The connection between controller and physical model of thermal process was realized with NI USB-6211 acquisition card. Also, acquisition card was used for the identification of transfer function of the model. The goal was to test USB (Universal Serial Bus) card to work in Real-Time.

To work with Data acquisition card and Matlab Simulink it is necessary to have a program package Data Acquisition Toolbox, which enables Simulink to recognize the card and allow use of its inputs and outputs. Using a card for the purpose of closed loop control is shown on the simple thermal process, which has large time constants. Therefore, sampling frequency does not need to be large. In this case there is no latency.

II. DATA ACQUISITION

Collecting data or data acquisition is process of measuring electrical or physical values such as voltage, current, temperature, pressure, or sound with computer. In comparison with analog acquisition systems, digital systems have less expressed problems with noise in signals and drift. Also, it is much easier to transmit and store digital signals. Acquisition system is consisted of sensors, measurement hardware and computers with programmable software. Compared to traditional measurement system, computer-based acquisition uses the processing power, flexibility and provides financially profitable solution for measurement and data storage. The task of acquisition card is analog to digital and digital to analog conversion. The following picture shows the flow of data collecting:

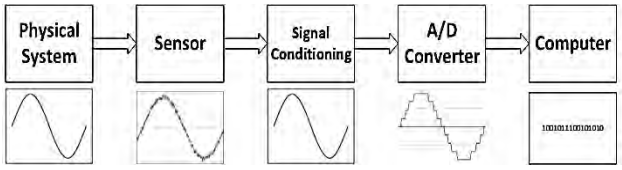


Figure 1Flow of data collecting

III. DATA ACQUISITION CARD NI USB-6211

National Instructions USB-6211 belongs to M series of multifunctional acquisition devices that allow collecting and processing various types of signals. Having a USB port makes card mobile and easy to use, which simplifies work on the various plants and situations where other measurement equipment is difficult to use. The card requires 5 V power supply which is available on PC (Personal Computer) port. Maximum allowable voltage on analog inputs and outputs is ±10 V. Specifications of the cards are:

- 16 analog inputs (16-bit, 250kS/s by TDM-a)
- 2 analog outputs (16bit, 250kS/s by TDM-a)
- 4 digital inputs
- 4 digital outputs
- 2 counters of 32-bits



Figure 2 NI USB-6211

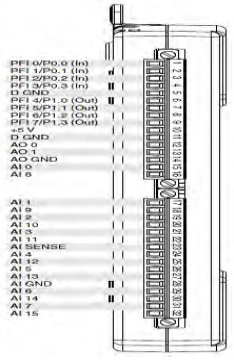


Figure 3 Card connectors

IV. MATLAB® AND SIMULINK® FOR WORKING WITH A CARD

The card is produced and designed to work with National Instruments software packages but it can also be used to work with Matlab's Simulink. To work with Data acquisition card NI USB-6211 it is necessary to have 32-bit version of MATLAB® and OS, because the card is not compatible with 64-bit version of program. Using 64-bit version of Matlab® allows only session-based interface while legacy interface is not available. It would be desirable that the operating system is 32-bit because it is impossible to synchronize 32 bit Matlab® with 64 bit operating system. In this work the Matlab® 2013a was used. In new version of Matlab Simulink Desktop Real-Time is available but as mentioned in this paper synchronization was made with Windows XP with Matlab® 2013a. Also a problem with Windows XP is that it is unable to separate cores to work with only some processes. From Windows 7 on, it is possible to separate cores and use some of them only for one application.

V. REAL-TIME

For real-time operation it is necessary to synchronize Matlab and PC. To do that, we need to install Simulink Real- Time Windows Target program library which synchronizes program work and Kernel. Kernel is the core of the operating system and it controls access of users programs to the systems hardware and software. By the fact that Kernel does so much work, it was necessary to access it in some way and use it to do what the user wants. In this case Simulink Real-Time Windows Target supports real-time sampling rate up to 100 Hz, which is satisfactory for a slow process. After the Real-Time Synchronization Block is opened, two parameters can be modified. Sample time, which determines how often we want the block to make interaction with I/O (Input/Output) ports, and Maximum missed ticks, which defines how many samples can be missed. The block parameters are shown in Figure 4. When more samples are missed then is defined, error occurs and simulation stops.

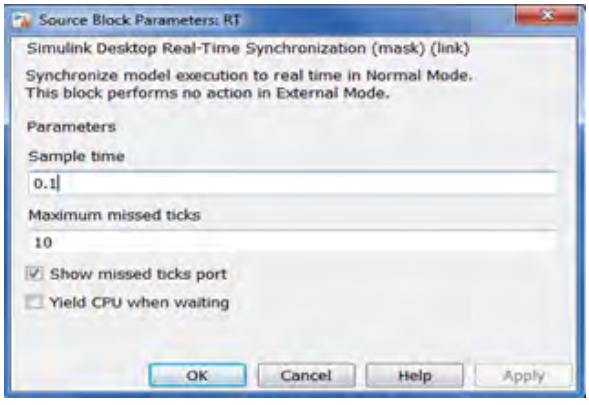


Figure 4 Real-Time Synchronization Block

VI. REGULATED THERMAL PROCESS (INCUBATOR)

A. *TESTING OF PHYSICAL MODEL*

Closed-loop system is realized with reference value and controller from Simulink and model of heating process. Signal goes from controller to the output of the acquisition card and enters the chopper, which gives on the output 0-24 V for the control range form 1-4 V, and thereby controls the heating element (24 V/60 W) which heats the space inside the wooden box. At the opposite side of the heating element there is a PT100 temperature probe with conversion to 0-10 V information for 0-100 ºC. Measuring signal from the probe comes to the cards input and on that way provides a closed-loop system. First step is to determine transfer function of the process in order to identify time constants and gain which are necessary to adjust the controller. Determining the parameters of the transfer function is performed using SIT (System Identification Toolbox). SIT provides the identification of the parameters by comparing response obtained by measuring to some of the known responses types from library. Figure 5 shows test model and a way how to obtain measured values from process.

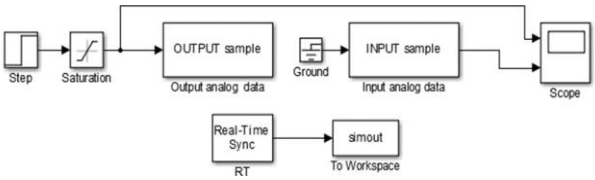


Figure 5 Simulink model for first testing

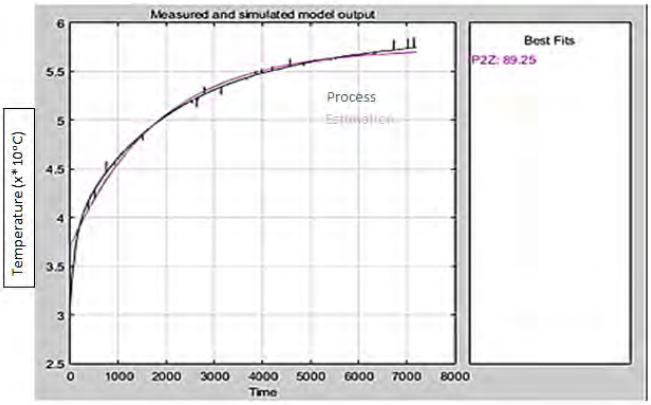


Figure 6 Response of real and estimated process with 89% overlap

From the obtained transfer function (Figure 7), it is type of PDT2 (Proportional plus derivative second order process with damping bigger than unity. Process includes minor lag and lead time constant and large time constant) process. Comparison of the estimated response to the recorded response is shown in Figure 6. Overlapping of response is a bit more than 89%, which is a satisfying percentage. Heating object is performed as 11dm^3

wooden chamber. Heating element (60 W) heats air inside and also heats chamber walls. Also a part of heating energy gets lost through the walls to the environment. This causes large lag time constant and a minor lead time constant of heating transfer function.

B. TESTING OF *SIMULATION MODEL AND CONTROLLER ADJUSTMENTS*

To select and adjust the process controller it will be used simulation model (Figure 7), which allows tracking effect of changes of specific controller parameters on Bode diagrams (Figures 8 and 10) and response of the process. To ensure the quick response of the process, it is necessary to properly set the parameters of selected PI controller. The controller is adjusted for small changes in the environment of the working point and for the purpose of the process it should follow the reference and satisfactorily quickly compensate a disturbance.



Figure 7 Simulation model

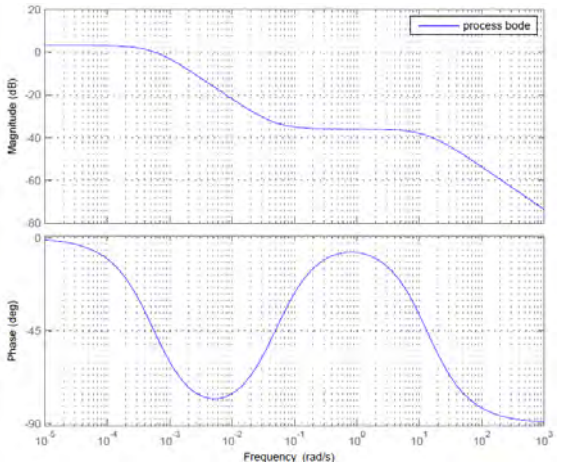


Figure 8 Bode diagrams of the process model without a controller

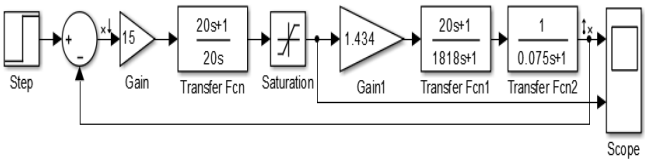


Figure 9 Simulation model with PI controller

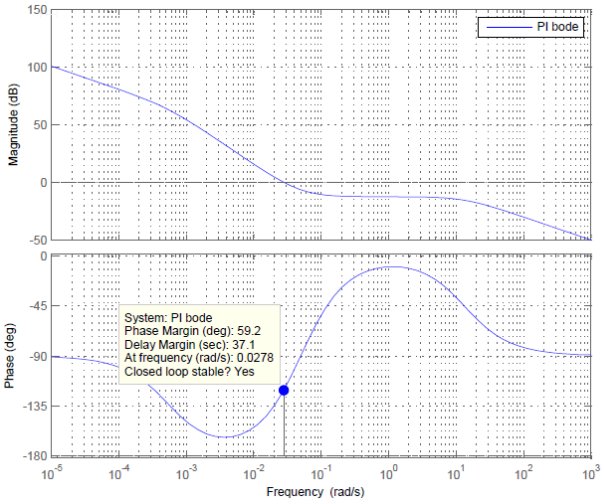


Figure 10 Bode diagrams of the process model with PI controller

C. *TESTING OF THE PHYSICAL MODEL WITH CONTROLLER*

To see the impact of the selected controller on the real model it was made a test during which has been modified the reference value (Figure 11). The goal was to get a quick response without the windup effect. Test model with controller is shown in Figure 12.

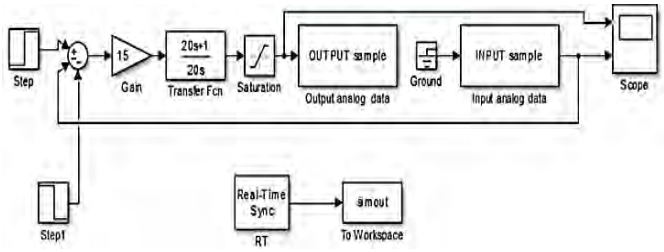


Figure 11 Real process with controller

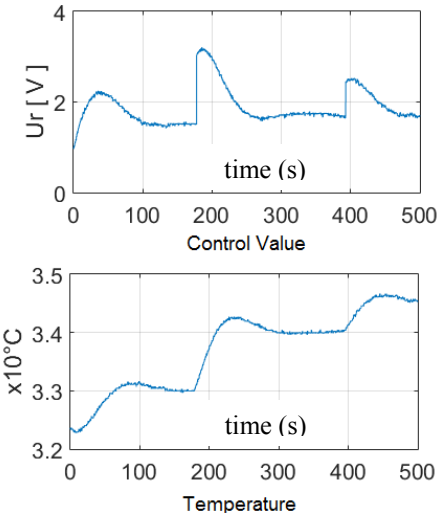


Figure 12 Reaction of controller and temperature response on reference changes

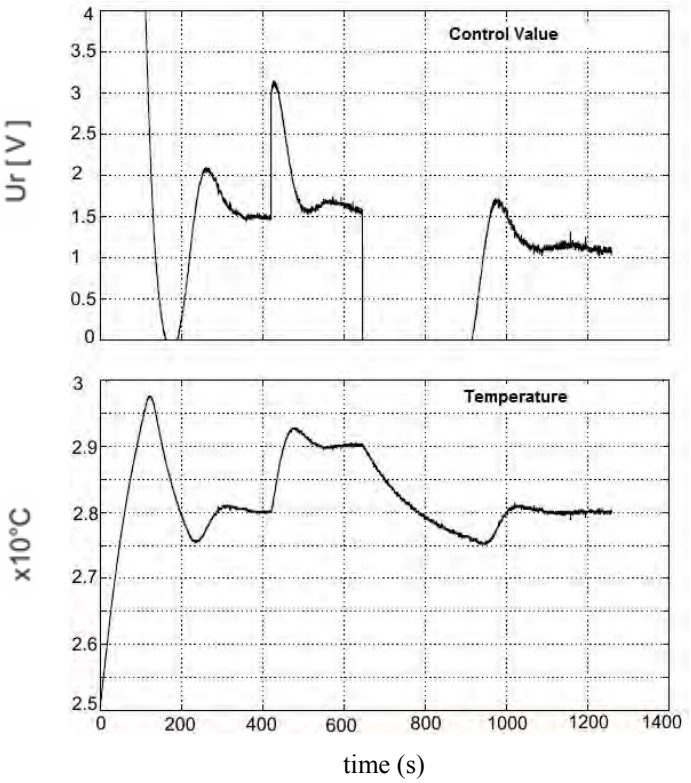


Figure 13 Heating and impact disturbances

Figure 13 shows that a big step of temperature from 25 ºC to 28 ºC is set on the beginning, and that output of controller goes to saturation with temperature overshoot of almost 2 ºC. Reaching desired temperature or entry into 2% has been achieved within 400 s. After the temperature had been stabilized at 28 ºC, it is set to 29 ºC. New temperature changes does not cause the controller output to saturate and temperature overshoot is less than 0.5 ºC. After the temperature had been stabilized at 29 ºC, the room starts to cool because new reference is set to 28 ºC. Disturbance was realized by opening the wooden chamber. After the temperature has reached 27.5 ºC the wooden chamber was closed and with the reference value set at 28 ºC the controller turned on and heated the wodden chamber to desired temperature.

A conclusion after examining is that the controller satisfactory maintains desired temperature.

VII. CONCLUSION

Testing of NI USB-6211 acquisition card with Matlab® and Simulink® shows that card's specifications can not be fully utilized for the realization of a regulated process because it requires work in real time. The problems that are manifested in limited number of sampels, up to 100 in a second, that are not enough to control fast processes such as motor speed control. In order to enable real time operations with high sampling rate, it will be necessary to separate processor cores in the purpose of using them only for a specific process. The problem is also that the card can not be fully used on 64-bit computers ie. It can be used for collecting and processing but not for sending data. For future work this method will be tested on 32-bit operating system that can seperate cores with 32-bit Matlab®. Based on the peroformed work it can be concluded that the testing and adjustmens of controller can be done wihout frequent trips to the factories which greatly facilitates enginners job.

REFERENCES

1. NI USB-621x NI USB-621x User Manual; http://www.ni.com/pdf/manuals/371931f.pdf, 10.1.2015.
2. http://www.mathworks.com/, 10.1.2015.
3. Perić, Nedjeljko. Automatsko upravljanje-predavanja, FER, Zagreb, 2006.
4. https://en.wikipedia.org/wiki/Kernel_(operating_system) , 10.1.2015.
5. A/D pretvorba i način prijenosa digitalnog signala; https://www.fer.hr/_download/repository/V03-PRIZVU2012.pdf, 10.1.2015.
6. https://www.fer.unizg.hr/_download/repository/SIMULINK_SKRIPTA.pdf, 10.1.2015.
7. https://bib.irb.hr/datoteka/640486.Diplomski_rad_Marenzi.pdf, 10.1.2015.
8. Fruk, Mato; Vujisić, Goran; Špoljarić, Tomislav. Parameter Identification of Transfer Function Using MATLAB // MIPRO 2013, Proceedings, Computers in Education – CE, Opatija, 2013, 697-702

Stabilization of Multi-AUV formation with digital control

S.A. Ul'yanov* and N.N. Maximkin*
* Institute for System Dynamics and Control Theory / Department of Information Control Systems, Irkutsk, Russia
sau@icc.ru

Abstract - Formation stability problem for a team of autonomous underwater vehicles (AUV) with digital control is investigated. A leader-follower formation control algorithm for multi-AUV system is proposed. Feedback control signal for the follower, which is piecewise constant, is computed on the basis of discrete measurements of the distance and the bearing angle to the leader. A discrete observer for estimation of the derivatives of the distance and the bearing angle are used to improve the control performance. The parameters of the controller and observer are synthesized using a digital control system design method based on the vector Lyapunov functions that allows one to take into account the uncertainties of the plant, measurement errors, and control constraints. Numerical simulations are carried out to demonstrate the effectiveness of the proposed control scheme.

I. INTRODUCTION

Formation control is one of the most actively studied problems in the field of multi-agent systems. It is due to a wide spectrum of potential applications, which are exploring and monitoring various physical environments (some of them are hard to access or dangerous), moving heavy loads and search-and-rescue activities. More and more attention is being paid to formations of autonomous underwater vehicles (AUV) due to the needs to explore the depth of the World Ocean and its resources.

There are three basic approaches to formation control [1]: the leader-follower approach, the virtual structure approach and the behavior-based approach. Formation control strategies for multi-AUV systems are, for the most part, based on the leader-follower approach (see [2-4]). According to this approach, all the agents in a multi-agent system are connected in pairs by the leader-follower relation so that there is an agent that is not a follower for any other agent, which is called the leader of the formation and determines its motion, whereas each other agent as a follower try to maintain a desired position with respect to its leaders.

Formation control problem becomes more practical if a digital or sampled-data control scheme is used, since most of the modern control systems are digital. The models of such systems combine the continuous-time dynamics of the plant and the discrete-time dynamics of the controller, i.e., they are heterogeneous in nature. Known approaches to the design and analysis of digital control systems are usually based on the transition to the discrete time via preliminary discretization of the heterogeneous system. However, with this transition, some peculiarities of digital systems are fully or partly ignored, which ultimately may decrease the accuracy of the analysis of such systems.

The formation control problem for multi-agent systems in sampled-data settings (with heterogeneous dynamics of agents) has been scarcely studied so far. Particularly, this problem is investigated in [5-8]. In these studies, the sample-and-hold mechanism is used to model the information exchange between agents. More precisely, each agent samples and transmits the information about its state (position, orientation, and velocities) to the local neighbors or followers (if the leader-follower scheme is used), and these latter compute the control actions on the basis of the received data. In this paper, we propose a formation control scheme for multi-AUV systems where any kind of communications between AUVs is not provided and the control protocol for each AUV uses only sampled information about its relative position with respect to the leaders. So, all of the AUVs in the formation act autonomously, what is important in the underwater environment where only unreliable and slow communication channels can be exploited.

In real environments, it is very important to provide formation stability subject to external disturbances, navigation and sensor errors, and restrictions on control. Known conceptions of the formation stability and proposed formation control algorithms (see [9] and references therein), which, for the most part, consider the linear dynamics of agents and/or imply that for each agent all variables of its motion and the motion of the leader can be measured, do not take these factors into account. To tackle this problem, formalizations of the desired behavior of the formation in typical modes of operation as dissipativity and practical stability properties are introduced in [10-11]. On the basis of vector Lyapunov functions, sufficient criteria for the mentioned dynamical properties are derived. They provide background for a computer-aided technology for rigorous analysis and synthesis of nonlinear multi-component systems. It is based on the use of the vector Lyapunov function method [12] and the reduction method [13-14]. In this paper, we adopt this technology for analysis of nonlinear sampled-data models of multi-AUV formations.

The rest of the paper is organized as follows. Section II presents the 3 DOF dynamical model of AUV and dynamical model of the leader-follower formation. In Section III, we propose a sampled-data formation control

scheme, in which individual discrete controller for each follower AUV uses only sampled measurements of the distance and the bearing angle to the leader as well as estimates of their derivatives computed by use of a discrete observer. A computer-aided technology for rigorous analysis and synthesis of nonlinear digital control systems based on the VLF method are presented in Section IV. This technology is applied to the controller design for multi-AUV formation. We demonstrate simulation results in Section V and provide conclusions in Section VI.

II. FORMATION DYNAMICS

Assume that each AUV in the formation operates at the same constant depth so that its heave, pitch, roll motions can be neglected. In that case, the equations of motion of the AUV can be presented as

$$\begin{aligned} \dot{x}_i(t) &= v_i(t)\cos(\varphi_i(t)), \\ \dot{y}_i(t) &= v_i(t)\sin(\varphi_i(t)), \\ \dot{\varphi}_i(t) &= \omega_i(t), \\ m_i\dot{v}_i(t) &= u_i(t) + p_i(\ldots), \\ J_i\dot{\omega}_i(t) &= u_i^{\varphi}(t) + p_i^{\varphi}(\ldots), \end{aligned} \tag{1}$$

where $i = \overline{0,N}$, N is the number of AUVs in the formation (index "0" corresponds to the leader of the formation), x_i, y_i denote the inertial coordinates of the center of mass of the i-th vehicle, v_i is the linear velocity, φ_i is the heading angle, which presents yaw orientation around z-axis, ω_i is the angular velocity, m_i is the mass of the AUV, J_i is the total moment of inertia around vertical z-axis; u_i, u_i^{φ} are the control force and moment. Functions $p_i = p_i(t,\ldots)$ and $p_i^{\varphi} = p_i^{\varphi}(t,\ldots)$ specify the resistence force and the resistance moment as well as other disturbances that depend on a set of variables including t.

Suppose that the mass and the moment of inertia of the AUV can not be accurately determined so that

$$m_i = m_i^* + \Delta m_i(\ldots), \quad J_i = J_i^* + \Delta J_i(\ldots),$$

where m_i^*, J_i^* are their nominal values, $\Delta m_i(\ldots)$, $\Delta J_i(\ldots)$ are time-varying uncertainties that satisfy

$$|\Delta m_i(\ldots)| \le m_i^0, \quad |\Delta J_i(\ldots)| \le J_i^0,$$

m_i^0, J_i^0 are positive constants.

The formation control strategy proposed in this paper is based on the leader-follower approach. The follower seeks to maintain a certain position with respect to its leaders. We suppose that every follower in the formation has only one leader.

We impose the following quite strict restrictions on the work of the formation. It is assumed that each AUV knows nothing about the geometric description of the path of any other vehicle or how its variables of motion change with respect to time. Moreover, any kind of explicit communication in the group of AUVs is forbidden.

Assume that each AUV is equipped with sensors capable of measuring the parameters of its motion (coordinates, velocities and so on) and also the distance to the leader $s_i \equiv \sqrt{(x_{j_i} - x_i)^2 + (y_{j_i} - y_i)^2}$ and the bearing angle $\theta_i \equiv \varphi_i - \arctan\dfrac{y_{j_i} - y_i}{x_{j_i} - x_i}$. Index j_i denotes the leader for the i-th AUV.

The dynamics of a leader-follower pair in terms of variables s_i and θ_i can be described by

$$\begin{aligned} \dot{s}_i &= v_{j_i}\cos(\varphi_{j_i} - \varphi_i + \theta_i) - v_i\cos\theta_i, \\ s_i\dot{\theta}_i &= s_i\omega_i + v_i\sin\theta_i - v_{j_i}\sin(\varphi_{j_i} - \varphi_i + \theta_i). \end{aligned} \tag{2}$$

As an unperturbed motion of the formation, we take the motion with $s_i(t) = s_i^*$, $\theta_i(t) = \theta_i^*$, $v_i(t) = v^*$, $\omega_i(t) = 0$, $\dot{v}_i(t) = 0$, $\dot{\omega}_i(t) = 0$, where s_i^*, θ_i^* are constants that describe the desired position of the i-th AUV with respect to its leader, v^* is the nominal linear velocity of the formation maintained by each AUV.

Denote by $\Delta s_i \equiv s_i - s_i^*$, $\Delta\theta_i \equiv \theta_i - \theta_i^*$ the variables that define the relative position error of the follower AUV. Using Taylor expansions of the right sides of (2), the error dynamics can be written as

$$\begin{aligned} \Delta\dot{s}_i &= \cos(\theta_i^*)\Delta v_{j_i} - \cos(\theta_i^*)\Delta v_i - v^*\sin(\theta_i^*)\Delta\varphi_i + \Delta\tilde{s}_i(\cdot), \\ \Delta\dot{\theta}_i &= \omega_i + \frac{\sin(\theta_i^*)}{s_i^*}\Delta v_i - \frac{\sin(\theta_i^*)}{s_i^*}\Delta v_{j_i} - \frac{v^*\cos(\theta_i^*)}{s_i^*}\Delta\varphi_i + \\ &\quad + \Delta\tilde{\theta}_i(\cdot), \end{aligned} \tag{3}$$

where $\Delta v_i \equiv v_i - v^*$, $\Delta v_{j_i} \equiv v_{j_i} - v^*$, $\Delta\varphi_i \equiv \varphi_{j_i} - \varphi_i$, $\Delta\tilde{s}_i(\cdot)$ and $\Delta\tilde{\theta}_i(\cdot)$ are some polynomials that depend on the components of vector $\zeta \equiv (\Delta s_i, \Delta\theta_i, \Delta\varphi_i, \Delta v_i, \Delta v_{j_i})^T$ and satisfy

$$|\Delta\tilde{s}_i(\zeta)| \le \Omega_i(|\zeta|), \quad |\Delta\tilde{\theta}_i(\zeta)| \le \Omega_i^{\varphi}(|\zeta|),$$

$\Omega_i(\cdot)$, $\Omega_i^{\varphi}(\cdot)$ are nondecreasing, upper-semicontinuous functions.

From (1) it is easy to derive the equations for $\Delta\varphi_i$, Δv_i, ω_i as

$$
\begin{aligned}
{}_{\Delta}\dot{\varphi}_i &= \omega_{j_i} - \omega_i, \\
m_i\,{}_{\Delta}\dot{v}_i &= u_i + p_i, \\
J_i\dot{\omega}_i &= u_i^{\varphi} + p_i^{\varphi},
\end{aligned} \tag{4}
$$

System (3), (4) presents a formation model in deviations which will be used in control design in the sequel.

Finally, we impose constraints on the maneuvers of the leader of the formation as

$$
|v_0 - v^*| \le v_0^0, \quad |\omega_0| \le \omega_0^0, \quad |\dot{v}_0| \le w_0^0, \quad |\dot{\omega}_0| \le \varepsilon_0^0,
$$

where v_0^0, ω_0^0, w_0^0, ε_0^0 are some positive constants.

III. CONTROL ALGORITHM

Based on the measurement capabilities of AUV, we build orientation and translational controllers as (in the sequel, we will omit the subscript index denoting an individual vehicle)

$$
u = u_s + u_c, \qquad u^{\varphi} = u_s^{\varphi} + u_c^{\varphi}, \tag{5}
$$

where $u_c = u_c(t,\ldots)$, $u_c^{\vartheta} = u_c^{\vartheta}(t,\ldots)$ are compensating control terms aimed to overcome the resistance force and moments as well as to counteract the disturbances so that

$$
|u_c - p| \le p^0, \quad |u_c^{\varphi} - p^{\varphi}| \le p^{\varphi 0},
$$

p^{φ}, $p^{\varphi 0}$ are some positive constants; $u_s = \Phi(\sigma)$, $u_s^{\varphi} = \Phi^{\varphi}(\sigma^{\varphi})$ are feedback control terms aimed to stabilize the formation.

For the feedback control, we use piecewise constant signal which is updated at time moments $t_k = kh$, $k = 0,1,2,\ldots$, h is the control step, which is supposed common for all AUVs in the group. Also we assume that several measurements of variables s and θ are available for the follower AUV on every control interval $T_k \equiv [t_k, t_{k+1})$. Let $t_{k,\nu} \equiv t_k + \tau_\nu$ be a time instant when the ν-th measurement is obtained, $0 \le \tau_\nu \le h$, $\nu = \overline{1,\mu}$, μ is the number of measurements available on the control step.

Thus, the feedback control law may be taken as

$$
\begin{bmatrix} \sigma(t) \\ \sigma^{\varphi}(t) \end{bmatrix} = K\eta_k, \quad t \in T_k,
$$

where K is 2×4 -matrix of feedback coefficients, $\eta_k \equiv ({}_{\Delta}\hat{s}_k, {}_{\Delta}\hat{\theta}_k, z_k, z_k^{\varphi})^T$,

$$
{}_{\Delta}\hat{s}_k = \sum_{\nu=1}^{\mu} \lambda_\nu (\Psi(s(t_{k-1,\nu})) - s^*),
$$

$$
{}_{\Delta}\hat{\theta}_k = \sum_{\nu=1}^{\mu} \lambda_\nu^{\varphi} (\Psi^{\varphi}(\theta(t_{k-1,\nu})) - \theta^*),
$$

are estimates of ${}_{\Delta}s$ and ${}_{\Delta}\theta$ obtained at t_k; z_k, z_k^{φ} are the outputs of discrete observers given by difference equations

$$
z_{k+1} = \sum_{\nu=1}^{\mu} a_\nu (\Psi(s(t_{k,\nu})) - s^*) + bz_k, \qquad z_0 \equiv z(t_0) = 0,
$$

$$
z_{k+1}^{\varphi} = \sum_{\nu=1}^{\mu} a_\nu^{\varphi} (\Psi^{\varphi}(\theta(t_{k,\nu})) - \theta^*) + b^{\varphi} z_k^{\varphi}, \qquad z_0^{\varphi} \equiv z_0^{\varphi}(t_0) = 0.
$$

The observers is aimed to estimate time derivatives ${}_{\Delta}\dot{s}$, ${}_{\Delta}\dot{\theta}$ at moments t_k.

To define the output characteristics of the actuators and sensors $\Phi(\cdot)$, $\Phi^{\varphi}(\cdot)$, $\Psi(\cdot)$, $\Psi^{\varphi}(\cdot)$, we introduce the following definition.

A function $\theta : R^1 \to R^1$ is a function of class $SN(k, \xi^0, \bar{\xi}, k^0)$ [12,15] if

$$
\begin{aligned}
\min\left\{k\left(\xi - \xi^0\right) - k^0|\xi|;\, k\bar{\xi}\right\} \le \theta(\cdot,\xi) \le \\
\le \max\left\{k\left(\xi + \xi^0\right) + k^0|\xi|;\, -k\bar{\xi}\right\}, \\
(0 \le \xi^0 < \bar{\xi} \le \infty),
\end{aligned}
$$

Parameters of the class k, k^0 specify the nominal value and admissible uncertainties of the gain coefficient; ξ^0 specifies the admissible error, dead zones, and quantization level; $\bar{\xi}$ characterizes the control saturation and the operating ranges of the sensors. Graphical representation of the SN class is depicted in Fig. 1. A

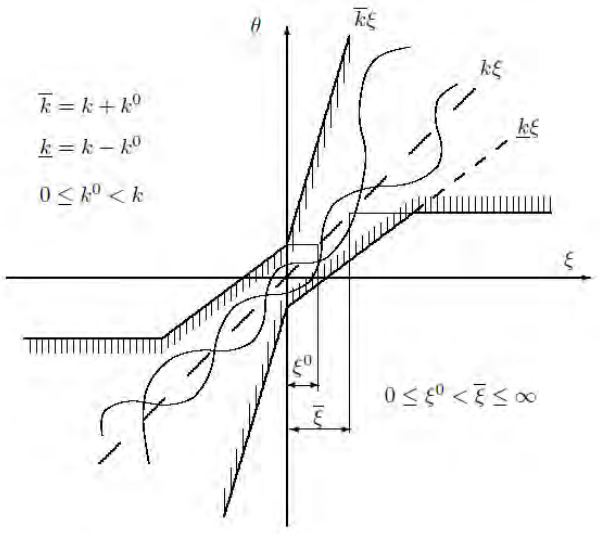


Figure 1. Graphical representation of the SN class

SN class function lies inside of the area bounded by the hatched lines. It should be noted that this class contains many typical nonlinearities of the absolute stability theory.

Now, we define the output characteristics of the actuators and sensors as

$$\Phi \in SN(1,0,\bar{\sigma},0), \quad \Phi^{\varphi} \in SN(1,0,\bar{\sigma}^{\varphi},0),$$
$$\Psi \in SN(1,s^0,\infty,0), \quad \Psi^{\varphi} \in SN(1,\theta^0,\infty,0).$$

The formation control design problem consists in finding the feedback coefficients in matrix K and observer's parameters λ_v, λ_v^{φ}, a_v, z_k, a_v^{φ}, b^{φ} that provide the dissipativity of closed system (3)-(5) subject to the given uncertainties and disturbances, and minimize ultimate stabilization error of the formation

$$J_f = \left\| \operatorname*{col}_{i=\overline{1,N}} \left(\limsup_{t\to\infty} \limsup_{t'\geq t} |\, {}_{\Delta}s_i(t') \,| + \alpha \limsup_{t\to\infty} \limsup_{t'\geq t} |\, {}_{\Delta}\theta_i(t') \,| \right) \right\|.$$

IV. Control Design Method

In this section, we give a brief review of a computer-aided technology for strict analysis and synthesis of nonlinear digital control systems based on sublinear vector Lyapunov functions.

Let the disturbed motion of a plant together with the dynamics of state observers and other elements of the control system be described by differential equations of the form

$$\begin{aligned}(J+\Delta J(t,x))\dot{x} = (A+\Delta A(t,x))x + (B+ \\ +\Delta B(t,x,u))u + GF(t,x,u) + H\Phi(t,x,u), \quad x \in R^n.\end{aligned} \tag{6}$$

Control $u = u_k = \varphi(\cdot,\sigma_k) \in R^m$, $\sigma_k = Cz_k \in R^m$ is piecewise constant on intervals $T_k = [t_k, t_{k+1} = t_k + h)$, $k = 0,1,2...$ and is formed on account of the output $z_k \in R^{n_d}$ of the discrete (dynamic) controller

$$\begin{aligned}z_{k+1} = A_d z_k + C_d \psi(\cdot,\eta_k) + H_d \Phi_d(\cdot), \\ \psi(\cdot,\eta_k) \in R^l, \quad \eta_k = \sum_{i=\overline{1,\mu}} D_i x(t_{k_i}) \in R^l.\end{aligned} \tag{7}$$

Here, $t_{k_i} = t_k + \tau_i$ are the time points of discrete measurements, $\mu \geq 1$ is the memory depth; vector functions $\Delta J(\cdot)$, $\Delta A(\cdot)$, $\Delta B(\cdot)$, $F(\cdot)$, $\Phi(\cdot)$, $\Phi_d(\cdot)$ specify uncertainties, nonlinearities, perturbations of the control system, and satisfy the constraints

$$|J^{-1}\Delta J(\cdot)| \leq J^0, \quad |\Delta A(\cdot)| \leq A^0, \quad |\Delta B(\cdot)| \leq B^0,$$
$$|F(\cdot)| \leq U(|w|) \in R^p, \quad w = Wx \in R^q,$$
$$\Phi(\cdot)| \leq \Phi^0, \quad |\Phi_d(\cdot)| \leq \Phi_d^0,$$

where $U(\cdot)$ is a nondecreasing upper-semicontinuous function; functions $\varphi(\cdot)$, $\psi(\cdot)$ is of the SN class.

It is easy to show that closed system (3)-(5) can be presented in the form of model (6), (7).

We formulate the control design problem as to find parameters of system (6), (7) so that it was stable or, strictly speaking, dissipative [10,11] for all admissible uncertainties and disturbances and the highest accuracy of stabilization is achieved.

Let it be required to determine the optimal values of a vector of parameters $\beta \in \mathfrak{B}$ (usually, these are elements of the feedback matrices C, C_d, etc.), where by $\mathfrak{B}$ we denote the area of admissable values for β. Denote by $\bar{\gamma}_f$ the estimate of ultimate accuracy of the control system (a vector that satisfies

$$\bar{\gamma}_f \geq \overline{\lim_{t\to\infty}} |\, \gamma_f(t) \,| \equiv \limsup_{t\to\infty} \limsup_{t'\geq t} |\, \Gamma_f x(t', \mathrm{x}_0) \,| \tag{8}$$

and by $\mathfrak{D}$ the dissipativity domain defined as a set of initial states of the system containing the neighborhood of the origin, for which the relation (8) is satisfied for any admissible disturbaces.

Now, the control design problem can be formulated as follows: find the values of parameters $\beta \in \mathfrak{B}$ that minimize the criterion

$$J_f = \alpha^T \bar{\gamma}_f \tag{9}$$

($\alpha \in R_+^v$ is a given vector of weighting coefficients) provided that system (6), (7) is dissipative, i.e., $\mathfrak{D} \neq \varnothing$ (domain of dissipativity is not empty).

The posed problem cannot be rigorously solved, since, in the general case, it is impossible neither to get the exact value of the criterion, nor to test the feasibility of the restrictions. The idea of the approach used here is, instead of exact indicators of performance, from which the criterion and constraints depend on, to use their estimates. For system (6), (7), such estimates can be effectively computed using sublinear VLFs. To construct VLFs and corresponding comparison systems (CSs), to verify a dynamic property, and to compute estimates, we use modifications of algorithms from [15-17].

The sublinear VLF is constructed in the form of vector norm

$$\begin{aligned}v(x,z,\beta) = \operatorname*{col}_{j=\overline{1,s}} \left\| S^j(\beta) \begin{pmatrix} x \\ z \end{pmatrix} \right\|_j \in R^s, \\ 1 \leq s \leq \bar{n} = n + n_d,\end{aligned} \tag{10}$$

where S^j are row blocks of $(\bar{n} \times \bar{n})$ matrix $S(\beta) = [S_c(\beta), S_d(\beta)]$, which transforms matrix L of the system

$$\begin{pmatrix} x_{k+1} \\ z_{k+1} \end{pmatrix} = L(\beta) \begin{pmatrix} x_k \\ z_k \end{pmatrix} \quad (11)$$

to a near diagonal form. System (11) represents an exact discretization of the "unperturbed" linearized model

$$\dot{x} = J^{-1}Ax + J^{-1}BK_\sigma C(\beta)z_k,$$
$$z_{k+1} = A_d z_k + C_d(\beta)K_\eta \sum_{i=1,\mu} D_i x(t_{k_i}),$$

at time points t_k ($x_k = x(t_k)$). The comparison system, whose solutions majorize the behavior of the VLF at time moments t_k , may be obtained as

$$y_{k+1} = P_\beta y_k + Q_\beta g_\beta(y_k) + R_\beta \triangleq f_\beta(y_k) \in R^s, \quad (12)$$

where P_β, Q_β, R_β are, respectively, nonnegative matrices and the vector; $g_\beta(y)$ is a nondecreasing subdifferentiable function, $g_\beta(0) = 0$.

The control design problem in terms of sublinear VLF (10) and CS (12) can be reformulated as to minimize the criterion (9) with $\overline{\gamma}_f = |\Gamma_f S_c^-(\beta)|\, q_\beta$ under the condition that the spectral radius of non-negative matrix $L_\beta^c \triangleq P_\beta + Q_\beta \partial g_\beta(q_\beta)$ is smaller than 1. Here, $\partial g_\beta(q_\beta) = \{g_\beta^i(q_\beta, e^j)\}$ is the generalized Jacobian; q_β is the rest point of CS ($f_\beta(q_\beta) = q_\beta$).

Procedures for constructing abovementioned VLFs, correspinding CSs, computing the indicators of dynamic performance (stabilization accuracy, attraction domain, reachable domain, control time, etc) and synthesis of control parameters for the considered continuous-discrete system has been implemented in software package VLF-REDUCTOR.

The main advantages of the described technology are, on the one hand, the strictness of results obtained and guaranteed nature of the judgments and estimates, and, on the other hand, its constructiveness. It is for the accuracy of numerical results VLF method is advantageous over other known methods. It is important that analysis problems are formulated directly in terms of system requirements, which are convenient for engineers; the results are also obtained in this form. After reduction of an investigated model to some standard form, subsequent computations are performed, for the most part, automatically.

V. NUMERICAL COMPUTATIONS AND SIMULATION

Numerical computations are conducted for a group of four AUVs (one leader and three followers) uniformly distributed around a circle of radius 10 m. We assume that all the AUVs are identical and the model parameters are $v^* = 1$ m/s (nominal velocity of the formation), $m_i^* = 150\,\text{kg}$, $J_i^* = 100\ \text{kg}\cdot\text{m}^2$ (nominal values of the AUV's inertial parameters), $m_i^0 = 2$ kg, $J_i^0 = 1\ \text{kg}\cdot\text{m}^2$ (uncertainties of the inertial parameters), $p_i^0 = p_i^{0\varphi} = 1$ (admissible disturbances), $\overline{\sigma}_i = 200\,\text{kg}\cdot\text{m/s}^2$, $\overline{\sigma}_i^\varphi = 100\,\text{N}\cdot\text{m}$ (control constraints), $s_i^0 = 0.3$ m, $\theta_i^0 = 0.03$ rad (errors of the distance and bearing angle sensors), $v_0^0 = 0.1$ m/s, $\omega_0^0 = 0.05\,\text{rad/s}$, $w_0^0 = 0.01\ \text{m/s}^2$, $\varepsilon_0^0 = 0.01\ \text{rad/s}^2$ (constraints on the maneuvers of the formation leader), $h = 0.5\,\text{s}$ (control sampling interval), $\mu = 2$, $\tau_1 = 0$, $\tau_2 = 0.5$ (sampling times for measurements).

In simulation, the leader navigates with constant linear velocity $v(t) = 1\,\text{m/s}$ and its angular velocity changes as follows

$$\omega_0(t) = \begin{cases} 0, & 0 \le t < 5 \text{ and } t \ge 85, \\ 0.065, & 5 \le t < 45, \\ -0.065, & 45 \le t < 85. \end{cases}$$

All the follower AUVs are static at $t = 0$ and their initial positions are $x_0(0) = 10$, $y_0(0) = 0$, $x_1(0) = 5$, $y_1(0) = 15$, $x_2(0) = -5$, $y_2(0) = 0$, $x_3(0) = -5$, $y_3(0) = 15$. Hence, the task of the formation is to change the direction of motion from the East to the North.

Fig. 2 shows the trajectory of AUVs for the given simulation scenario. Stabilization errors in distance and angle are shown in Fig. 3 and Fig. 4 respectfully. From Fig. 3-4 one can see that the stabilization errors in distance and bearing angle for each AUV tend to zero.

VI. CONLUSION

In this paper, the problem of stabilization of multi-AUV formation operating in 2D is studied taking into account the quantization of measured data and control signals, errors of sensors, restrictions on control actions, and external disturbances. We propose a decentralized control scheme for multi-AUV formation, in which local

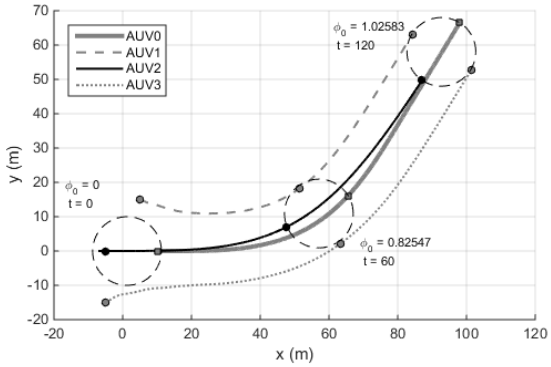


Figure 2. Trajectory of the formation. AUV0 is the leader of the formation

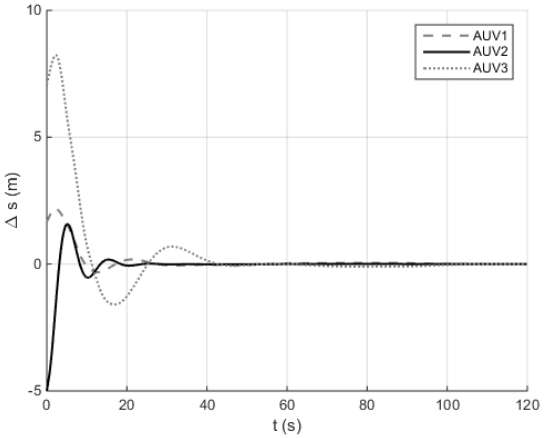


Figure3. Errors in the distance

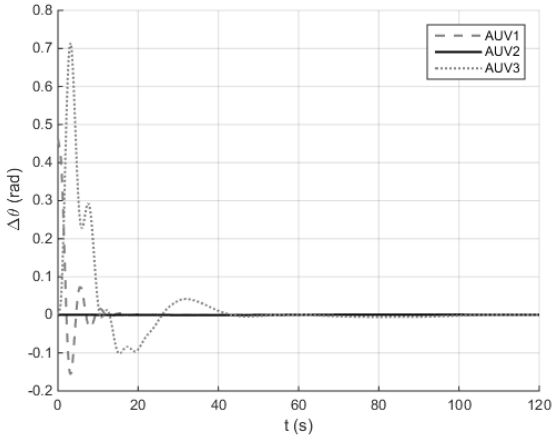


Figure 4. Errors in the bearing angle

controller for an individual follower AUV uses only sampled data of its relative position with respect to the leader, i.e., the information exchange between AUVs is not provided. A VLF-based approach to analysis and synthesis of continuous-discrete systems is described and results of its application to the formation controller design problem is given. The results of the numerical study of the designed control system are presented to demonstrate its performance.

ACKNOWLEDGMENT

The research was carried out at the expense of the Russian Science Foundation (project No. 16-11-00053).

REFERENCES

[1] Y. Q. Chen, Z. M. Wang, "Formation control: a review and a new consideration," Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3181–3186, 2005.

[2] X. Li, D. Zhu, Y. Qiun, "A survey on formation control algorithms for multi-AUV system," Unmanned Systems, 2(4), pp. 351–359, 2014.

[3] R. Cui, S. S. Ge, B. V. E. How, Y. S. Choo, "Leader-follower formation control of underactuated autonomous underwater svehicles," Ocean Eng., 37(17), pp. 1491–1502, December 2010.

[4] Y. Dai, S. G. Lee, "The leader-follower formation control of nonholonomic robots," International Journal of Control, Automation and Systems, 10(2), pp. 350–361, April 2012.

[5] T. Hayakawa, T. Matsuzawa, S. Hara, "Formation control of multi-agent systems with sampled information relationship between information exchange structure and control performance," IEEE Conferrence on Decision and Control, pp. 4333–4338, December 2006.

[6] P. Millan, L. Orihuela, I. Jurado, F. Rodriguez Rubio, "Formation control of autonomous underwater vehicles subject to communication delays," IEEE Transactions on Control System Technology, 22(2), pp. 770–777, March 2014.

[7] C. Ma, Q. Zeng, "Distributed formation control of 6-DOF autonomous underwater vehicles networked by sampled-data information under directed topology," Neurocomputing, vol. 154, pp. 33–40, April 2015.

[8] B. Zheng, X. Mu, "Formation-containment control of second-order multi-agent systems with only sampled position data," International Journal of Systems Science, Novermber 2015.

[9] K. K. Oh, M. C. Park, H. S. Ahn, "A survey of multi-agent formation control," Automatica, vol. 53. pp. 424–440, March 2015.

[10] S. N. Vassilyev, R. I. Kozlov, S. A. Ulyanov, "Analysis of coodinate and other transformation of models of dynamical systems by the reduction method," Proceedings of the Steklov Institute of Mathematics, vol. 268, pp. 264–282, April 2010.

[11] S. N. Vassilyev, R. I. Kozlov, S. A. Ul'yanov, "Mutimode formation stability," Doklady Mathematics, 89(2), pp. 257–262, March 2014.

[12] R. Z. Abdyllin, L. J. Anapolski, R. I. Kozlov, A. I. Malikov, V. M. Matrosov, A. A. Voronov, A. S. Zemljakov, "Vector Lyapunov functions in stability theory," University of Michigan: World Federation Publisher Company, 1996.

[13] S. N. Vassilyev, "Method of reduction and qualitative analysis of dynamic systems: I," Journal of Computer and Systems Sciences International, vol. 45, issue 1, pp. 17–25, January 2006.

[14] S. N. Vasil'ev, "Method of reduction and qualitative analysis of dynamic systems: II," Journal of Computer and Systems Sciences International, vol. 45, issue 2, pp. 167–179, March 2006.

[15] I. V. Bychkov, V. A. Voronov, E. I. Druzhinin, "Synthesis of a combined system for precise stabilization of the SPECTR-UF observatory: II," Cosmic Research, 52(2), pp. 145–152, 2014.

[16] R.I. Kozlov, O.R. Kozlova, "Investigation of stability of nonlinear continuous-discrete models of economic dynamics using vector Lyapunov function. I," Journal of Computer and Systems Sciences, no. 2, pp. 370–378, 2009.

[17] S. V. Burnosov, R. I. Kozlov, "Investigation of the dynamics of nonlinear systems with uncertainty and perturbations on the basis of the vector Lyapunov functions II," J. Comp. Syst. Sci. Int., 34(2), pp. 82–90, 1996.

A hybrid approach to solve the dynamic patrol routing problem for group of underwater robots

M.Yu. Kenzin, I.V. Bychkov and N.N. Maksimkin
Matrosov Institute for System Dynamics and Control Theory SB RAS, Irkutsk, Russia
gorthauers@gmail.com

Abstract - We suggest an approach to solve the dynamic patrolling problem as the variation of periodic vehicle routing problem with time windows. In order to do this, firstly a set of waypoints (targets) should be located in space in such way that their spherical neighborhoods, which corresponds to the size of the robots vision/detections fields, would form a complete coverage of the protected perimeter. Regular inspections of each waypoint by robots of the group not less than once during specified period would guarantee the effective perimeter patrolling. Thus, the routing problem is to a feasible group route ensuring well-timed inspections for all waypoints of the mission. To solve the described problem we propose a hybrid approach, based on the combined use of genetic algorithms and methods and heuristics of local search. The proposed approach allows robots to effectively generate required solutions under given constraints and quickly readjust it in unforeseen situations due to the dynamic nature of the problem. A software modeling system implementing all the necessary computational procedures has been developed; the results of computations are given.

I. INTRODUCTION

Over the past few years, the use of autonomous underwater vehicles (AUVs) has grown significantly, especially in such underwater works as seabed mapping, taking samples, searching and deactivating mines, etc. It is obvious, that application of coordinated and distributed groups of such robots may significantly improve the effectiveness of the large-scale and long-duration underwater operations. It is a problem of considerable practical interest to use AUVs group for continuous cooperative patrol of large surveillance areas in order to achieve the required level of coverage.

In this work, we propose to consider the problem of planning such a mission as the dynamic group routing problem with time windows. The challenge is to build effective routes for each vehicle taking into account all spatio-temporal constraints imposed by the specific nature of the water environment. Presence of the time-windows requirement here allows us to determine desired observation frequency for each region of a given area.

In recent years several researchers [1-3] proposed multi-agent solutions to a simpler version of the patrolling problem, considering that: 1) the territory can be represented by an undirected graph; 2) targets are fixed; 3) the patrolling task consists of continuously visiting all graph nodes in order to minimize the time lag between two visits [4]. Our formulation of the patrolling problem is an extension to these works. We admit that in case of real underwater environment targets and agents behave dynamically, making it necessary to constantly adjust group strategy in real time. To react rapidly to all unforeseen changes, group coordination is achieved by transferring data between members of the group through hydro-acoustic channel. Thus, it is an additional and important restriction to generate group routes ensuring possibility of regular communication sessions within the entire group under the limited range of communication channel. Furthermore, we assume that different sub-regions of the patrolled area could have different priority relative to each other, and that AUVs of the group may differ by their coursing speed.

The proposed problem has similarities both to the periodic vehicle routing problem (VRP) with time windows and to the team orienteering problem with time windows (TOPTW). In the latter, the goal is to maximize the total profit by a fixed number of routes such that the customers are visited within a time window and the maximum tour length is limited. The main difference between our problem and the TOPTW is that we do not have a fixed "profit" associated with each control points. Conversely, instead of profit, an agent collects penalties for non-well-timed inspections. Comparing to the former, all specialized VRP variations are limited by one common feature: each control point (customer) must be visited exactly once (sometimes, not more than once) within one concrete route. We had to abandon this restriction to successfully formulate the routing problem for continuous observation mission [5].

This paper proposes a formulation for dynamic patrol routing problem and provides an algorithmic scheme based on a hybrid evolutional approach to address it effectively. We combine the work of genetic algorithms with local search methods and specialized heuristics, which allows on-board generating of the group trajectories in real time. The rejection of using the prescribed cyclic trajectories allows group to patrol areas of complex shapes, to efficiently adapt to the dynamic environment, and makes it complicate to analyze the traffic patterns by a potential intruder, if any.

The reported study was partly funded by RSF (research project No. 16-11-00053) and by RFBR according to the research project No. 14-07-00740-a.

II. GRAPH REPRESENTATION OF AN AREA

In order to formulate the dynamic patrol routing problem as a combinatorial optimization problem, firstly a set of waypoints should be located in space in such way that their spherical neighborhoods, which corresponds to the size of the robots vision/observation field, would form a complete coverage of the protected perimeter. In assumption that technical equipment of an AUV allows it to observe equipotent in all directions, regular continuous inspections of all described waypoints would guarantee the effective perimeter patrolling.

It is a known open problem to cover an area of arbitrary shape by minimum amount of fixed radius circles. Fortunately, we do not need the exact solutions here, just an efficient ones, which can be obtained at very little computational cost.

We suggest overlapping the area map with regular hexagon tiling with sides of hex cells equal to the observation equipment's working radius (Fig. 1a). Then, we should exclude all hexes laying entirely inside of non-patrolled zones and obstacles (Fig. 1b). Those hexes among the remained, which centers also lie on unpassable regions (outlined on Fig. 1b), should be shifted in the direction of the nearest acceptable zone of the patrolled territory. It is a simplistic and sufficient way to generate rational coverage by a set of waypoints and their neighborhoods (Fig. 1c).

Thereafter, the given patrolled territory can be translated to an undirected graph where each vertex corresponds to a single waypoint. It leads us to the graph-based solution to the patrolling routing problem [6].

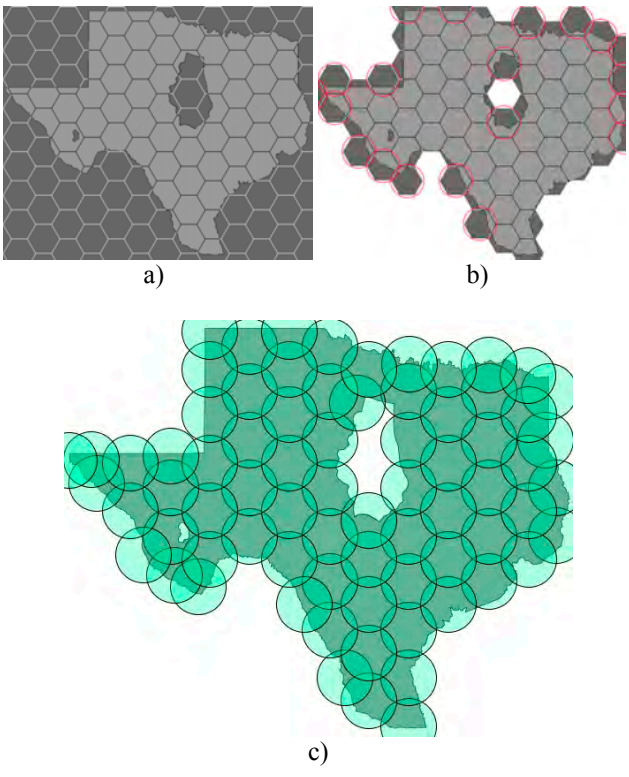


Figure 1. Procedure of setting up waypoints on the patrolled area

III. GENERAL MODEL

Dynamic patrolling missions require group of AUVs to visit and examine the set of waypoints at scheduled intervals. The problem of routing during such missions is to find a feasible group route ensuring, as far as possible, the well-timed inspection of the majority of targets. It is formally defined as follows.

Within the given water area there is a set of waypoints (targets) algorithmically located to cover all the patrolled territory. These targets are defined not only with their location in the space, but also with the demanded time interval for their periodical inspections. We denote set of targets with $N = \{1,...,\mathrm{n}\}$ where each target $i \in N$ receives its periodicity value p_i. It means that the duration of the time interval between its two successive inspections must not exceed p_i. In other words, each target should be inspected not less than once in its periodicity. We denote s as the time required to inspect each target by any vehicle. Additionally, for each target we define t_i' as the time since its last inspection. We suggest that at the beginning of the mission $t_i' = 0,\ i = 1,...,\mathrm{N}$.

The group of robots performing the mission consists of M functionally equivalent vehicles, which however may differ by their cruising speed v^k and by the range of hydro-acoustic communication channel $c^k, k = 1,...,M$. Additionally, we define the dummy nodes n_0^k for those AUVs, which initial positions does not fit in with any of the targets.

Let $V = N \cup \{n_0^k\}$ denote the set of all waypoints and AUVs states. We also let the $\varepsilon = \{(\mathrm{i},\mathrm{j}) : \mathrm{i},\mathrm{j} \in \mathrm{V}, \mathrm{i} \neq \mathrm{j}\}$ define the set of edges. The connected graph $G = (V, \varepsilon)$ represents the "roadmap" for the given area coverage. We denote the length of the shortest path from vertex i to j as $l_{ij} > 0,\ i, j \in V, i \neq j$. Thus, the shortest travel time for k–th AUV to move between two targets would be $t_{ij}^k = l_{ij} / v^k$.

The route of single vehicle $r = \langle V_1(\mathrm{r}), V_2(\mathrm{r}), ..., V_h(\mathrm{r}) \rangle$ is a list of vertices (waypoint's numbers) in the consecutive order of their planned visit. It should be noted that any vertex could be included more than once into the route of a single robot. One of the main characteristics of the route is its time duration $t(\mathrm{r})$:

$$r = \sum_{i=1}^{h} \left(t^k_{V_{i-1}(\mathrm{r})V_i(\mathrm{r})} + \mathrm{s} \right), \qquad (1)$$

where k is a number of performing AUV, $V_0(r) = \mathrm{n}_0^k$. The group route $R = \{r_1, ..., \mathrm{r}_M\}$ is a combination of routes of all single AUVs, $t(R) = \max_{i=1,...,M} (\mathrm{t}(\mathrm{r}_i))$.

The effectiveness of the patrolling is provided by maintaining the regularity of scheduled inspections. Situations, when AUV arrives too late and delays inspection of the waypoint, are undesirable and should be penalized via efficiency criteria. If it is physically

impossible to visit all waypoints in time with given robots group, two possible types of solutions should be considered:

- Routes guaranteeing total absence of delays by excluding some waypoints;
- Routes that provide minimal sum of delay time with obligatory visiting of all of the targets.

Before proceeding with our model development, we should note that the process of mission implementation is self-contained, i.e. all calculations are performed exclusively on AUVs board computer systems. Group coordination thereby is provided by transferring data between robots. Complete synchronization of actual data within the group could be achieved only if each pair of vehicles would be able to transfer data to each other directly or through other AUVs (Fig. 2). In what follows the group routes are called *communicatively stable*, if they guarantee the ability to synchronize data regularly.

Communication stability requirement arises due to the dynamic nature of underwater missions: firstly, the observation results may require changing parameters of some targets and even their number; secondly, the uncertainty of external environment may lead to unexpected changes in the status of working group. All these changes may occur in real time, making it necessary to adjust the current route (re-plan) in order to maximize the group efficiency in new conditions. Among the events that require route replanning are:

- The patrolled area expanding/contracting (adding new waypoints or removing existing ones);
- Change of periodicity p_i of the targets;
- Change of the AUVs group's composition.

Our objective is to construct the best communicatively stable patrol routes that provide the minimal total amount of delays during continuous and regular inspections of the targets up to the moment of the mission's end.

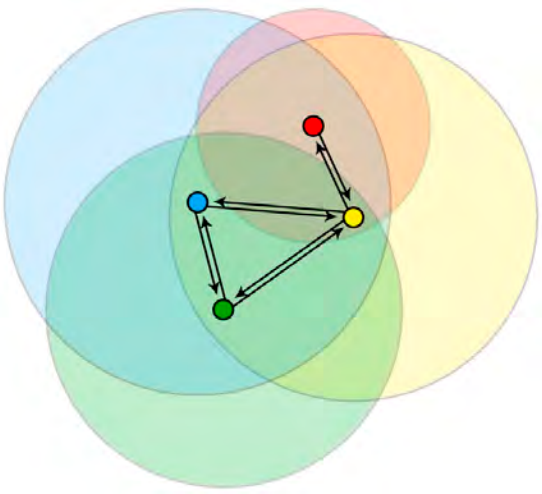


Figure 2. Building the communication graph for the heterogeneus group of AUVs

IV. PROBLEM STATEMENT

One of the main feature of the problem is uncertainty of the whole mission duration – the group should perform patrolling until it get some specific external command. This special aspect along with dynamic and unpredictable nature of the underwater mission leads us to the following decomposition approach to address these difficulties effectively: we suggest dividing the process of mission implementation on finite periods (planning period) with data synchronization within the group at the end of each period.

According to this scheme, each vehicle while following its pre-planned route on the current planning period computes to find the best feasible group route for the next period. Full data synchronization at the end of each period allows robots to receive and update all information obtained by the group, to reconfirm the current state of each unit of the group and to exchange its best-found solutions. The most effective group route among suggested is selected to become approved for the new planning period. The possibility of telecommunication session within the whole group is provided by communication stability restriction while constructing the group routes. This approach allows also to parallelize and distribute all calculations among the robots in a natural way.

Planning the route for a single period with determined and limited duration (horizon) allows us to construct the efficiency criterion for the group route in an explicit form to evaluate and compare the effectiveness of cooperative patrol. In order to do this, we define an additional function $a_i(\mathrm{t})$ corresponding to each target i of the mission that defines the "hotness" of corresponded target at specified moments of time:

$$a_i(\mathrm{t}) = \begin{cases} \bar{a}(t-(t_{ik}+\mathrm{s}))/\,p_i, t \in [t_{ik}+\mathrm{s}, t_{ik+1}] \\ 0, \qquad\qquad t \in [t_{ik}, t_{ik}+\mathrm{s}] \end{cases}, \quad \mathrm{k}=1,2,..., \tag{2}$$

where $t_{i1}, t_{i2}, ...$ is a sequence of moments, when target i is expected to be visited by AUVs according to the current route, and $\bar{a}$ is a constant value common for all targets. This way, the inspection of a target i resets its "hotness" to a zero value, following that the function (2) begins to increase and reaches threshold value $\bar{a}$ in period of p_i (Fig. 3).

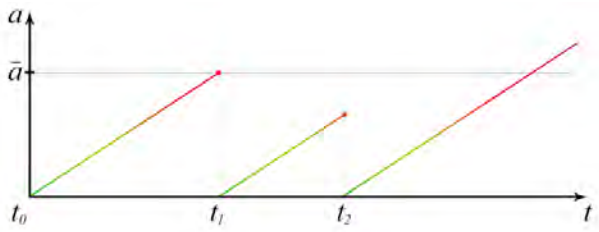


Figure 3. Graphical representation of the "hotness" function

Now we define a function to calculate a penalty for delaying the inspection of the target:

$$\varphi(a_i, \mathrm{t}) = \begin{cases} a_i(\mathrm{t}) - \overline{a}, a_i(\mathrm{t}) > \overline{a} \\ 0, \quad a_i(\mathrm{t}) \le \overline{a} \end{cases}. \tag{3}$$

Hence, the total penalty for the group route R is sum of penalties (3) for each inspections by all AUVs within the route:

$$\Phi(\mathrm{R}) = \sum_{i=1}^{M} \sum_{j=1}^{h} \varphi(a_m, \mathrm{t}_m), \tag{4}$$

where m stands for $V_j(\mathrm{r}_i)$ and t_m values represent the time moments of expected inspections of corresponded targets. The sequence of moments t_m for each target is calculated via (1).

The problem with function (4) is that it considers only those waypoints being included in the group route R. For this reason, it is needed to consider an additional function, which would "care" not about well-timed inspection like (4), but about the whole situation by the end of planning period:

$$\Psi(R) = \sum_{i=1}^{N} \mathrm{a}_i(t_0 + t(R)), \tag{5}$$

The function (5) delivers some sort of forecasting value for the next planning period. Using (5) as a part of ultimate criteria would allow us to provide the group routes with two positive features instantly: it encourages group to inspect regularly each waypoint of the mission and indirectly normalizes durations of routes (1) of all vehicles in the group.

Hence, the final efficiency criterion for the group route is as follows:

$$f(R) = \omega \cdot \Phi(R) + \Psi(R), R \in \mathrm{Z}, \tag{6}$$

where Z is the set of all possible communicatively stable routes and ω is the weight number that allows one to expertly choose the operating mode of the group in case of impossibility of well-timed inspection of all waypoints in the mission:

- With $\omega \cong 1$ the group will always examine all the mission waypoints but with some delays.
- With $\omega >> 1$ the group will continue to inspect only those targets, which can be visited strictly in time and the remaining waypoints would be ignored.

It is also should be noted that the function (2) is constructed in such way, that "hotness" of targets with lesser periodicity grows faster. In that way, if the first operating mode ($\omega \cong 1$) is chosen in (6), delays of inspections would happen preferably with waypoints of biggest periodicity.

Now that we have the final criterion in the explicit form, we can construct and compare different group routes on the graph G in terms of effective patrol.

V. HYBRID EVOLUTIONARY APPROACH

For a broad combinatorial class of vehicle routing problems there are no algorithms solving it in polynomial time, which leads us to the class of approximation algorithms that allow obtaining rational sub-optimal solutions in low computational time. In fact, due to the high dynamic of underwater environment we do not really need an exact solution since vehicles world hardly be able to execute it thoroughly.

A comprehensive survey of various approaches for solving the VRP concludes that evolutionary algorithms (EAs) generally outperform any other heuristic or metaheuristic [7]. Its main advantage is ability to find solutions for poorly structured problems and problems with complex constraints, as EAs require a relatively small amount of information about the nature of the problem.

We propose a hybrid evolutionary approach featuring specialized genetic operators, local search heuristics and some other schemes of solutions improvement to address the given constraints and expectable large-size of the problem. Another layer of complexity is implied by a "bad" neighborhood structure of the described patrol routing problem, making it difficult to allocate and find qualitative and feasible solutions, as they may not be in the neighborhood of other feasible high-quality solutions in the search space.

A. Initial Population

The construction of the initial population of solutions is the first and crucial step to achieve good rational solutions rapidly. The aim of this step is to construct a set of chromosomes ensuring both covering a significant portion of the search space and containing a variety of good solutions. This requirement is achieved by the simultaneous use of three different construction heuristics in ratio of 40/40/20: two parallel insertions and a random sequential insertion. Both parallel insertion heuristic are implementations of greedy randomized constructions – one is spacial and another is temporal. The random insertion is to bring the diversity of genetic material into the population. The constructed population is evaluated by the objective function (6). According to the results of ranking, the tournament selection chooses a set of solutions for procreation and mutations.

B. Crossover & Mutation

We suggest using a number of specialized genetic operators that has shown themselves to be the most effective in the computational experiments: two different variants of crossover and a multimode mutation. The first crossover is a heuristic modification of the two-point crossover but with additional rule to choose second point only within the route of the same vehicle as the first point. The second crossover operator seeks to combine the general characteristics of the parental individuals with the qualities of the best-known solution. The multimode mutations consists of a five operators: add new vertex into the route, delete a vertex from the route, change a vertex, swap two vertices, insert new vertex from the neighborhood of another vertex from the route.

C. *The Local Search and Replacement*

In order to further improve our newly received individuals the modifications of deterministic heuristic search techniques of 2-opt exchange and λ-interchange are applied periodically to the whole population for a deeper exploitation.

The mechanisms of parallel populations with immigration (island model), clone removal and elitism provide faster algorithm convergence rate while preventing premature convergence to local optima.

D. *Bringing Solutions to Feasibility*

To ensure overall communication stability of the offspring solutions, each constructed route is iteratively modified according to the following procedure until requirements are met:

- The graph formed by the final positions of AUVs $V_h(\mathrm{r}_i), i = 1,\ldots,M$ is checked for connectivity only on those arcs whose length does not exceed the range of communication channels of AUVs corresponded to both vertices.
- If the graph is disconnected, the center of gravity g of all AUVs final positions is calculated.
- The most distant vertex from g is defined and replaced in the route of corresponded vehicle by another target, that is closest to g.

This procedure (Fig. 4) guarantees that the communication stability requirement for the current group route would be achieved in no more than M iterations.

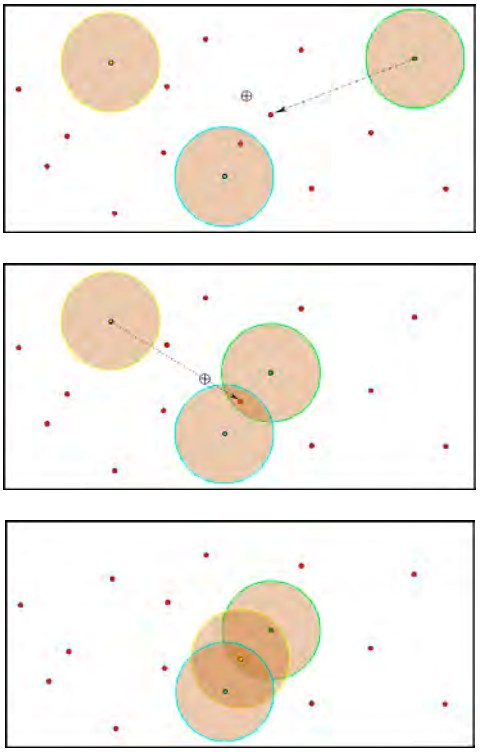
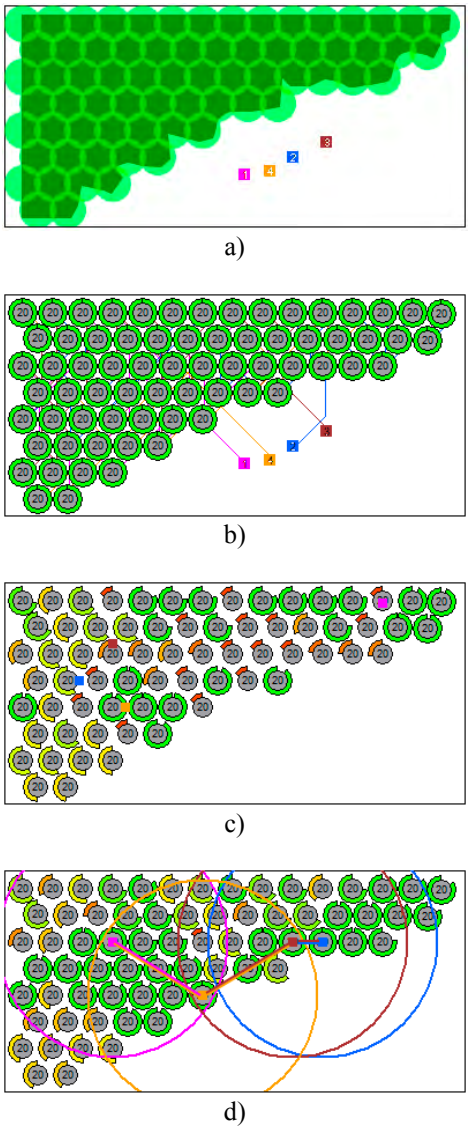


Figure 5. The procedure of bringing routes to feasibility

VI. RESULTS OF TEST CALCULATIONS

On the basis of a series of simulation experiments using the developed framework AUV Mission Planner the best initial parameters of the algorithm were found to guarantee its efficiency on various types of test problems.

In the following example (Fig. 5), we demonstrate how the algorithm manages to generate the group route for patrolling the given area with four AUVs. Firstly, the coverage on the territory by circles of fixed radius has been fulfilled (Fig. 5a). Based on the coverage the set of 69 waypoints has been determined (Fig. 5b) with periodicities of 1200 time units, while planning horizon is 3000 time units (more than 150 scheduled visits of waypoints). Within computing time of 2 minutes, the algorithm had constructed the group route providing well-timed inspection of the majority of targets with a total delay of 50 time units while optimal solution provides 14. Fig. 5c shows an AUV Mission Planner screenshot during the simulation of the group movement. The route is built to ensure the possibility of data synchronization within the group at the end of the planning period (Fig. 5d).



a)

b)

c)

d)

Figure 4. An example of the patrol mission planning

VII. CONLUSION

The dynamic routing problem model described in the paper allows formulating the patrol missions for the group of robots as a separate subclass of VRP. The suggested coverage approach provides us with the graph representation of the patrolling problem that could be then solved as a combinatorial problem using evolutionary algorithm of the proposed design. Decomposition scheme of the planning periods allows group both to reduce computational complexity of the whole mission-planning problem and to react rapidly to all unforeseen changes. Due to the constructed efficiency criterion, the algorithm searches for routes not only ensuring the patrolling efficiency on the current period, but also estimates its impact on the further situation.

We have expanded our simulation framework AUV Mission Planner to test and work out newly developed patrolling approach for groups of heterogeneous autonomous robots with different types of vehicles, targets, and restrictions. The high efficiency of the suggested approach is shown through simulation studies.

The algorithm is proved to be fast, accurate and reliable in group routing under strict constraints. It allows fast construction of reasonable and feasible solutions and their further modification to achieve fast convergence to satisfying conditions.

REFERENCES

[1] A. Machado, G, Ramalho, J-D. Zucker, and A. Drogoul, "An empirical analysis of alternative Architectures," In Proceedings of the 3rd International Workshop on Multi-Agent Based Simulation, pp. 155–170, 2002.

[2] Y. Chevaleyre, "Theoretical analysis of the Multiagent patrolling problem," In Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology, pp. 302-308, 2004.

[3] H. Santana, G. Ramalho, V. Corruble, and B. Ratitch, "Multi-agent patrolling with reinforcement learning," In Proceedings of the Third International Conference on Autonomous Agents and Multi-agent Systems, pp. 1122-1129, 2004.

[4] J. Perron, J, Hogan, B. Moulin, J. Berger, and M. Belanger, "A hybrid approach based on multi-agent geosimulation and reinforcement learning to solve a UAV patrolling problem," Proceedings of the 2008 Winter Simulation Conference, pp. 1259-1267, 2008.

[5] M.Yu. Kenzin, I.V. Bychkov, and N.N. Maksimkin, "Hybrid evolutionary approach to multi-objective mission planning for group of underwater robots", Communications in Computer and Information Science, Vol. 549: Mathematical Modeling of Technological Processes, pp. 73-84, 2015.

[6] B.B. Keskin, S. Li, D. Steil, and S. Spiller, "Analysis of an integrated maximum covering and patrol routing problem," In: Transportation research Part E: Logistics, Vol. 48, pp. 215-232, 2012.

[7] O. Braysy, and M. Gendreau, "Vehicle Routing Problem with Time Windows, Part I: Route Construction and Local Search Algorithms," Transportation science, Vol. 39, No.1, pp. 104–118, 2005.

Multi - heater induction heating system with sandwich material heater

Albin Smrke*

* MSX d.o.o., Matjaževa cesta 17 ,1000 Ljubljana, Slovenija info@msx-technology.com

Abstract

Induction heater is often used in process of cooking in the household. Advantages is very quick and energy efficient heating of food. In standard cooking appliances heater are placed below the glass surface and they are energy controlled with electronic regulation. Each induction coil is powered with one power board. The power is normal around 2000 W. In case of 4 heaters the max Power is 8000 W. Power regulation is not precise and user can very rear find the right level of energy in different processes of cooking. It was a need to reduce max power and regulation. A new optimal multi heater system is developed. On one power board are connected more induction coils and new electronic board is developed with temperature control. Multi heater system is presented in this article, together with some results of applied multi heater system in real cooking. Glass surface of cooker is replaced with new sandwich material Stainless steel – aluminium – stainless steel. Temperature sensor are developed with hybrid technology and printed on the surface. A new step forward in induction cooker is integration of hybrid heater with power max 100W and temperature max 125 ^{0}C is implemented on top of induction coil. Developed is Control algorithm based on needed cooking process, where very little power is needed in process of cooking after food is heated to boil. Such algorithm enables the use of cooker with maximum power 2000 W for cooker with 4 coils and very precise automatic cooking control. Software and hardware are explained in article. Results of test are done with thermodynamically analysis of real cooking process.

I.INTRODUCTION

Induction heater is often used in process of cooking in the household because of quick and energy efficient preparing of food. Standard MSX single induction cooker E/2 is on Picture 1. E/2 cooking is a way of smart automatic cooking by utilizing various sensors installed on pots. Sensors connect via wireless radiofrequency (RF) link to E/2 plug-in module mounted inside stove. E/2 plug-in uses received sensor data to implement automatic and energy efficient cooking process.
E/2 plug-in module Figure 2 is designed for enabling E/2 cooking on regular stoves. In principle, E/2 enabled stove is still completely regular stove and can be used normally for manual cooking. E/2 plug-in adds the possibility of automatic cooking by usage of remote RF sensors when user chooses E/2 mode on a specific cooking zone.

Wireless sensors work on our proprietary protocol E/2.



Figure 1: Midea E/2 induction cooker with lid sensor

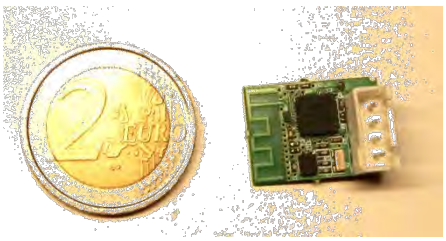


Figure 2: E/2 plug-in module

E/2 plug-in embedding requires minimal hardware modifications of existing stove:

- Power to E/2 module should be provided.
- Communication channel between main stove's microcontroller and E/2 plug-in must be implemented.
- Depending on existing user interface, some additions to user interface may be required for E/2 selection and indication.

In addition, stove's controller firmware must enable following to E/2 plug-in:

- Indirect access to user interface to support added E/2 functionality.
- Indirect access to heater must be granted to E/2 plug-in when needed.

Main controller is always in control of the heaters and user interface and still enables normal manual cooking. E/2 accesses user interface and heaters indirectly through main controller when E/2 smart cooking is in use on a specific cooking zone. Figure 3 shows the situation.

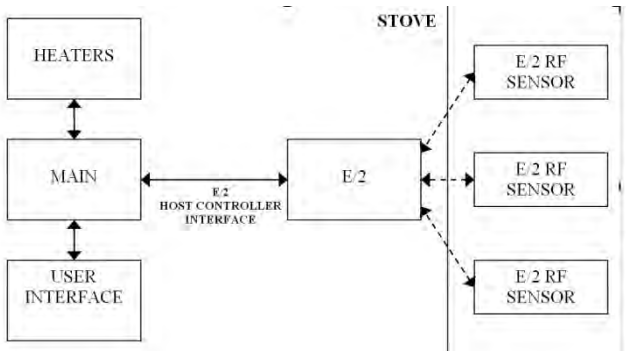


Figure 3: Stove with E/2 plug-in

E/2 interface is universal for different types of stoves with different user interfaces and supports single cooking zone stoves and multiple cooking zones stoves.

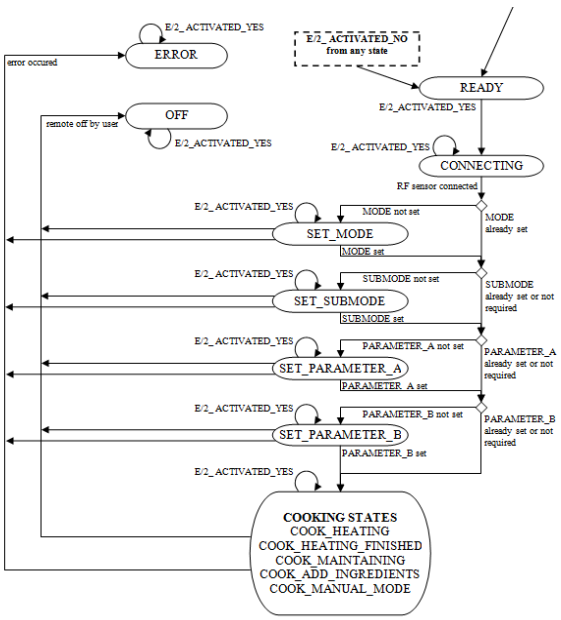


Figure 4: Regular Power board for induction heater

Induction heater is placed below the glass surface. Each induction coil Figure 5 is powered with one power board Figure 6. We developed cooking algorithms for E/2 Midea with very precise smart energy control. The coil specification is in Table 1 and power specification in Table 2. Power regulation is not precise and user can very rear use the right level of energy in different processes of cooking.

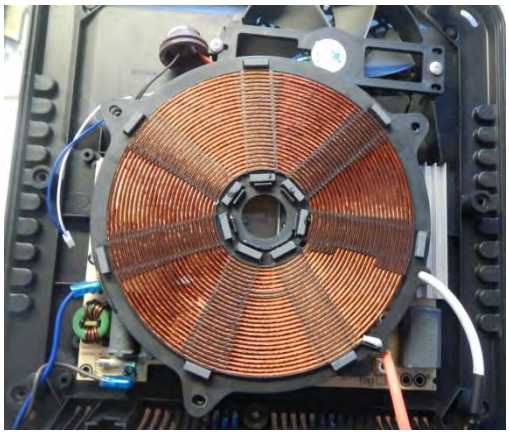


Figure 5: Regular induction heater in a induction cooker

coil specification	
lead wire longth (mm)	inner 95mm/outer 85mm
wire number	26 whires
Material	copper wire
wire diameter (mm)	0.31mm
coil type	single layer
coil diameter	155mm
plastic disc diameter	177mm
coil laps	18 laps
coil inductance (uH)	55±2uH

Table 1: Coil specification

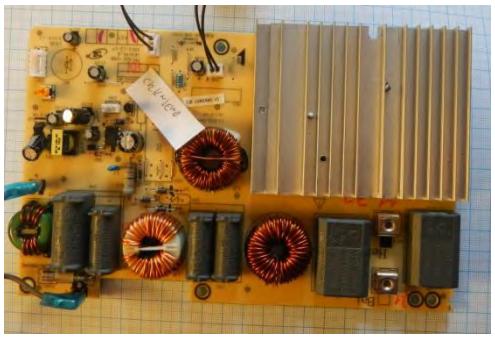


Figure 6: Regular Power board for induction heater

Power level	W
1	120
2	300
3	400
4	600
5	800
6	1000
7	1100
8	1500
9	1800

Table 2: Power settings on power board

Average Pulse power:
120W （2S/10S), 300W （3S/10S), 400W （4S/10S), 600W（6S/10S), 800W（8S/10S), 900W（9S/10S)

Minimum continues power is 1000W at level 7. This system is not good for cooking and this is the main reason that we developed E/2 MSX smart sensor system for cooking that is also on market under Iqcook trade name from Gorenje [1].

It was a need to reduce max power and regulation. A new optimal multi heater system is developed.

II. MULTI HEATER SYSTEM

The Idea was to connect on one power board more induction coils. A new electronic board is developed with multi-heater control algorithms. Principle of idea is shown in schematic on Figure 7. We used relay system with OFF POWER status of heating in moment of changing heater.

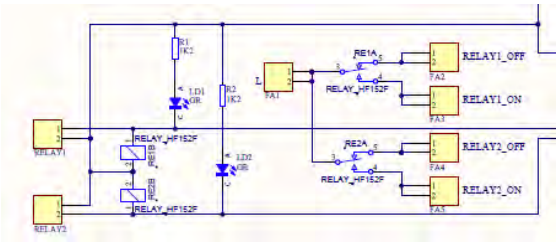


Figure 7: Schematic solution for multi-heater



Figure 8: Multi-heater interface

For testing and changing parameters of heating we developed new software that is shown on Figure 9

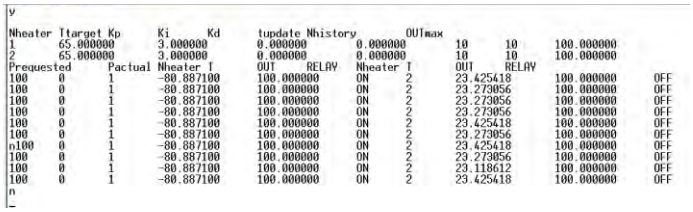


Figure 9 Input –output Form

Two main parameter are:

- heat_Tlimit – temperature which the system touches when switched on. If the measured value is less than this temperature always turns heat heating program that has parameters of heat PID [5] (heat_Ttarget , heat_Kp , heat_Ki , heat_Kd , heat_OUTmax)

- heat_dTlimit - a parameter that takes into account the speed of the cooling plates in a given time (parameters of heat -t period) . If the value is smaller, as a result of poorer heat conduction in a situation where food is already heated to a certain degree, it will go to the maintenance system warm with parameters.

- heatOUTmeasure – power requested from 0-100%

4. NEW SANDWICH MATERIAL SS-AL-SS

Glass surface of cooker is replaced with new sandwich material Stainless steel – aluminium – stainless steel.

a) S/S 304 + AL + S/S 430
Thickness: 2.0mm - 3.0mm S/S 430
(Induction Compatible Stainless Steel)
AL (High Thermal Conductivity Aluminium)

For optimisation of induction heating system the metal plate is very important. We perform a research on different possible materials.

Important parameter is temperature but speed must be taken into account.

For induction heating heater devices works extremely responsive and the temperature is difficult to be held in needet range. Metal sandwich panel with aluminum make the regulation possible with extremely precise regulation. Figure 10 show sandwich materials with same parameter than shown in Figure 11 for standard iron 3mm plate.

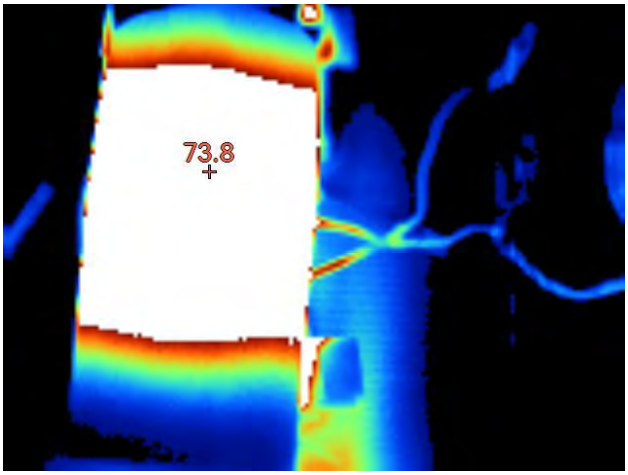


Figure 10: Thermal picture of sandwich materials

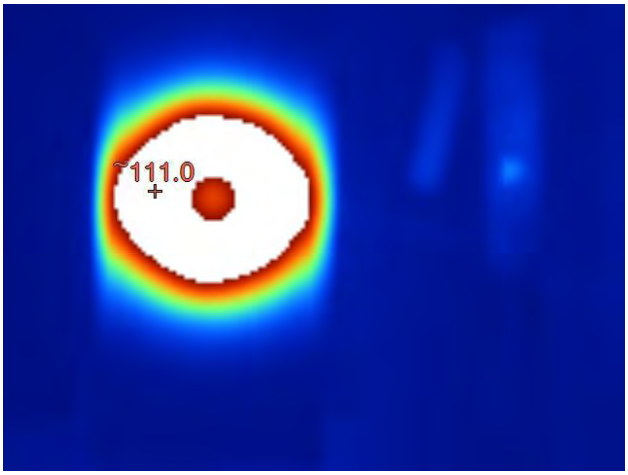


Figure 11: Thermal picture of iron materials

Heating control allows precise setting of the desired temperature in the food and allows optimum supply of energy.

In terms of healthy food preparation is accurate temperature control very important due to destruction of vitamins, proteins and other valuable food ingredients, which are in principle sensitive to temperature. Another problem arises due to high temperatures in the heated fat, roasting meat and the like, where the fat disintegrate and form harmful ingredients that we consume.

Impulse response of the control system is relatively short, and it clearly indicates towards a very precise setting for desired temperature of food. This is good because the probe is not attached directly on the food it is on back side of metal plate.

Results of measurement direct an food and on probe are shown in Figure 12.

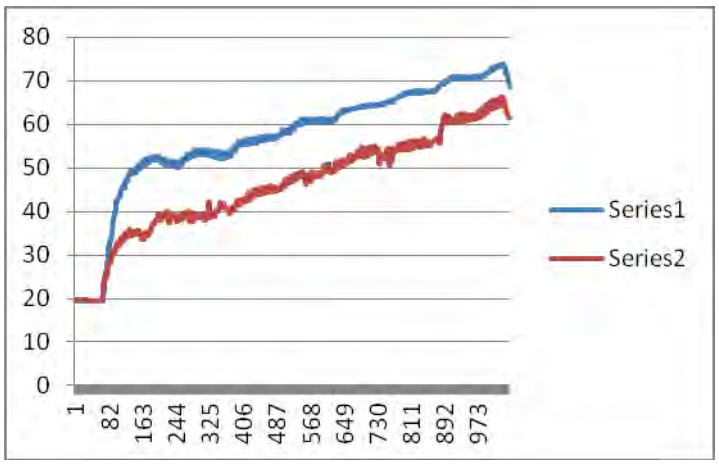


Figure 12: Real experiment on food and on probe below sandwich

Series 1 is probe and series 2 is food.

5. NEW LOW POWER HEATER AND SENSOR

For understanding need of development low power heater shown in Figure 13 on top of powerful induction heater, we will give short description of today's problem that is in all state of the art cooking appliances.

Heat and temperature are the most misunderstanding terms in laboratory of producer of hob. Amount of energy that is need to cook from science point depends on: Specific heat of food and means heat needed to heat a 1 kg of food for 1° - latent heat of vaporisation of water has even bigger influence. Heat flow depends on the temperature levels and many other factors. Transfer of energy can be in different thermal processes like conduction, convection, radiation and also the phase

change by boiling, simmering, melting are not so simple. Cooking time is night mare for scientific oriented cooks. Cooking time is proportional to the square of radius of food stuff and inversely proportional to square of the difference in the temperature.



Figure 13: Induction heater with low power hybrid heater

Food heats up and we say that the food is cooked when it has been on higher temperature for a specified time period. Usually cookery books are imprecise when talking about the heat and the temperature (low heat, middle level, high temperature...) [8]

Important factor for user: No over boiling or burning, using only needed energy, helping cook to save time.

E/2 sensors with automated control, optimize the use of the minimum energy required for each cooking process. Complete optimization of the energy enables savings of up to 50% and when the cooking process is longer even greater savings. With steam cooking the savings are up to 80%.

Food will not be overcooked or burned. E/2 sensors will recognize the boiling point and with automatic process regulate the energy so that cooking will be without unpleasant steam.

All cooking modes and aloso for cooking with large amount of water will bring the contents of the pot to the boil and then maintain the boiling temperature through communication with the E/2 sensors. Normal used Power in heatre today are from 2 – 3 kW. The main point is that for heating we need 100% Power and for maintaining only few %. Induction heater have a big problem in range below 50% Power.

Temperature heater is developed with hybrid technology and printed on the surface of ceramic. A new step forward in induction cooker is done with integration of hybrid heater with power max 100W and temperature max 125 ^{0}C is implemented on top of induction coil.

Developed is also control algorithm based on needed cooking process, where very little power is needed in process of cooking after food is heated to boil. Such algorithm enables the use of cooker with maximum power 2000 W for cooker with 4 coils and very precise automatic cooking control.

6. EMISSION COMPLIANCE AND MAGNETIC MEASUREMENT RESULTS AND DISCUSION

Industry is very sensitive for every change in construction of appliance for cooking. For this reason we did all test [7]. Principle is shown on Figure 13.

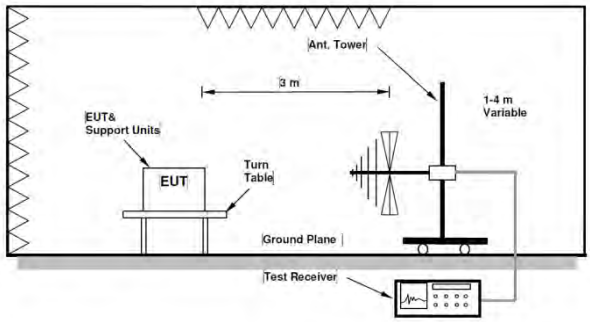


Figure 13: Measuring method for emission compliance

According to EN 55014-1:2006/A1:2009 measuring method ″b″ was used to show emission compliance in frequency range from 30 MHz to 1000 MHz

One of the result is depicted on Figure 14.

RADIATED EMISSION - MAGNETIC FIELD MSX MULTIHEATER Uin 230V/50Hz, MAX POWER, BLUE CURVE - 20cm WMF pot; BLACK CURVE - 24cm Vzepter pot. VERTICAL COMPONENT

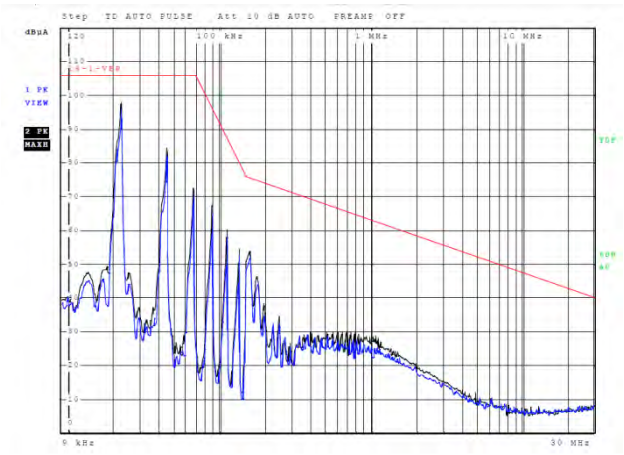


Figure 14: Vertical magnetic component for 20cm and 24cm pot

7. CONCLUSION

Climate change is now a recognised global issue, 'energy efficiency' is no longer just a fashionable phrase, but has become a serious consideration for governments, businesses and consumers alike. Even though the majority of domestic appliances, such as washing machines, dishwashers and refrigerators have all achieved through development some improvement in energy efficiency, there remains one very energy-intensive appliance which, to date, has had few alternatives: the cooker.

Now however, thanks to E/2 a revolutionary technology that can be installed in new cooker hobs, [electric, induction and gas] the energy saving hob has become a reality.

- Delivering a saving of up to 50% depending on the type of food being cooked, E/2 was developed following the discovery that significant energy savings can be achieved. Data gathered from the E/2 sensor is then run through algorithms in the software of the hob controller which regulates precisely the amount of energy required.

- The new E/2 technology also has another significant practical advantage: as the controller automatically regulates the energy, the cooker can be left unattended during the cooking process without fear of overcooking or burning the food.

- E/2 (sometimes E/2 MSX technology) is an interoperability function featuring a wireless communication protocol between cooking appliances and cookware.

- E/2 is proprietary open access wireless cooking sensor network technology.

- E/2 is a wireless multisensor technology and comes with protocol software universal for stoves with all types of energy (induction, gas, electric) and all possible accessories like universal sensors with E/2 add-in, plug-in for healthy steam cooking, boiling control, pressure cooking, sous vide cooking, wok and all other.

- E/2 proprietary wireless communication guarantees complete automatically controlled hob operation, without the presence of user.

- E/2 makes real automatic cooking possible with closed loop regulation.
- It encourages interoperability and open-access to data between manufacturers of cooking hobs and manufacturers of pots (stainless still, enamel, aluminium.....).

- Low power heater for maintaining of cooking can be incorporated into the induction cooking hob with very low cost of production (less than 2 EUR).

- Sandwich plate instead of glass makes possible to people to use every pot not only special expensive for induction. This makes also big price difference. Price of glass is 10 times of sandwich metal .

REFERENCES

[1] http://www.gorenje.si/izdelki/veliki-gospodinjski-aparati/kuhanje/lastnosti/kuhalisca/iqcook

[2] Bejan, A., et.all, *Thermal Design and Optimization*; John Wiley & Sons, New York, 1995.

[3] Čudina, M., *Tehnična akustika*.; FS Ljubljana, 2001.

[4] Kuščer, I.,: *Termodinamika*; Društvo mat, fizikov in astronomov, Ljubljana, 1974 Allyn and Bacon, Inc, Boston, 1986

[5] Prelec, Z., *Energetika u procesnoj industriji*; Školska knjiga, Zagreb,1994.

[6] Smrke, A., *US Patent 5,951,900, EP 0820573, RUS 97118228, JPN504701, SLO 9600097, SLO 9500106 , Kitajska HM 971678,* 1995-2002

[7] IEC 60335-2-75 Issue:2009/04/29 Ed: 2.2 Household and Similar Electrical Appliances - Safety - Part 2-75: Particular Requirements for Commercial Appliances

[8] http://cooking-sensor.com/E-2%20brochure.html

Two-Rate Motion Control of VTAV by NARMA-L2 Controller for Enhanced Situational Awareness

I. Astrov*
* Tallinn University of Technology/Department of Computer Control, Tallinn, Estonia
igor.astrov@ttu.ee

Abstract – This paper focuses on a critical component of the situational awareness (SA), the motion control of autonomous vertical flight for vectored thrust aerial vehicle (VTAV). With the SA strategy, we proposed a two-rate flight control procedure using two autonomous control subsystems to address the dynamics variation and performance requirement difference in initial and final stages of flight trajectory for a nontrivial VTAV type mini-aircraft model. This control strategy with using of NARMA-L2 neurocontroller for chosen model of VTAV has been verified by simulation of take-off and forward maneuvers using software package Simulink and demonstrated good performance for fast stabilization of motors, consequently, fast SA with economy in energy can be asserted during search-and-rescue operations.

I. INTRODUCTION

Situation awareness has been formally defined as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" [1]. As the term implies, situation awareness refers to awareness of the situation. Grammatically, situational awareness (SA) refers to awareness that only happens sometimes in certain situations.

SA has been recognized as a critical, yet often elusive, foundation for successful decision-making across a broad range of complex and dynamic systems, including emergency response and military command and control operations [2].

The term SA have become commonplace for the doctrine and tactics, and techniques in the U.S. Army [3]. SA is defined as "the ability to maintain a constant, clear mental picture of relevant information and the tactical situation including friendly and threat situations as well as terrain". SA allows leaders to avoid surprise, make rapid decisions, and choose when and where to conduct engagements, and achieve decisive outcomes.

In [4], a novel design of a multi-rotor wing platform which has great potential for both military and civilian application areas is presented. These platforms can provide the greater SA in the urban terrain.

The vectored thrust aerial vehicle (VTAV) is one of the key tools to gather the information to build SA for all leaders. The VTAV is the ground maneuver commander's primary day and night system. The VTAV provides the commander with a number of capabilities including:

- Enhanced SA.
- Target acquisition.
- Battle damage assessment.
- Enhanced battle management capabilities (friendly situation and battlefield visualization).

The combination of these benefits contributes to the commander's dominant SA allowing him to shape the battlefield to ensure mission success and to maneuver to points of positional advantage with speed and precision to conduct decisive operations. Some conditions for conducting aerial reconnaissance with VTAVs are as follows.

- Time is limited or information is required quickly.
- Detailed reconnaissance is not required.
- Extended duration surveillance is not required.
- Target is at extended range.
- Threat conditions are known; also the risk to ground assets is high.
- Verification of a target is needed.
- Terrain restricts approach by ground units.

The VTAV offers many advantages, including low cost, the ability to fly within a narrow space where the objects in the proximity of the vehicle are protected from the propeller and vice versa, and the unique hovering and vertical take-off and landing (VTOL) flying characteristics.

Autonomous vertical flight is a challenging but important task for VTAVs to achieve high level of autonomy under adverse conditions. The fundamental requirement for vertical flight is the knowledge of the height above the ground, and a properly designed controller to govern the process.

In [5], a three stage flight control procedure using three autonomous control subsystems for a nontrivial nonlinear helicopter model on the basis of equations of

vertical motion for the center of mass of helicopter was proposed. The proposed control strategy has been verified by simulation of hovering maneuvers using software package Simulink and demonstrated good performance for fast SA.

This paper concentrates on issues related to the area of [5], but demonstrates another field for application of these ideas, i.e., research technique using control system modeling and simulation on the basis of state-space equations of motion of VTAV.

In this paper our research results in the study of vertical flight (take-off and forward cases) control of VTAV which make such SA task scenario as "go-search-find-return" possible are presented.

The contribution of the paper is twofold: to develop new schemes appropriate for SA enhancement using VTAVs by control of vertical flight, and to present the results of flight maneuvers for chosen model of VTAV for fast SA in simulation form using the MATLAB/Simulink environment.

II. VTAV MODEL

In [6], a model of VTAV was proposed. The VTAV which has a triangular shape consists of three ducted fans. The axis of the front ducted fan is fixed perpendicular to the plane of the triangle, whilst the rear two fans are able to rotate around an axis which is common to both fans. The independent vectoring of these two fans allows the yaw motion and forward flight with minimal or no pitch movement. This VTAV has a weight $5.5kg$ and a distance $0.231m$ from the center of gravity to all the fans.

The dynamic model of this VTAV demonstrates unstable equilibrium and is significantly simpler than the dynamic model of a helicopter. The fact that the system is in unstable equilibrium also allows us to develop control systems that will provide greater maneuverability than would have been possible with quad-rotor systems or co-axial helicopters. Unlike single ducted fan systems, this design provides ample space for a sizable payload bay. Minimal pitch and roll movements ensure that the VTAV is suitable to carry vision sensors and laser range sensors for terrain mapping where the attitude of the sensors play a major role on the reliability of acquired data.

The dynamic model for control yields the general form of state equations for the VTAV [6]

$$\dot{x}(\tau) = Ax(\tau) + Bu(\tau) + v(\tau) \tag{1}$$

$$y(\tau) = Cx(\tau) + w(\tau) \tag{2}$$

where $x(\tau), u(\tau), y(\tau), v(\tau), w(\tau)$ are the state, control input, output, process noise and measurement noise vectors, respectively.

The variables of this model are:

$$x^T = (\dot{x}_c,\ \dot{y}_c,\ \dot{z}_c,\ \phi,\ \theta,\ \psi,\ \omega_x,\ \omega_y,\ \omega_z),$$

$$u^T = (\omega_{f_1},\ \omega_{f_2},\ \omega_{f_3},\ \theta_{t_2},\ \theta_{t_3}) \tag{3}$$

where x_c, y_c, z_c are coordinates of center of mass of VTAV in the earth-frame, ϕ, θ, ψ are roll, pitch and yaw angles; $\omega_x, \omega_y, \omega_z$ are roll, pitch and yaw rates; $\omega_{f_1}, \omega_{f_2}, \omega_{f_3}$ are angular speeds of the first, second, and third motors; $\theta_{t_2}, \theta_{t_3}$ are vectoring angles of second and third fans.

The matrix structure of A, B, C for the state-space model of system (1)-(2) is given by

$$A = \begin{bmatrix} a_{11} & 0 & 0 & 0 & a_{15} & 0 & 0 & 0 & 0 \\ 0 & a_{22} & 0 & a_{24} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & a_{72} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ a_{81} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$B = \begin{bmatrix} 0 & 0 & 0 & b_{14} & b_{15} \\ 0 & 0 & 0 & 0 & 0 \\ b_{31} & b_{32} & b_{33} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & b_{72} & b_{73} & 0 & 0 \\ b_{81} & b_{82} & b_{83} & 0 & 0 \\ 0 & 0 & 0 & b_{94} & b_{95} \end{bmatrix}, C = I_9 \tag{4}$$

The parameters a_{11} through a_{81} and parameters b_{14} through b_{95} in (4) are given by:

$a_{11} = -0.0052$, $a_{15} = -9.8100$, $a_{22} = -0.0052$, $a_{24} = 9.8100$, $a_{72} = -0.0311$, $a_{81} = 0.0056$;

$b_{14} = -3.2700$, $b_{15} = -3.2700$, $b_{31} = -0.3448$,
$b_{32} = -0.3448$, $b_{33} = -0.3448$, $b_{72} = -16.5640$,
$b_{73} = 16.5640$, $b_{81} = 3.4254$, $b_{82} = -1.7127$,
$b_{83} = -1.7127$, $b_{94} = 39.2257$, $b_{95} = -39.2257$.

Then, we have

$$x(\tau) = \int_0^\tau \dot{x}(t)dt, y(\tau) = \int_0^\tau \dot{y}(t)dt, z(\tau) = \int_0^\tau \dot{z}(t)dt, \tag{5}$$

where

$$x_c(0) = 0, y_c(0) = 0, z_c(0) = 0.$$

From (1)-(5) we can see that the attitude vector $(x_c, y_c, z_c)^T$ for given model of VTAV can be computed.

III. TWO-RATE SUBSYSTEMS

In [7], the approach to design of decomposed multirate stochastic linear systems, which consist of naturally grouped entrance and target signals that are caused by their characteristic frequencies, is offered.

Setting $z(\tau) = Tx(\tau)$, where T is a nonsingular matrix, we see that (1)-(2) are transformed into the equations

$$\dot{z}_1(\tau) = \Lambda_1 z_1(\tau) + B_1 u(\tau) + T_1 v(\tau) \tag{6}$$

$$\dot{z}_2(\tau) = \Lambda_2 z_2(\tau) + B_2 u(\tau) + T_2 v(\tau) \tag{7}$$

$$y(\tau) = C_1 z_1(\tau) + C_2 z_2(\tau) + w(\tau) \tag{8}$$

$$|\lambda(\Lambda_1)| \langle \gamma_{s_{\max}}, |\lambda(\Lambda_2)| \rangle \gamma_{f_{\min}} .$$

Definition 1: A function with a large derivative, which is quickly decreasing, is said to be the "fast" function, a function with a small derivative, which is slowly decreasing, is said to be the "slow" function.

Consider the first time interval $0 \langle \tau_f \le \tau_{fs}$. According to Definition 1, the variable z_1 can be considered as a "slow" function on this interval. Hence, assuming that $\dot{z}_1(\tau_f) = 0$, from (6), we see that

$$z_1(\tau_f) = -\Lambda_1^{-1} B_1 u(\tau_f) - \Lambda_1^{-1} T_1 v(\tau_f) . \tag{9}$$

From (7)-(9), we find that the state equations for a "fast" subsystem may be written as

$$\dot{z}_f(\tau_f) = A_f z_f(\tau_f) + B_f u_f(\tau_f) + T_f v_f(\tau_f) \tag{10}$$

$$y_f(\tau_f) = C_f z_f(\tau_f) + D_f u_f(\tau_f) + w_f(\tau_f) \tag{11}$$

$$A_f = \Lambda_2, B_f = B_2, T_f = T_2, C_f = C_2,$$

$$D_f = -C_1 \Lambda_1^{-1} B_1, V_f = -C_1 \Lambda_1^{-1} T_1,$$

$$z_f(\tau_f) = z_2(\tau_f), u_f(\tau_f) = u(\tau_f), v_f(\tau_f) = v(\tau_f),$$

$$y_f(\tau_f) = y(\tau_f), w_f(\tau_f) = w(\tau_f) + V_f v(\tau_f) .$$

Consider the second time interval $\tau_s \rangle \tau_{fs}$. According to Definition 1, the variable z_2 can be considered as a "fast" function of time, achieving on this interval a steady meaning. Hence, assuming that $\dot{z}_2(\tau_s) = 0$, from (7), we find

$$z_2(\tau_s) = -\Lambda_2^{-1} B_2 u(\tau_s) - \Lambda_2^{-1} T_2 v(\tau_s) . \tag{12}$$

From (6), (8) and (12), we find that the state equations for a "slow" subsystem may be written as

$$\dot{z}_s(\tau_s) = A_s z_s(\tau_s) + B_s u_s(\tau_s) + T_s v_s(\tau_s) \tag{13}$$

$$y_s(\tau_s) = C_s z_s(\tau_s) + D_s u_s(\tau_s) + w_s(\tau_s) \tag{14}$$

where

$$A_s = \Lambda_1, B_s = B_1, T_s = T_1, C_s = C_1,$$

$$D_s = -C_2 \Lambda_2^{-1} B_2, V_s = -C_2 \Lambda_2^{-1} T_2,$$

$$z_s(\tau_s) = z_1(\tau_s), u_s(\tau_s) = u(\tau_s), v_s(\tau_s) = v(\tau_s),$$

$$y_s(\tau_s) = y(\tau_s), w_s(\tau_s) = w(\tau_s) + V_s v(\tau_s) .$$

IV. SIMULATION RESULTS

Consider the control of given VTAV decomposed model (10)-(11), (13)-(14) for the case of take-off and forward maneuvers.

The goal of the following simulations is twofold. First, we verify that this control system is able to control the flight trajectory of given VTAV. Second, we observed the effect of enhancing SA because of increasing/decreasing speeds of all three ducted fans equally can be changed the possible flight trajectory of given VTAV.

Initial position and desired height for control system are chosen to be:

$$x(0) = 0m, y(0) = 0m, z(0) = 0m, z_1^0 = 6m$$

Simulation results of the offered block scheme (see Fig. 1) are shown in Figs. 2-6.

In [8], the two approximations to the nonlinear autoregressive moving average (NARMA) model called the NARMA-L1 and the NARMA-L2 are proposed. From a practical stand-point, the NARMA-L2 model is found to be simpler to realize than the NARMA-L1 model.

The neurocontroller used in this section are based only on the NARMA-L2 approximate model [8].

The NARMA-L2 Controller tuning can be accomplished quickly and accurately using internal windows for this block.

Some advantages of this example are as follows.

- Possibility to consider a terrain restriction in a place of destination.
- Possibility of lag in one selected height position.
- Using of one control system to control the take-off and forward trajectories of flight.
- Fine and simplified adjustment of this neurocontroller for any change of desired height position.

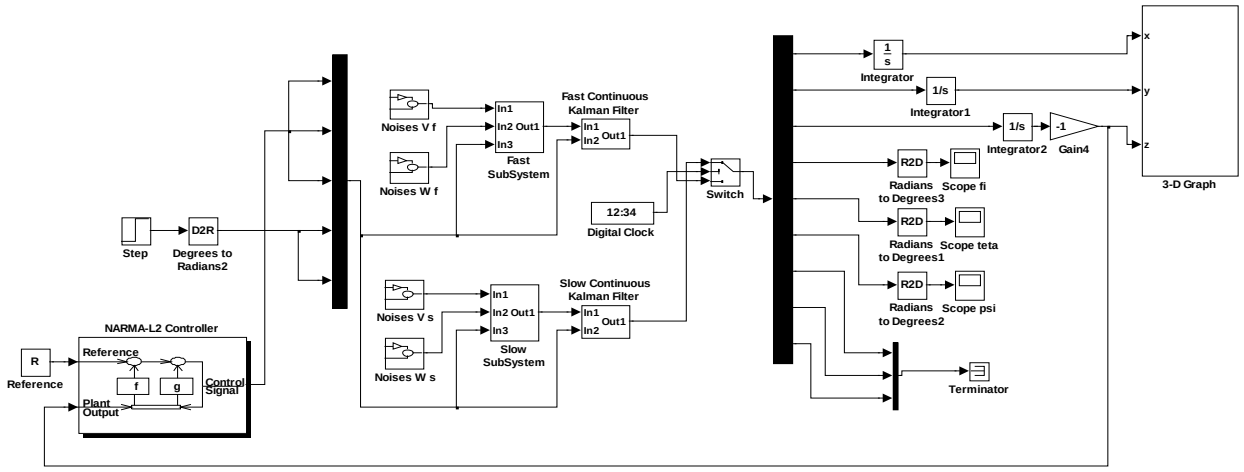


Figure 1. Block diagram of two-rate control system.

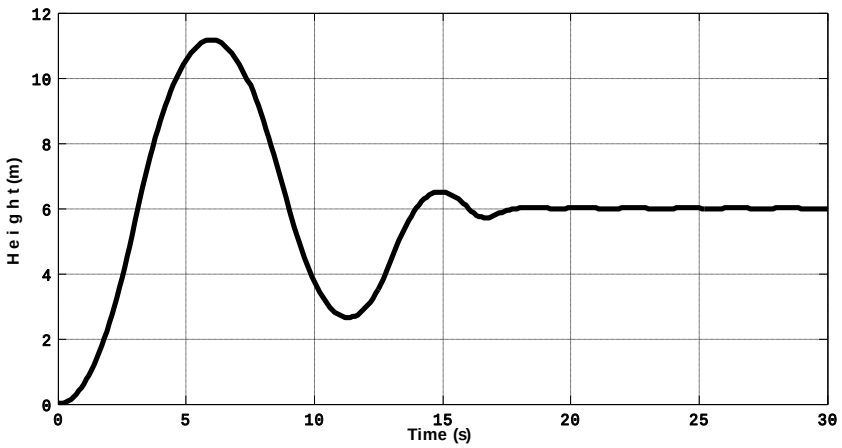


Figure 2. Height trajectory of VTAV.

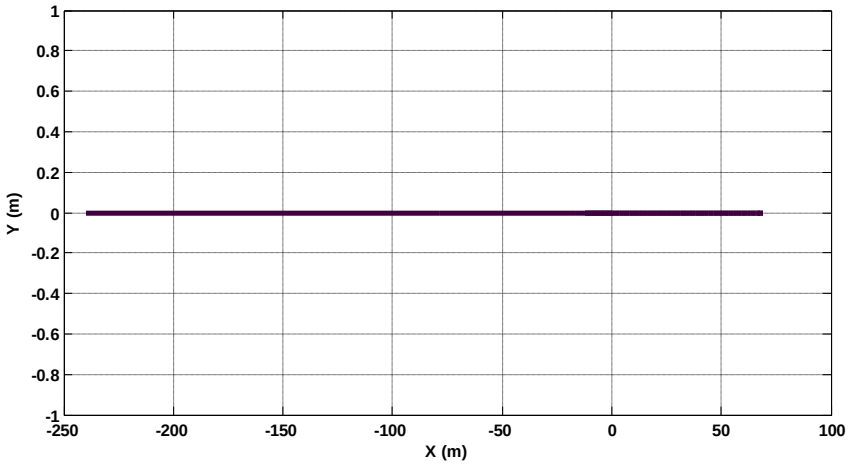


Figure 3. X-Y view of VTAV's trajectory.

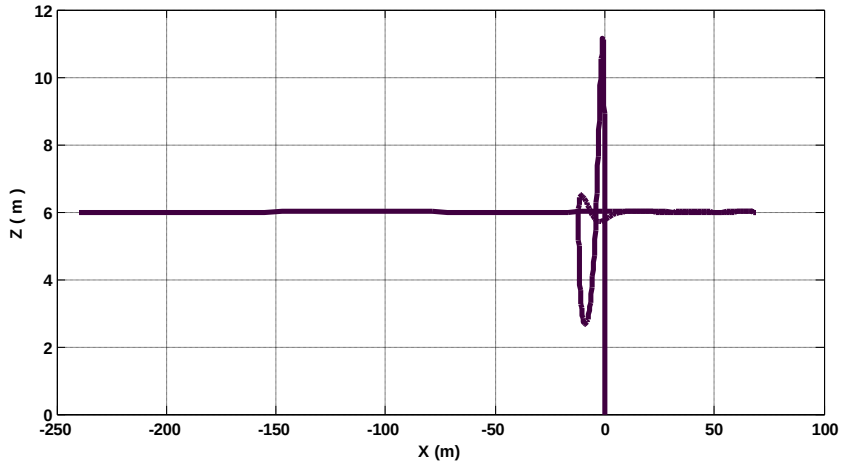


Figure 4. X-Z view of VTAV's trajectory.

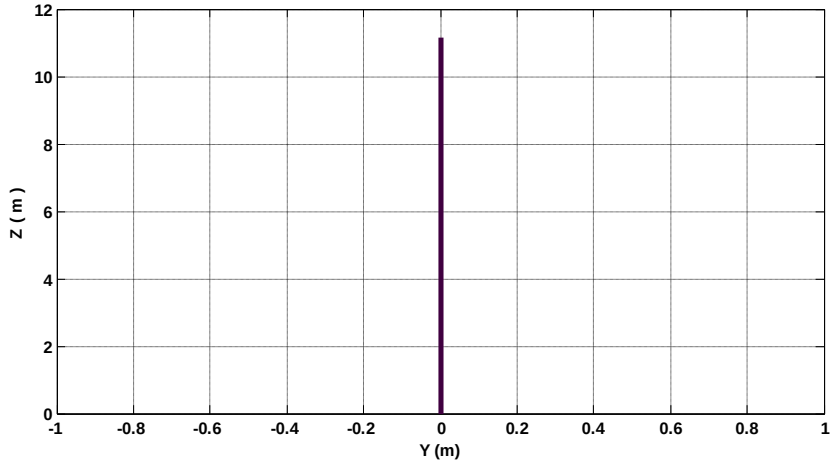


Figure 5. Y-Z view of VTAV's trajectory.

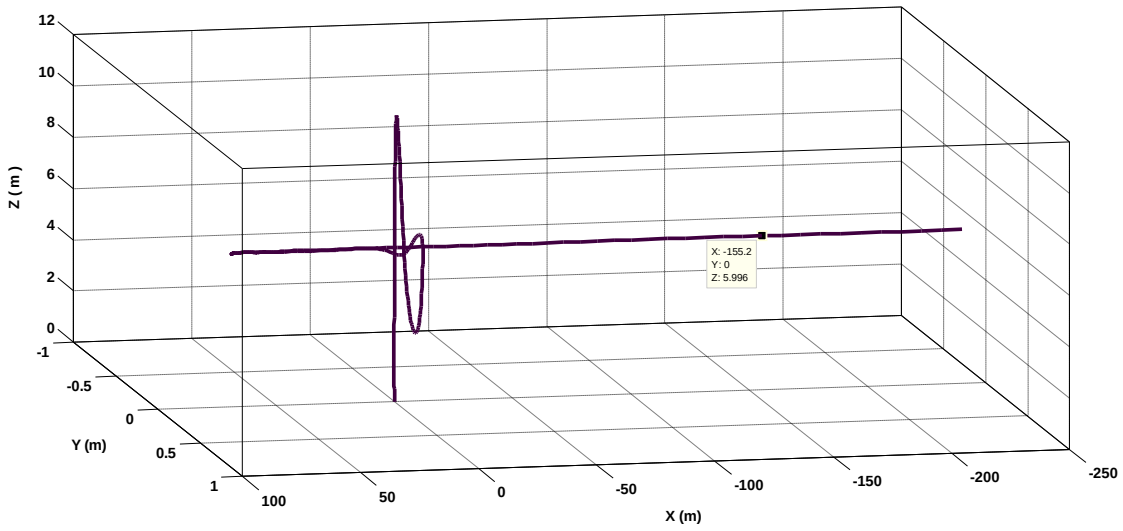


Figure 6. 3-D motion of VTAV.

V. CONCLUSION

A new research technique is presented in this paper for enhanced SA in possible missions of VTAV.

The need for highly reliable and stable hovering for VTOL class VTAVs has increased morbidly for critical situations in real-time search-and-rescue operations for fast SA.

By following the proposed methodology, the VTAV model structure (1)-(2) is decomposed into two subsystems: the "fast" subsystem used in the initial phase of trajectory (take-off motion) and the "slow" subsystem used in the final phase of trajectory (approach motion).

For fast, stable and smooth flight maneuvers, we used a NARMA-L2 controller. The effectiveness of this NN control has been verified in field of flight simulation tests for chosen model of VTAV using software package Simulink.

Note that the two-rate neural network flight control works more qualitatively than the single-rate neural network control [9] under the same conditions.

From the applications viewpoint, we believe that this flexible and effective control furnishes a powerful approach for enhancing SA in applications to VTAV class aerial vehicles in real-time search-and-rescue operations.

Future work will involve further validation of the performance of the proposed research technique and exploring other relevant and interesting missions of VTAVs.

REFERENCES

[1] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," Human Factors, vol. 37, pp. 32–64, March 1995.

[2] J. Gorman, N. Cooke, and J. Winner, "Measuring team situation awareness in decentralized command and control environments," Ergonomics, vol. 49, pp. 1312–1325, October 2006.

[3] Interim Brigade Combat Team Newsletter. [Online]. Available: http://www.globalsecurity.org/military/library/ report/call/call_01 -18_ toc.htm

[4] S. D. Prior, S. T. Shen, A. S. White, S. Odedra, M. Karamanoglu, M. A. Erbil, and T. Foran, "Development of a novel platform for greater situational awareness in the urban military terrain," Proc. 8th International Conf. Engineering Psychology and Cognitive Ergonomics, San Diego, USA, 2009, pp. 120–125.

[5] I. Astrov and A. Pedai, "Control of hovering manoeuvres in unmanned helicopter for enhanced situational awareness," Proc. International Conf. Industrial Mechatronics and Automation, Chengdu, China, 2009, pp. 143–146.

[6] W. Yuan, H. Jayakody, and J. Katupitiya, "Dynamic modelling and analysis of a vectored thrust aerial vehicle," Proc. Australasian Conf. Robotics and Automation, Wellington, New Zealand, 2012, pp. 1–8.

[7] I. Astrov and E. Rüstern, "Simulation of two-rate hybrid control for stochastic model of missile autopilot," Proc. International Conf. Computer, Communication and Control Technologies, Orlando, Florida, USA, 2003, pp. 150–154.

[8] K. S. Narendra and S. Mukhopadhyay, "Adaptive control using neural networks and approximate models," IEEE Trans. Neural Networks, vol. 8, pp. 475–485, May 1997.

[9] I. Astrov and N. Berezovski, "Neural network motion control of VTAV by NARMA-L2 controller for enhanced situational awareness," International J. Computer, Electrical, Automation, Control and Information Engineering, vol. 9, pp. 1382–1386, August 2015.

LADDER Program Solution for Multi-Probe Monitoring and Control in Simple Cooling Process

Tomislav Špoljarić*, Mirko Špoljarić**
*University of Applied Sciences - Department of Electrical Engineering, Zagreb, Croatia
**Frigo-Ing d.o.o., Kašina, Croatia
tomislav.spoljaric@tvz.hr, frigoing@yahoo.com

ABSTRACT - Simple cooling process is a thermal process that usually takes place in industrial refrigeration systems used for storage of different types of goods. This type of process is based on control that usually uses one or two temperature probes for temperature reference. It is also a static process that consists mostly of two repeating cycles: one of refrigeration and one of defrost. Multi-probe control uses more than two temperature probes set in different areas of a cold room. In this type of control temperature reference can be calculated via mean value algorithms with different weight factors attached to different probes. This type of control can be used in storage of sensitive goods since the accuracy of temperature measurement is greatly increased with number of probes.

I. INTRODUCTION

A simple cooling process is a self-repeating process consisting of a single refrigeration period and a defrosting period, if one is needed. It has predictable control flow and is used in situations where temperature of ambient air is used as a monitoring signal for controlling the operation of main elements of a refrigeration system: compressor, condenser fans, evaporator fans and defrosting heaters.

Temperature is central characteristic in a refrigeration regulation. Almost all refrigeration systems have the purpose of reducing the temperature, whether it is an ambient air temperature or directly a temperature of a goods being cooled, stored or frozen. Other characteristics associated with refrigeration systems are fluid pressure, heat transfer and cooling and electrical power of a system. Large scale refrigeration systems comprising of a several compressors connected in a multiset with several cold rooms rely also on a pressure control. These systems are driven by simple regulating devices with pressure as a reference and act independently from temperature regulation.

In today's increasing demands for quality of goods, a refrigeration control needs to be expanded with multi-probe control technique. Here, several temperature probe readings in the same cold room are used in adequate algorithms for gaining a greater accuracy towards a real temperature of ambient air or goods being stored.

II. SIMPLE COOLING PROCESS

A. *Basic elements of a refrigeration system in a simple cooling process*

Basically, a refrigeration process is a heat transferring between ambient air and a refrigerant which flows through refrigerant circuit. This circuit consists of basic elements: compressor, evaporator, condenser and expansion valve. As in every pipeline, additional valves are required. Usually in refrigeration systems solenoid valves are used for automated refrigerant flow control and shut-off valves are used for manual control.

Refrigerant that flows through a circuit changes its physical condition and pressure. In liquid form it will absorb heat when it evaporates and it is this conditional change that produces cooling in a refrigerating process. Heat will be taken up from the surrounding air and evaporation will occur at a refrigerant temperature corresponding to atmospheric pressure. This is analogous to the original temperature: it is pressure dependent. This stage of process takes place in a device called *evaporator*.

To maintain a lower pressure and with it a lower temperature it is necessary to remove vapor. This is done by the *compressor*, which sucks vapor away from the evaporator. In simple terms, the compressor can be compared to a pump that conveys vapor in the refrigeration circuit.

The refrigerant gives off heat in the *condenser*, and this heat is transferred to a medium having a lower temperature, usually air. The amount of heat given off is the heat absorbed by the refrigerant in the evaporator plus the heat created by compression input.

Liquid from the condenser runs to a collecting tank, the *receiver*. This can be likened to the tank mentioned under section 3.1 on the evaporator. Pressure in the receiver is much higher than the pressure in the evaporator because of the compression (pressure increase) that has occurred in the compressor. To reduce pressure to the same level as the evaporating pressure a device must be inserted to carry out this process, which is called throttling, or expansion. A device used for such operation is usually an *expansion valve*. By suddenly reducing

pressure a conditional change will occur and a liquid will evaporate which takes place in evaporator. This makes the refrigeration circuit complete [2].

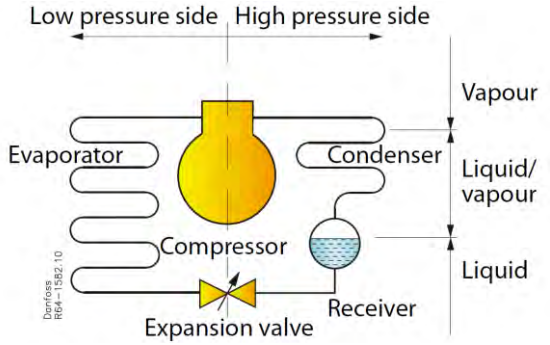


Figure 1. Simplified diagram of a refrigeration circuit and basic elements of a refrigeration system [2]

B. *Simple regulation techniques in a simple cooling process*

In today's refrigeration systems, mostly low cost regulators are used as temperature regulating devices for small and medium scale systems. These devices are based on simple electronic PIC technology with implemented simple algorithms for calculating the difference between temperature probe reading and a reference. Some regulators may implement a virtual probe reading, meaning that they use mean value of two probes and may use a percentage calculation for defining a priority of a probe. These regulation techniques are defined as follows:

$$\vartheta_{room} = \vartheta_{probe} \tag{1}$$

$$\vartheta_{room} = \frac{w_1 \cdot \vartheta_1 + (100 - w_1) \cdot \vartheta_2}{2} \cdot 100 \tag{2}$$

In a most simple manner a reading from a probe is used as a regulating information: the difference between probe reading and reference temperature ϑ_{ref} is used for switching the components on and off. Usually a temperature hysteresis is used for avoiding unwanted frequent switching which may cause failures in switching equipment. Some regulating devices have loaded algorithm for management of two probes. With percentage factor that defines priority of a first probe, this algorithm presents a "virtual probe". This term is defined by calculation of a mean value with a percentage factor (2).

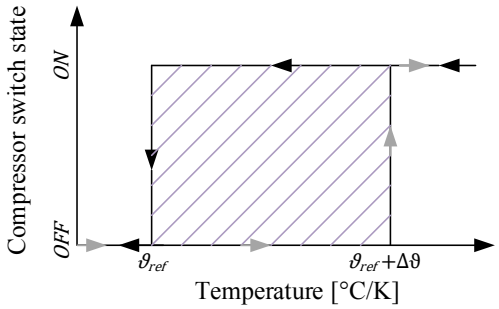


Figure 2. Diagram of a compressor switching state according to room temperature – direct mode

As shown on Fig. 1 refrigerating elements (such as compressor) are switched on if a logic condition $\vartheta_{room} > \vartheta_{ref} + \Delta\vartheta$, and after successful cooling operation they are switched off when $\vartheta_{room} \leq \vartheta_{ref}$. This type of control in refrigeration is known as *direct mode*.

Defrosting is needed when required room temperature needs to be maintained below 0°C (in some cases even below +5°C), due to icing on evaporator section. This is done by switching the electrical heaters that need to be inserted into the evaporator section or, when electrical defrost is not available, by defrosting with hot gas. Defrosting is done separately from other refrigeration system elements (compressor, fans, etc.) and is regulated by a separate temperature probe that is inserted in evaporator [3]. This type of control is known as *reverse mode*.

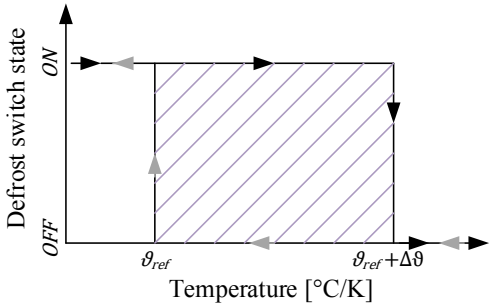


Figure 3. Diagram of a heater switching state according to evaporator temperature – reverse mode

According to time, simple cooling process is a static process where one (refrigeration) or two periods (refrigeration and defrost) alternate from time when a system is switched on until it is switched off. Diagram of such a process is shown on Fig. 4.

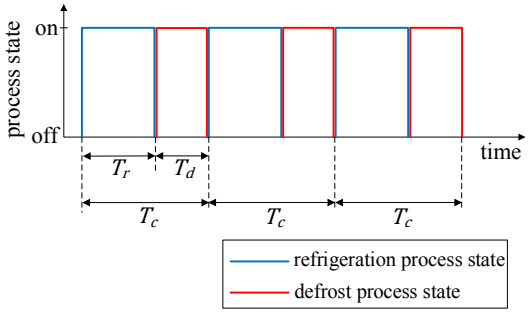


Figure 4. Time flow diagram of a simple cooling process

A repeating time cycle duration is described with simple equation:

$$T_c = T_r + T_d, \tag{1}$$

where T_r is a refrigeration period duration and T_d is a defrost period duration. When refrigeration process is active, temperature regulation according to room temperature probe is active. When defrost process is active, temperature regulation according to evaporator temperature probe is active [3].

C. *Multi-probe monitoring and control in a simple cooling process*

Certain goods require strict quality control. This requires more accuracy in temperature monitoring and regulation. One way to ensure strict temperature control is with duct probes which are jabbed or inserted into a product, and in this way a goods' temperature control can be monitored with satisfactory accuracy. The problem with this type of control is in the type of goods controlled and cooled. With meat, fish, fruit and vegetables this is plausible, but with certain solid and liquid products, such as pharmaceuticals, drinks or even blood for blood banks, this type of control is impossible. So, an indirect type of accurate temperature measurement needs to be implemented.

Other way to gain temperature accuracy in cold rooms used for storage is to ensure more than two measurement places. The problem with this type of regulation is a that specific regulators do not support more complex control than with one or two probes. This is where PLC needs to be implemented as a main regulating device.

PLC's have several inputs and outputs and they can be modular. Additional modules can be added for specific purposes, such as analog input modules used for measuring devices (voltage, current, resistor, thermistor), digital input modules for alarming purposes (switching failures) or digital output modules used for switching appropriate actuating devices. PLC's are also useful for connection to BMS (Building Management Systems) since they have the possibility to connect to main plant monitoring system through several industry protocols via appropriate connection modules. Specific refrigeration regulators do not have that possibility and are often limited to one type of connection protocol, usually a protocol specific to the manufacturer of that regulator.

In this article a system with PLC as a controlling device is observed. Cold room for storing pharmaceutical products is monitored with five probes of which three temperature probes are used for refrigeration control, one temperature probe is used for defrost control and a humidity probe is used only for monitoring a humidity level.

III. LADDER PROGRAM FOR MULTI-PROBE CONTROL AND MONITORING

In this section a LADDER program review is given. This program is used for controlling the ambient conditions of a cold room used for storing medical products. Demands for such control include temperature monitoring of three sections of a cold room and regulation of switching refrigerating elements according to these three measurements. A simple mean value algorithm which incorporates weight factors is used in program. Also a specific demand was implementation of a security temperature hysteresis for each probe used in refrigeration control. This hysteresis is used for switching between one probe and three probe control. If temperature in certain section of a room deviates above that security hysteresis, one probe regulation takes over. Basically, refrigeration control is then referenced only from deviating section's probe reading. This is particularly useful in moments after the room door closes (after managing the goods by the operator). Cold room is cooled quickly, and heat is transferred properly across the room when a three probe regulation takes over.

Defrost control uses only one PT100 probe and switches heaters on evaporator according to that probe.

Humidity probe (4-20mA) is used only for monitoring of ambient humidity.

A. *Program flow*

LADDER program was made with VisiLogic software package that is specifically made for PLC Series Vision made by Unitronics®. LADDER Program is divided into main routine and several subroutines. Main routine is executing initially, and somewhere in its code it is possible to call for subroutine, part of the program which code is divided from main routine's code. This can be helpful when dividing a code by its practical use, and when a subroutine needs to be called for more than one time.

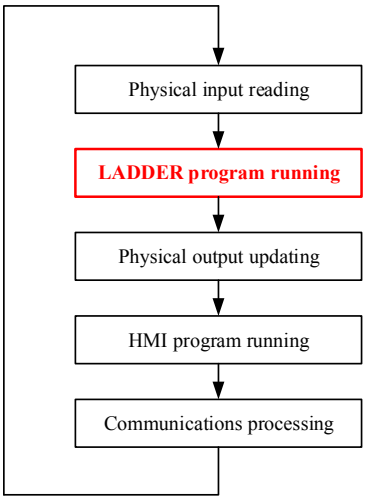


Figure 5. LADDER program running in PLC sequencing diagram

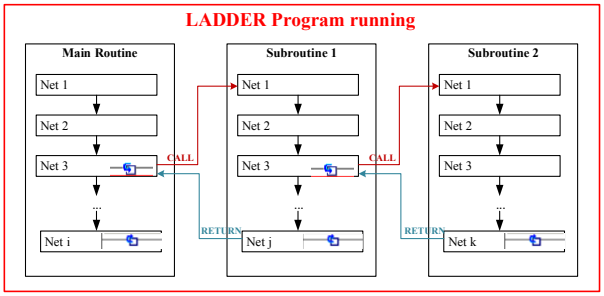


Figure 6. Subroutine call and return in LADDER program workflow

B. Main routine

Main routine is composed of several call subroutine orders after a power up bit and a physical input connected to first digital input of PLC. First subroutine consists of reading and processing analog inputs: humidity and temperature probes. Second subroutine consists of orders for loading and processing timers for regulation. Third subroutine is used for alarms. Fourth subroutine is used for regulation, and fifth is used for sending and receiving data from HMI display. This workflow is processed continuously until first digital input is set to 0. Then a program stops, and a cooling process is terminated.

C. Subroutine 1 – input data

Loading and processing analog input data starts with humidity probe loaded to linear block which gives an output as memory integer. In this way 4-20mA active input set to address MI4 is converted to integer form as 10 bit (4mA equals to integer of 204, 20mA equals to integer 1023). Output is given in integer form: 0%rH equals to integer of 4 and 99%rH equals to integer of 990. These are the limits of linear measurement given by active humidity probe. When output integer shows value below 0 or above 990, an alarm is triggered for probe failure after a time set for alarm. This is done with compare block. Humidity probe loading and processing is shown on Fig.7 and Fig. 8.

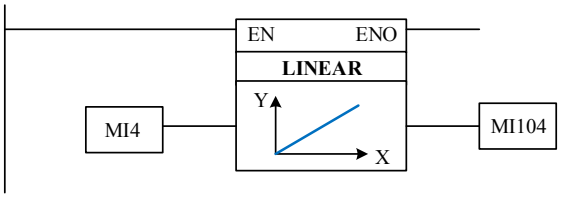


Figure 7. Linear block for processing humidity probe reading

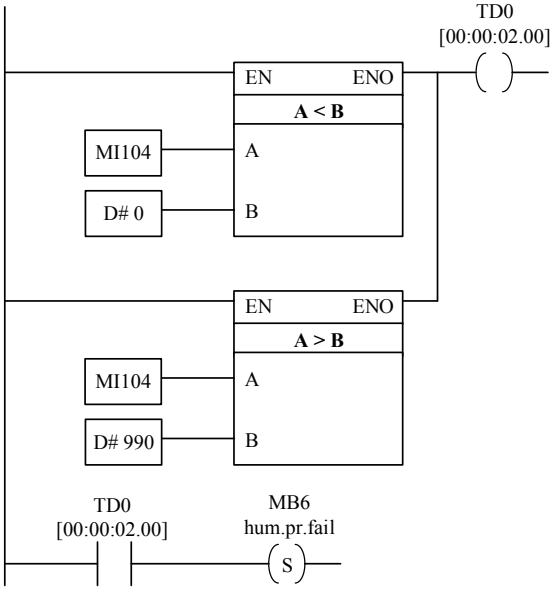


Figure 8. Comparison block for triggering an alarm in a case of humidity probe failure

For temperature monitoring a IO-PT4x module with four PT100 inputs is used. 12-bit values are stored in memory integers with 0.1°C resolution (25°C equals to integer of 250, -10.3°C equals to signed integer of -103). Since no linear block is needed, as in case with active inputs, only a comparison block is introduced for triggering an alarm, as shown on Fig. 9. This section is same for all four PT100 probes with MI100, MI101, MI102 and MI103 as values read from probes 1 to 4 and corresponding alarm timers are addressed in TD1 to TD 4, and set to trigger after 2 seconds. Limits for triggering are -50°C and 120°C.

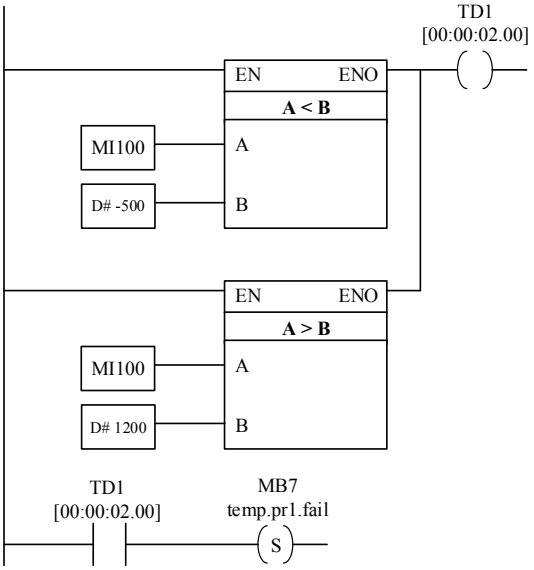


Figure 9. Comparison block for triggering an alarm in a case of temperature probe 1 failure

Also, a mean value with weight factors is used for defining a virtual probe reading. Since weight factors are limited from 0 to 100, a linearization block needs to be implemented - minima are same: zero, and maximum of output is 100 times smaller than input integer.

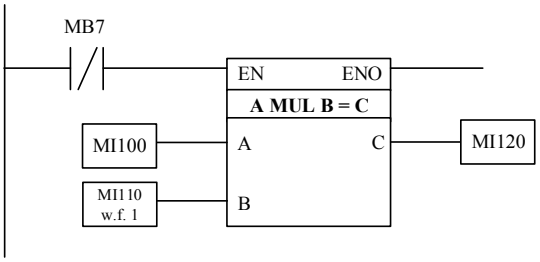


Figure 10. Multiplier block for assigning a percentage priority to probe 1

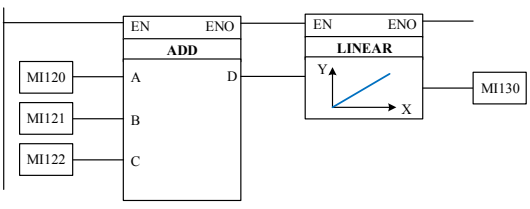


Figure 11. Addition of three values and linearizing the summation

D. Subroutine 2 – timers

In this subroutine a block called LD TP (load timer) is used three times with preset values (refrigeration, defrost and alarm reset period). One such block is shown on Fig. 12.

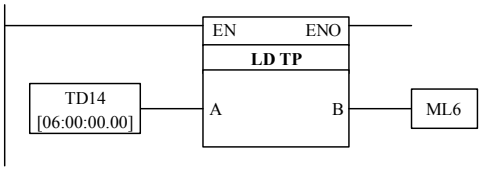


Figure 12. Load timer block for refrigeration time

Operand A contains value from address TD14 with preset value that equals to 6 hours and moves it into operand B on address ML6 which will be processed in HMI display subroutine. ML6 is 32-bit value of time with base in 10 milliseconds. This subroutine has two more LD TP blocks with T2 and T3 as operands A and ML7 and ML8 as operands B. In this way on addresses ML6, ML7 and ML8 are stored refrigeration time, defrosting time and pause interval. Pause interval is preset to two minutes and defrosting interval lasts for half an hour.

E. Subroutine 3 – alarms

Alarm managing is simple. All switching equipment is connected to digital inputs which are directly read and written to appropriate memory bit addresses (MB#). Probe failures are discussed in previous section. Also, high and low temperature alarms are introduced, and they operate in same manners as probe failure detection shown on Fig. 8, but in case of extreme temperature alarm new memory integer addresses are introduced and compared with comparison block. If extreme temperature occurs a new timer address is used for triggering appropriate alarms. Also if one of these alarms occur, process stops by resetting the memory bit that starts cooling process (in this example it is MB4).

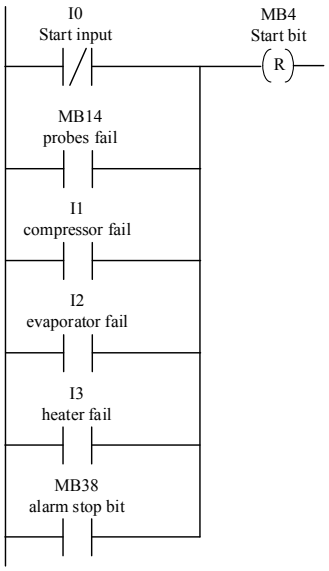


Figure 13. Resetting the start bit in alarms' subroutine

F. Subroutine 4 – regulation and writing to outputs

In regulation subroutine first networks define the situation when a process is disabled. Alarm triggering and manual disabling of bit that starts the process (via HMI display or physical input) are responsible for this part of code. All process bits and timers are reset when these conditions are met.

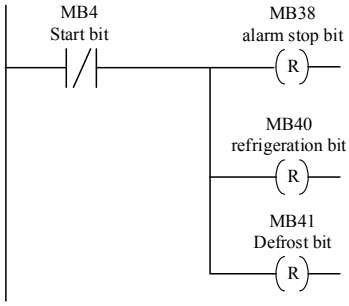


Figure 14. Resetting alarms and process bits

When a single shot impulse occurs in MB4 bit, a refrigeration bit MB40 is set on positive change of impulse. MB40 activates refrigeration timer TD14. When TD14 counts off impulse occurs, refrigeration bit is reset and defrost bit MB41 is set. MB41 activates defrost timer TD15. When TD15 counts off impulse occurs, defrost bit is reset and, by setting MB40 bit again, cycle repeats itself until start bit is set to zero. This type of process is shown on Fig. 15.

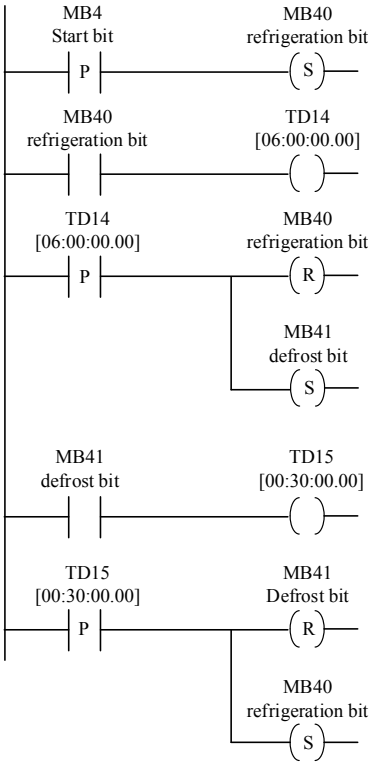


Figure 15. Continuous cooling cycle

As a reference in refrigeration, two hysteresis are introduced: standard hysteresis for reference (MI12) and security hysteresis (MI13). Add block for reference and standard hysteresis gives output in address MI25. Additional add block for MI25 and

security hysteresis MI13 gives output in MI26. These integers are later used as regulation references in two modes of regulation. Comparison block dictates which mode is enabled. If one of the probes deviates out of limits above temperature defined in MI26, regulation is set to work by that probe's reading. Otherwise, regulation works through mean value of prioritized probes.

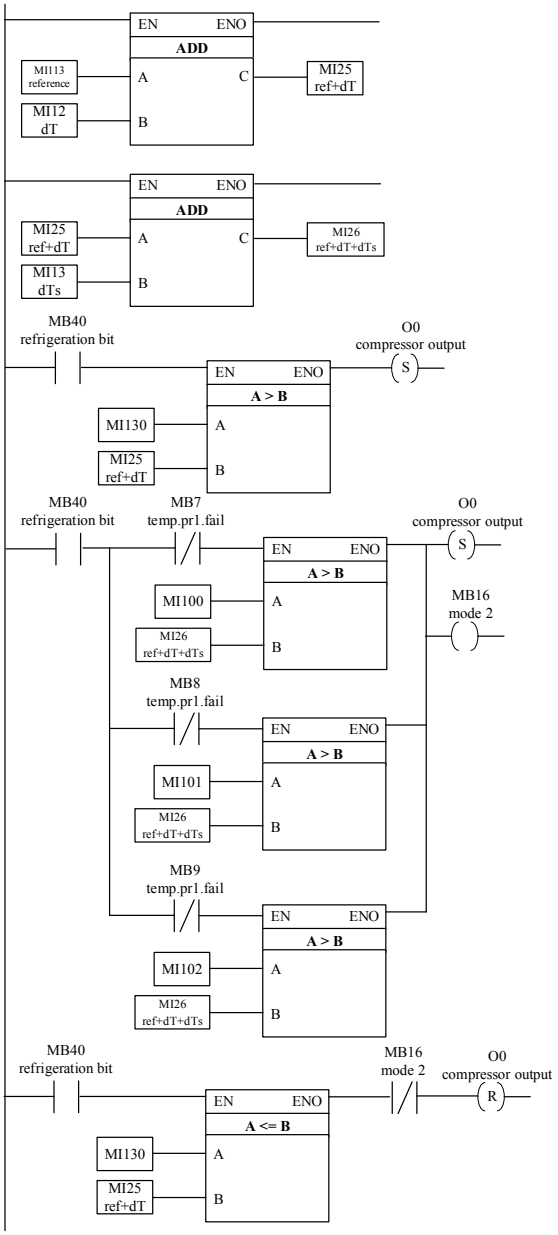


Figure 16. Regulation process for refrigeration

Defrost control is defined in a similar way. Only a standard hysteresis is used. Heaters are switched on during defrosting period if evaporator temperature falls below referenced temperature. They are switched off if temperature grows above the sum of defrost reference and hysteresis (2°C fixed).

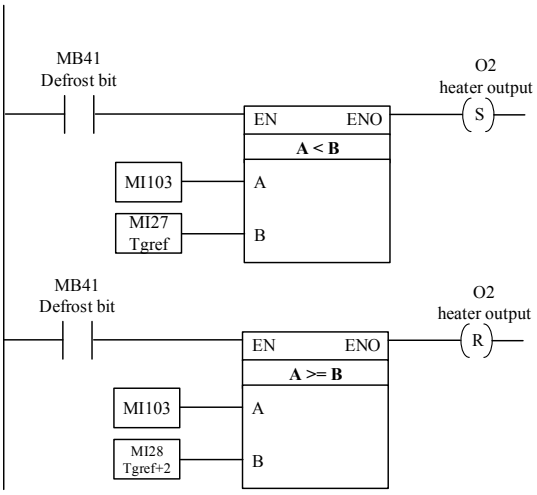


Figure 17. Regulation process for defrosting

G. *Subroutine 5 – HMI display*

This subroutine is used for loading and receiving data blocks such as memory integers, memory bits and output status into a graphical interface via HMI display. In this article a PLC+HMI all-in-one device V430-J-RH2 was used and a graphical interface was programmed in same software package as is a LADDER program. On Fig. 18 a monitoring screen setting is shown. On Fig. 19 alarm screen setting is shown.

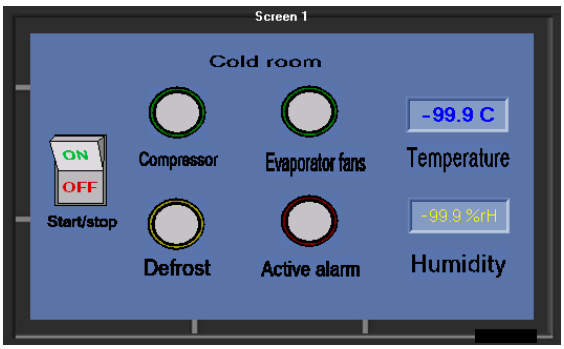


Figure 18. Monitoring screen on HMI display

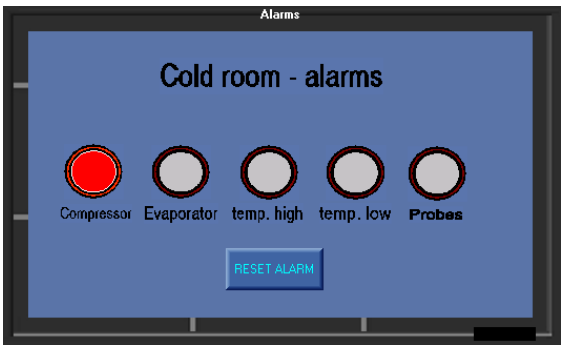


Figure 19. Alarm screen on HMI display

H. *Additional subroutine – Communication settings*

If a MODBUS or PROFIBUS network connection is present, a configuration block needs to be implemented after a power up bit on a start of the workflow, and on last place a subroutine for managing network variables is defined. This is used

for monitoring via main managing system, usually a BMS server.

As result, overview of LADDER program's flow is shown on Fig. 20 in a form of diagram.

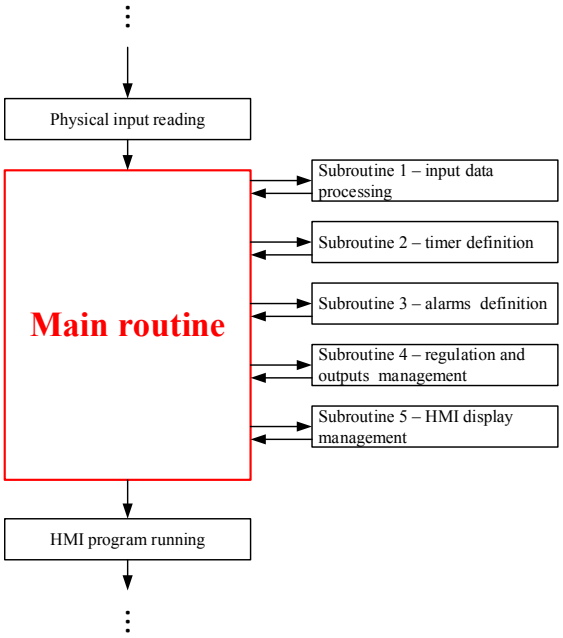


Figure 20. Program flow diagram of a LADDER solution to a simple cooling process with multi-probe monitoring and control

IV. Testing of a Multi-Probe Ladder Program – Experimental Results

A. Cold room description

Described program was tested in a real cold room control system used for storage of pharmaceutical and medical products, with 85% of its storage capacity fulfilled. Dimensions of a room and placement of various equipment is shown on Fig. 21. Independent temperature measurement was done with IR thermometer with ±1% accuracy deviation for ambient temperatures above 0°C (Fluke® 568). Independent measurements were first used for determining the best place for probes. Afterwards, with cold room in operation, independent measurements were used for determining the deviation between storage section temperature and temperature read by the PLC probes.

Since a multi-probe monitoring and control in described way was implemented in real cooling system, first task was to find appropriate weight factors for probe priority assignments. After few tests, weight factor of 45, 30 and 25 percent were used as satisfactory when parallel temperature measurement tests were made. Weight factor was chosen smaller with distance of installed probe farther from the entry door, since the cold air tends to flow out of the room when entry door opens. This makes the temperature of the space closer to the entry door higher than the rest of the room with opened door. With that in mind, a higher priority is assigned to a probe placed near entry door, so the cooling regulation will always operate with that condition.

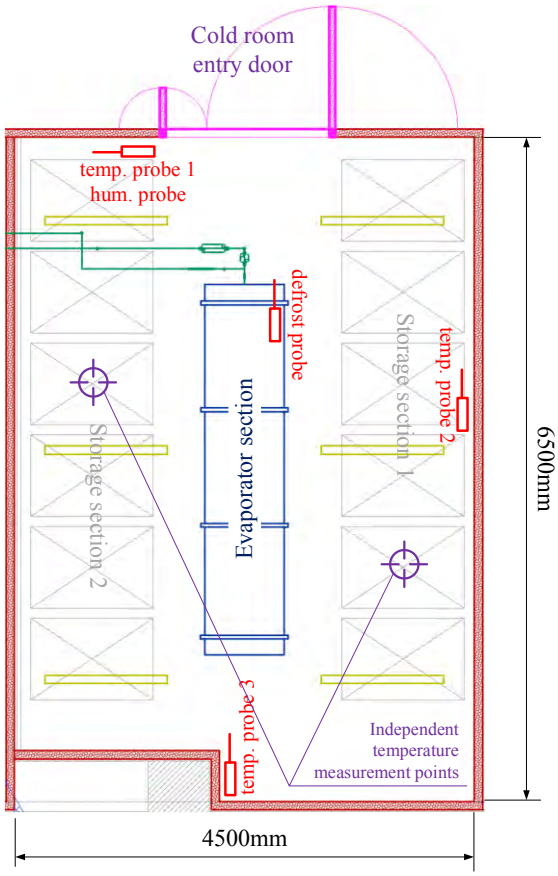


Figure 21. Disposition of a cold room used in testing

TABLE I. Weight factor assignment

Room probe *(address)*	**Weight factor *(address)***
Temperature probe 1 *(MI100)*	45% *(MI110)*
Temperature probe 2 *(MI101)*	30% *(MI111)*
Temperature probe 3 *(MI102)*	25% *(MI112)*

B. Cold room in operation – testing description

Testing period lasted for 10 hours (600 minutes). Temperature measurement resolution was 5 min per reading. During a testing period, one defrosting operation occurred (at 360th minute) and lasted for 10 minutes. Starting temperature was 16°C. Doors were opened three times: at 160th minute, 320th minute, and 5 minutes before the ending of a test, and they were left open for 10 minutes. During these periods independent measurements were taken in storage places (as shown in table II). Deviations were mostly below 6% with few peaks above that percentage. Also opening the door of room resulted in higher temperature in space closer to the door, because the cold air in that place got out first. This resulted in higher temperature reading from temperature probe 1. These peaks in readings are shown on Fig. 22.

When temperature rose above 6°C, regulation according to room temperature probe 1 took place. It lasted until reading from mentioned probe fell under 2°C. Then, after reading probes 2 and 3, and given a satisfactory input under 6°C to PLC from each probe, a three probe regulation was restored. Results can be seen on Fig. 22 and Fig. 23. Regulation probe is the term which defines the measuring value. According to difference between this measuring value and temperature reference system regulation is set. Regulation probe can be presented by each of temperature probes (if one of them deviates above reference and hysteresis) or by mean value calculation with weight factors which includes all three probes' readings, as described in previous section.

Compressor states in cooling mode of operation and door status is shown on Fig. 24. Temperature movement in higher peaks can be seen as result of opened door. Compressor state according to regulation probe measurements shows satisfactory operation of compressor and with it the operation of other refrigerating elements of the system.

TABLE II. REGULATION PROBE DEVIATIONS FROM INDEPENDENT MEASUREMENTS

Time point of measurement [min]	Regulation probe reading [°C]	Temperature measurement 1 [°C]	Deviation of reg. probe from measurement 1 [%]	Temperature measurement 2[°C]	Deviation of reg. probe from measurement 2 [%]
160	5	5.2	3.17	5.1	1.19
165	6.4	6.5	0.85	5.8	-10.01
170	8.7	8.7	-0.51	8.5	-2.8
320	4.3	4.5	5.76	4.2	-1.29
325	5.8	6	3.18	5.5	-5.42
330	9.5	10	8.9	8.9	-6.02
595	6.2	5.9	-4.3	6	-2.67
600	8	8.3	3.23	7.6	-5.47

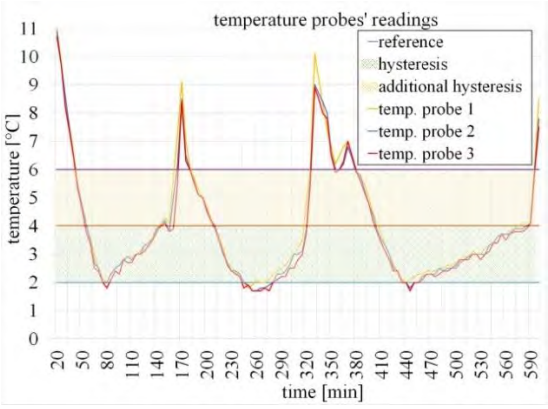


Figure 22. Readings from three temperature probes during the testing period

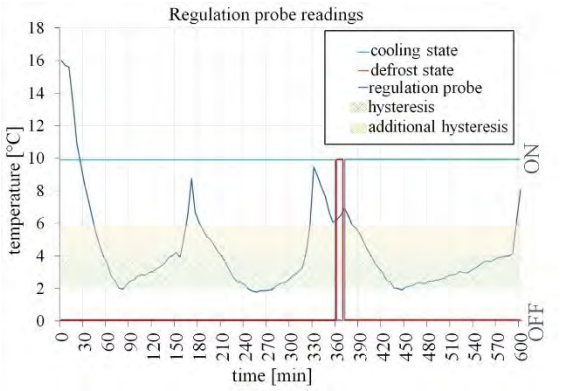


Figure 23. Regulation probe and cooling/defrost state readings during the testing period

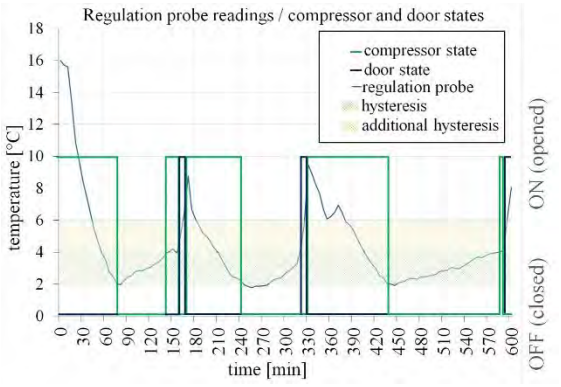


Figure 24. Regulation probe and compressor/door state readings during the testing period

V. CONCLUSION

This type of control has proven to be a challenge since control with more than two probes has not yet been implemented in standard refrigeration regulators. Installed PLC that was described in article has also proven to be on a satisfactory level according to task given to it. But even with low cost PLCs, such as one used in this case, a price for such control equipment presents a high demand for end user. Standard refrigeration controllers are low cost and storage systems usually do not demand regulation via more than two probes. This is the main reason why PLCs are not present more often in refrigeration market.

Trending time-temperature curves in form of graphs can be also monitored, but in these article they were not presented since BMS system which receives data via MODBUS communication protocol covered that area of monitoring in testing system. In local monitoring mode these trend curves can be very useful for end user.

Further possibilities include installment of frequency converters on fans for fine regulation of speed and with it a fine regulation of condensing and suction pressure levels. Indirectly, this means a fine regulation in temperature may be obtained with these devices. Since heuristic approach to weight factor definition was used in testing it would also be interesting if weight factors could be obtained using

intelligent software systems used for data processing, such as fuzzy control or neural networks, but this research also requires adequate time and finances and with little possibility of significant development in accuracy.

Described system that was installed in a real cooling system, a cold room used for storage of pharmaceutical products, was inexpensive and challenging. And today it is very successful in playing a part in regulation of a simple, yet demanding, cooling process.

REFERENCES

[1] "SM2 Cold Rooms, Pliva: Strojarski projekt – tehnički opis rada regulacije i nadzornog sustava", project's technical description, Frigo-Ing d.o.o., Kašina, 2015.

[2] "Refrigeration – An Introduction to the Basics", lecture notes, Danfoss, Nordborg, 2007.

[3] Jagetić G., Habazin M., Špoljarić T., "PLC/SCADA systems in automation control design for individual quick freezing process in cooling tunnels", 38th International Convention, MIPRO, Opatija, 2015.

[4] Špoljarić T., Fruk M., Vujisić G., "Program Solutions for The Complex Cooling Process Via LADDER Logic", 37th International Convention, MIPRO, Opatija, 2014.

[5] Vision OPLC: technical description, www.unitronics.com

[6] IO-PTx module: technical description, www.unitronics.com

An M2M Solution for Smart Metering in Electrical Power Systems

M. P. Shopov
Technical University of Sofia / Department of Computer systems and Technologies, Plovdiv, Bulgaria
mshopov@tu-plovdiv.bg

Abstract - The paper presents an implementation of smart metering application in electrical power systems. It makes uses of cloud computing and M2M technologies. The use of cloud computing technologies increase reliability and protection of collected data and eases the software developers in management of remote devices – the actual devices are managed through web services and provided REST-based interfaces. The smart meter is based on PST power transducer, developed at the Technical University of Sofia, an embedded platform A20-Olinuxino and DeviceHive M2M platform running on a private virtual infrastructure.

I. INTRODUCTION

With the development of sensor networks, wireless mobile communication, embedded system and cloud computing, the technologies of Internet of Things (IoT) have been widely used in areas such as Smart Meter, public security, Smart homes and so on [1]. There are three essential components of IoT [2]: embedded devices – consisting of both low cost/low power devices and high-end gateways; scalable connectivity – each embedded device should be connected; cloud-based mass device management – centralized management of distributed devices.

The joint use of distributed embedded systems and cloud computing technologies is an extension to the efforts of integrating monitoring and control within the business information systems. The aim is to support the binding of monitoring and management of physical processes from the environment with the business logic and business processes [3].

A. Machine-to-machine

Machine to machine (M2M) is a broad term used to describe any technology that enables networked devices – wired or wireless - to exchange information and perform actions without the manual assistance of humans. As such the M2M is an integral part of IoT. The support of legacy devices and devices without proper communication capabilities, will require the use of M2M gateways/aggregation points. These gateways could also be used for some local intelligence and value-added services [2].

National Science Fund of Bulgaria - contract E02/12.

B. Smart metering

One of the applications of M2M is smart metering. It integrates communication capabilities with electrical power systems and delivery infrastructure to automate monitoring and control. Dynamically linking utility supply with demand could result in optimization of resource consumption [2].

Smart meter is an advanced power meter that represent advanced metering infrastructure for enabling an automated, two-way communication between the utility meter and the utility provider. The smart meters are equipped with two interfaces: power reading interface and communication gateway interface [4].

C. Cloud-based applications

Cloud-based applications runs in a cloud computing environment and are delivered over the Internet. Some of the advantages of moving the server-side applications within the cloud are [5], [6], [7]:

- Application development: Development on the cloud will shorten development and prototyping time;
- Device heterogeneity: all the computation and storage is performed on the cloud infrastructure. The end-devices should only implement the communication protocol;
- Flexibility and scalability: Elastic resources allocation. Software updates does not affect end-devices;
- Mobility: Smart meters needs to be connected to Internet. All the computations are on the cloud infrastructure;
- Common user interface: Decoupling the computation infrastructure and the input system, enables multiple user interfaces to exist side by side allowing user-centric customization.

II. CLOUD-BASED SMART METERING APPLICATIONS

A. DeviceHive

The proposed solution for smart metering in electrical power systems uses DeviceHive [8]. It is an open-source M2M framework that contains a set of services and

components for establishing a two-way communication with the remote devices using cloud technologies as a middleware. The devices can be anything connected: sensor networks, smart meters, telemetry, smart home devices and etc [8].

DeviceHive framework consists of devices, organized in networks, server application working in the cloud and client applications (Fig. 1). Devices that are not capable of connecting directly to the cloud, uses an intermediate node called gateway, which is responsible for communication with DeviceHive cloud and for managing devices (through some custom or binary protocol). Each device has one or more equipments, i.e. sensors or actuators. Clients can manage only those devices that are part of a network, the client has permissions for.

The DeviceHive framework provides an application programmable interface (API). The API is used for interaction between individual components in real time and provides access to information about registered components in the system. The API is a RESTful service, which operates on several resource types (Fig. 2) [8]:

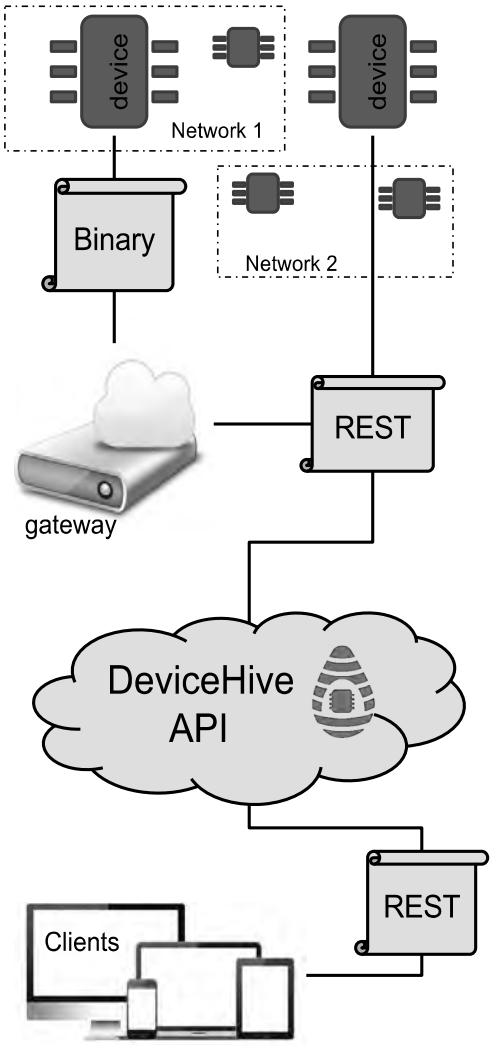


Figure 1: DeviceHive Framework

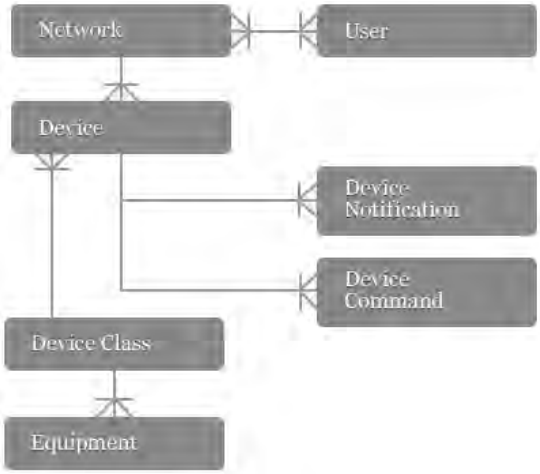


Figure 2: DeviceHive resource types relations

- User: represents a user with credentials and API access rights (client or administrator);
- Network: represents an isolation entity that encapsulates multiple devices with controlled access;
- Device: represents a unit that runs microcode;
- Device Class: represents an entity that holds all meta-information about particular type of device;
- Equipment: represents meta-information about one unit of equipment devices have onboard;
- Device Notification: represents a message dispatched by devices for clients;
- Device Command: represents a message dispatched by clients for devices;
- Access Key: represents an access key with specific API access rights.

The DeviceHive server application is deployed on Java servlet container running on a virtual machine in a private cloud infrastructure. It makes use of a publish-subscribe messaging system provided by Apache Kafka broker [9]. The broker defines a set of topics, producers and consumers that are used both for remote communication between clients and devices and for collecting remote measurements and raw logs. The measurements are stored on a Apache Hadoop database, where they can be further processed and analyzed, and the logs are consumed by a syslog server.

The Hadoop database is selected because it allows distributed processing of large data sets across clusters of computers. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, and delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures [10].

The private cloud used to run DeviceHive server application, Apache Kafka and Hadoop is based on VMWare and is implemented in Technical University of Sofia, branch Plovdiv. Its implementation and applied traffic optimizations are presented in [11] and (http://dsnet.tu-plovdiv.bg/energy/privatecloud).

B. Electrical power meter

The electrical power meter used is PST04. It measures the values of the main parameters of the three phase electric power system – voltage, current, frequency, active power, reactive power, power factor, active energy and reactive energy [12]. These transducers are developed at Development Laboratory for Semiconductor Circuit Engineering at Technical University of Sofia. They have two-wire instrumental serial interface [13].

C. Machine-to-machine gateway

The integration of electrical power meters with the DeviceHive uses a M2M gateway. The gateway is based on an A20-Olinuxino embedded hardware platform. It comes with Allwinner A20 dual core Cortex-A7 processor, 1GB RAM, Gigabit network interface, 2 USB ports, MicroSD card connector and GPIO ports [14].

On the hardware platform an Ubuntu Snappy Core [15] operating system is installed and running. Ubuntu snappy is a stripped down version of Ubuntu, designed to run securely on autonomous machines, embedded devices and other Internet-connected devices. The installation uses buildroot - a tool that simplifies and automates the process of building a complete Linux system for an embedded system, using cross-compilation, and the source repository provided by [16].

For connecting the gateway with the DeviceHive server application running in the private cloud the DeviceHive IoT toolkit [8] is used. It is provided as an Ubuntu snappy package. The IoT toolkit plays the role of cloud adapter and runs as a service in the operating system. It is responsible for connecting, re-connecting, authenticating and throttling messages to DeviceHive server application. Its configuration include the URL of DeviceHive server, access key for authentication, network name, network key (optional), device id and device key.

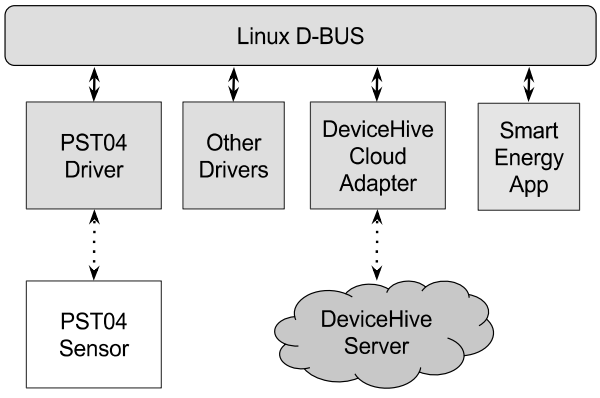


Figure 3: M2M Gateway software components

The communication with the DeviceHive cloud adapter service uses the standard Linux message bus (D-Bus). DBus is an inter-process communication and remote procedure call mechanism that allows communication between multiple computer programs (e.g. processes) concurrently running on the same machine. Two programs are developed and running on the gateway – a PST04 Driver and Smart energy application (Fig. 3).

The PST04 driver is responsible for communication with remote sensor. The sensor is connected through a serial interface and provides a simple custom protocol to fetch raw measurement data. These data are then processed by the driver to calculate the values of electrical power parameters. The measurements are then send through D-Bus and DeviceHive cloud adapter in the cloud. The driver sends measurements on periodic basis – the period can be configured by the user from the cloud, or asynchronously on client request.

The purpose of Smart energy application is to provide a local reasoning capabilities and automatically switching on/off consumers based on the power measurements and other sensors readings. In its current implementation it lights an led on when the power consumptions is above defined boundary.

D. Client applications

The development of client applications is not concerned with the actual end-devices. It should only implement the REST API provided by DeviceHive cloud. Client application allow administration of users, networks and device cloud, monitoring of devices and obtaining both real-time measurements (from DeviceHive) and statistical data (from Hadoop database). An implementation of a client application for real-time observation of measurement data from electrical power meter, sensor status and location is shown on Fig. 4. The implementation uses FreeBoard [17] software. It is an open-source project that can be extended and provides an interactive dashboards and visualizations for IoT.



Figure 4: Power meter dashboard

E. Security considerations

Power consumption data are sensitive to privacy and security issues and by using cloud-computing technologies for collecting and storing such data, several security considerations should be addressed [18].

Since the software components of the proposed solution is running in a private cloud infrastructure the

first level of security is the infrastructure security. It includes concepts of using shared computing resources, datastores and internal networks. That part relies on the security mechanisms and technologies provided by the underlining virtualization platform and will not be discussed in the paper.

On the application level, security includes identity and access management and communication security. In the proposed solution the DeviceHive application uses Oauth 2.0 for authentication and authorization. OAuth is an HTTP-based authorization protocol. The integration of Oauth consists of obtaining a session access token by using the application key, store it locally and renew it appropriately, the token is required in each request to the DeviceHive API. The communication with the REST API uses SSL. The different smart meters are grouped into networks on the DeviceHive and each network is protected by shared password.

On the devices level (smart meters) security mechanisms uses the Ubuntu Snappy system policies (policy groups). Currently, no encryption of the locally stored data is provided, but it could be easily added in future. Privacy consideration are addressed by encrypting the communication with DeviceHive REST API (using SSL) and depersonalization of smart meters – each smart meter is assigned an unique ID and their mapping with identity information is done on the DeviceHive server. This could be further improved by using external service to re-personalize the collected data. By depersonalizing the collected data it could be used by external services (even on public clouds) for statistical or other processing.

III. CONLUSION

The paper presents a practical implementation of a system for smart metering in electrical power systems that makes use of cloud technologies and IoT. The use of cloud computing technologies increase reliability and protection of collected data (it is stored and replicated in multiple secure, commercial-grade storage systems) and eases the software developers in management of remote devices – the actual devices are managed through web services and provided REST-based interfaces. The device cloud is based on DeviceHive M2M framework. It is an open-source platform and can be further developed, extended, and hosted in private and public environments. Although, it still has lack of some functionality, particularly in client-side services the solution presented in the paper shows the feasibility of applying cloud technologies in measurement and control of electrical power systems.

The future work includes realization of a smart energy sensor that combines the PST04 energy sensor and A20 Olinuxino gateway in a single board. Another directions for future work includes developing client applications for obtaining statistical information about power consumption for long periods of time, implementing local reasoning algorithms as part of the Smart energy application running on the gateway, and integration of private and public cloud infrastructures.

ACKNOWLEDGMENT

The presented work is supported by the National Science Fund of Bulgaria project “Investigation of methods and tools for application of cloud technologies in the measurement and control in the power system” under contract E02/12 (http://dsnet.tu-plovdiv.bg/energy/).

REFERENCES

[1] Beng, L., “Sensor cloud: Towards sensor-enabled cloud services,” Intelligent Systems Center Nanyang Technological University, 2009.

[2] Wu, Geng, Shilpa Talwar, Kerstin Johnsson, Nageen Himayat, and Kevin D. Johnson. "M2M: From mobile to embedded internet." Communications Magazine, IEEE 49, no. 4 (2011): 36-43.

[3] N. Kakanakov, S. Stoyanov, “Applying cloud technologies and sensor networks in monitoring and control of energy management systems,” Journal Automatica and Informatics, vol. 49, no. 1, pp. 11-14, 2015. [Н. Каканаков, Ст. Стоянов, “Използване на облачни системи и мрежи от сензори за следене и контрол в електроенергийната система,” Списание Автоматика и Информатика, бр. 1, стр. 11-14, 2015]

[4] Zubair Md. Fadlullah, Mostafa M. Fouda, Nei Kato, Akira Takeuchi, Noboru Iwasaki, and Yousuke Nozaki, “Toward Intelligent Machine-to-Machine Communications in Smart Grid,” IEEE Communications Magazine, vol. 49, no. 4, pp. 60 - 65, April 2011.

[5] Alamri A, Ansari WS, Hassan MM, Hossain MS, Alelaiwi A, Hossain MA, "A survey on sensor-cloud: architecture, applications, and approaches," International Journal of Distributed Sensor Networks, Article ID 917923, Feb 2013, doi:10.1155/2013/917923, 2013.

[6] Botta, A.; de Donato, W.; Persico, V.; Pescape, A., "On the Integration of Cloud Computing and Internet of Things," in Future Internet of Things and Cloud (FiCloud), 2014 International Conference on , vol., no., pp.23-30, 27-29 Aug. 2014.

[7] Nabi, Zubair, and Atif Alvi. "Clome: The practical implications of a cloud-based smart home." arXiv preprint arXiv:1405.0047 (2014).

[8] DeviceHive official site: devicehive.com/

[9] Apache Kafka offical site: kafka.apache.org/

[10] Apache Hadoop offical site: hbase.apache.org/

[11] Grozev D., M. Shopov, and N. Kakanakov “Practical Implementation of Private Cloud with traffic optimization” unpublished.

[12] Nikolov G., P. Yakimov, “Virtual Three Phase Power Transducer,” Annual Journal of Electronics, pp. 184, ISSN 1313-1842, 2009.

[13] Yakimov P., S. Ovcharov, N. Tuliev, E. Balkanska, R. Ivanov, “Three phase power transduser for remote energy management system application,” Annual Journal of electronics, vol. 4, no. 2, 2010, pp. 31-34, ISSN: 1313-1842.

[14] Olimex A20-Olinuxino offical site: www.olimex.com/Products/OLinuXino/open-source-hardware

[15] Ubuntu Snappy Core official site: http://www.ubuntu.com/internet-of-things

[16] Miroslav Bendík repository: github.com/mireq/buildroot-a20-olinuxino

[17] Official Freeboard github repository: https://github.com/Freeboard

[18] J. Liu, Y. Xiao, S. Li, W. Liang and C. L. P. Chen, "Cyber Security and Privacy Issues in Smart Grids," in IEEE Communications Surveys & Tutorials, vol. 14, no. 4, pp. 981-997, Fourth Quarter 2012.

Noise within a Data Center

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko.miljkovic@hep.hr

Abstract - Data centers with its numerous servers, network switches, routers and air conditioning equipment produce significant noise that influences work and communication of maintenance staff. Noise levels and spectrum within a data center are measured and analyzed. Daily and weekly variations of noise levels are also captured. From measured noise levels and spectrum speech intelligibility measures are calculated and communication distances within a data center are determined. The need for noise protection and methods for noise reduction are considered.

I. INTRODUCTION

Work in a data center exposes staff to some suboptimal working conditions. Noise, temperature and air circulation are three of the biggest environmental issues that data center workers face. Air-condition and equipment cooling fans necessary for the proper operation of IT equipment run continuously and create excessive noise that influence comfort, poses risk to hearing and impairs communication and concentration. With the introduction of smaller hardware, data center is even more densely populated and noise inside is becoming louder. Constant humming of air-conditioners and servers within a data center may create noise sometimes of excess of 80 dBA. Although maximum acceptable 8 hour exposure level in most countries is set at 85 dBA, as 10% of the level that does not show any effect (NOEL - No Observed Effect Level), this level is far above levels prescribed for mentally demanding work and office environment.

II. SOURCES OF DATA CENTER NOISE

Data centers consist of servers, routers, switches, storages, tape systems, UPS and cooling equipment (HVAC – Heat Ventilation Air Condition). Main sources of data center noise are shown in Fig. 1. HVAC with its supply ventilation ducts is a single main source of a data center noise, with noise level typically about 70 dBA. Within IT equipment main contributors to data center noise are cooling fans installed in numerous servers, routers, switches and storages, although some minor noise comes from disks, tape drives, transformers and human intervention. Individual servers produce noise levels in a 40-70 dBA range, depending on the brand and size, [1].

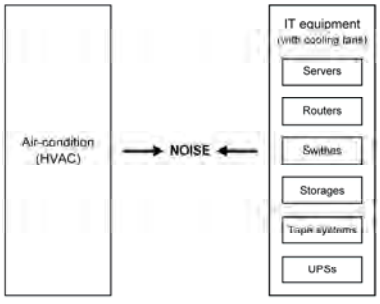


Figure 1. Main sources of data center noise

The sound power (not SPL) of centrifugal and axial fans (such are used in HVAC ventilation ducts) is given by (1), [2]:

$$L_W = K_W + 10\log_{10} Q + 20\log_{10} P + BFI + C_N \quad (1)$$

where

- L_W sound power level
- K_W specific sound power level depending on a type of fan (obtained from manufacturer data)
- Q volume flow rate
- P total pressure (inches of H_2O)
- BFI blade frequency increment, correction for pure tone (obtained from the graphs)
- C_N efficiency correction (out of optimum flow conditions, obtained in tables)

When there are hundreds of servers, with hundreds of cooling fans, noise levels are considerably higher than in the case of one server. Considering them as noise sources that are not coherent, level difference is given by (2):

$$\Delta L = 10 \times \log n \quad (2)$$

where

ΔL is the level difference
n is the number of equal loud sound sources

Permissible Exposure Time for various noise levels (according to NIOSH) may be calculated using (3), [3].

$$t = 480 / 2^{(L - 85)/3} \quad (3)$$

where

t is maximum exposure duration (seconds)
L is noise exposure level (dBA)
3 is exchange rate (dBA), i.e. for noise level increase of 3 dBA, permissible exposure time is cut in half

Permissible exposure time for noise - guidelines for level and duration (time) are shown in Table I, [4]. Noise exposure level/times exceeding those shown in table require use of hearing protection. When considering working conditions for technicians and visitors in a data center, high noise levels influence not only hearing and communication but negatively influence concentration needed for performing complex cognitive tasks. Some kind of protection may be needed, especially for workers working long shifts to increase concentration, comfort and

TABLE I. PERMISSIBLE TIME EXPOSURE

Level, dBA	85	88	90	92	94	95	100	105	110	115
OSHA PEL	16		8			4	2	1	0,5	0,25
NIOSH PEL	8	4			1		0,25			

Duration (in hours) of allowable exposure based on OSHA and NIOSH criteria. PEL = Permissible Exposure Limit; REL = Recommended Exposure Limit

prevent undesirable extra-aural effects (influence on the autonomous system), [5]. When noise levels exceed 85 dBA, hearing conservation programs are required. This includes baseline audiometric testing, noise level monitoring or dosimetry, noise hazard signage, education and training, [6]. Once levels exceed 87 dBA (in Europe) or 90 dBA (in the US), further measures are required like use of hearing protection, rotation of employees, or engineering controls must be taken, [6].

III. SPEECH COMMUNICATION UNDER NOISY CONDITIONS

During various installation and maintenance works communication between staff is of utter importance. Normal conversation tends to occupy 60-70 dBA range. Speech communication in data center is degraded by the masking effect of the background noise and changes in vocal effort are necessary for various background noise levels, [7], as shown in Fig. 2, [8], and Table II.

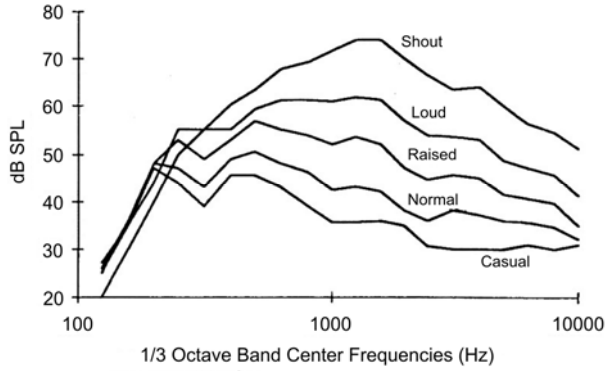


Figure 2. Long-term spectrum of speech under various vocal efforts (adopted from [8])

TABLE II. SPEECH LEVELS AT VARIOUS VOCAL EFFORTS

Voice	Average level dB/dBA
Casual	52.0/42.0
Normal	57.0/47.0 (private speech)
Raised	64.0/57.0
Loud	73.0/62.0
Shout	85.0/72.0

Auditory masking is intrusion of unwanted sounds that interfere with the speech signal. Masking effect of data center noise in spectral domain is illustrated in Fig. 3-5. When the low-frequency noise is louder than the speech signal it effectively masks speech. At high sound pressure levels such noise effectively masks both vowels and consonants. High-frequency noise masks only the consonants, and its masking effectiveness decreases as the noise gets louder, [9]. Noise exposure levels in a data center afford less than desired intelligibility (< 95%).

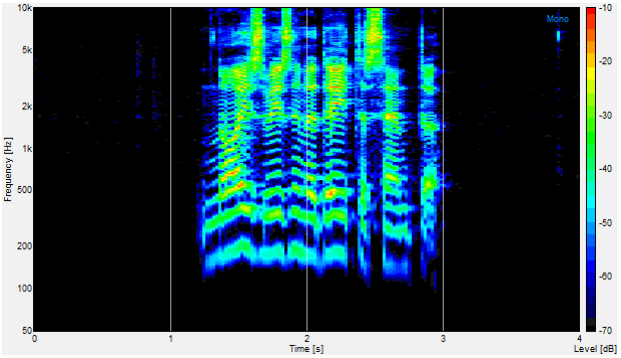


Figure 3. Spectrogram of speech ('Noise within a data center')

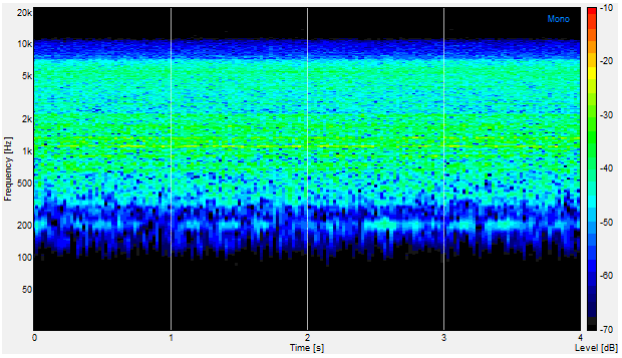


Figure 4. Spectrogram of data center noise

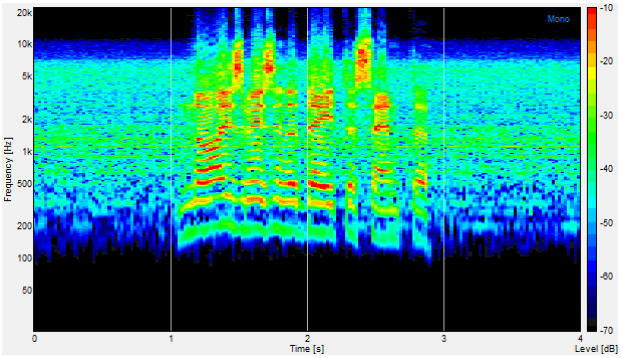


Figure 5. Spectrogram speech plus data center noise combined

IV. DATA CENTER NOISE MEASUREMENTS

Noise measurements were performed in mid-size corporate data center, of surface area about 170 m^2. IT equipment installed within a data center is listed in Table III. Temperature inside a data center is kept in 21-22 °C range with separate cabinet cooling systems blowing out air at 18-19 °C.

TABLE III. IT EQUIPMENT WITHIN A DATA CENTER

Type	Number of devices
Server (physical)	120
Storage	5
Tape subsystem	3
Routers/ Switches	15
Cabinet Cooling System	10

A. Measuring Equipment

Noise levels were measured using the CEM DR-805 Sound Meter. A-level weighting was used due to high correlation with people's subjective judgment of the loudness. Noise signals were recorded at 44.1 kHz sampling frequency and 16-bit resolution, using simple Ednet Desktop Microphone with 30 Hz-16 kHz range, HP ProBook 6570b laptop computer with internal audio card and YMEC sound measurement and analyzing software.

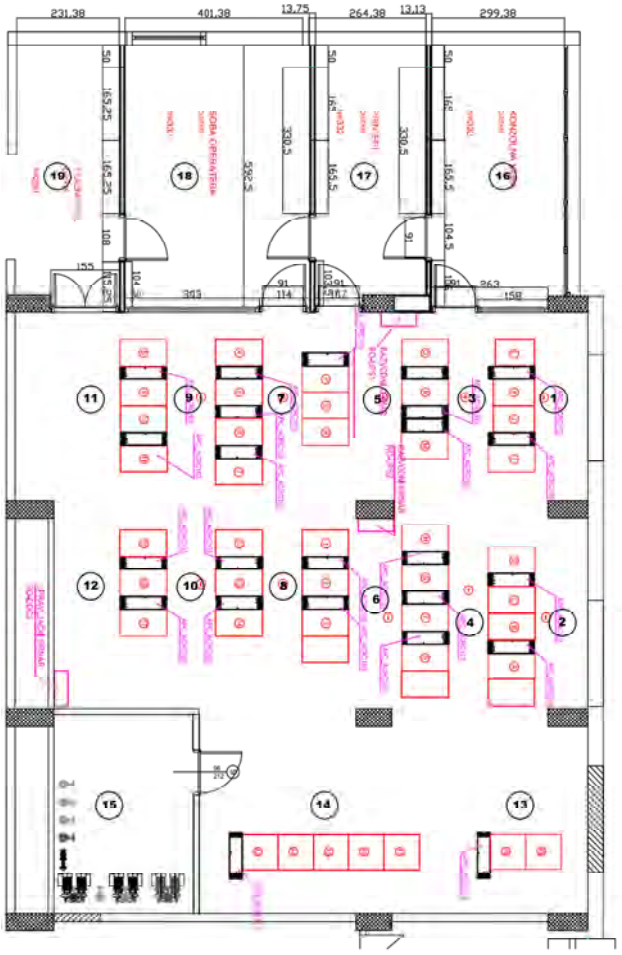


Figure 6. Data center schematic

B. *Measured Noise Levels*

Noise was measured at 19 different places, 15 in a data center and 4 in accompanying spaces (console room, archive, operators room and at the entrance hall) shown in data center schematic, Fig. 6. All noise measurements were performed at the head level while in standing position. Measurement results are shown in Table IV. Noise level depends if it is measured in a cold or hot aisle. The lowest level in a data center was around 70 dBA and the highest slightly above 80 dBA. Similar values are common in data centers, [10]. Higher levels were measured in hot aisles (where cooling air gets out of servers). The highest level was measured near routers with particularly noisy cooling fans. These levels are bellow maximum acceptable levels. However repeated exposure to noise levels between 75 and 85 dBA may pose a small risk to some people, [11]. Being so close to the limit and considering excess risk of long term exposure, [4], it may be wise to take no chances and wear some hearing protection. Noise in accompanying spaces was much lower (under 60 dBA). Although this is much less than in data center and staff got used to it, prolonged stay is still quite unpleasant for unaccustomed visitors.

C. *Noise Spectrum*

Spectrum of data center noise is shown in Fig. 7. Please note harmonic content at frequency range from 150 Hz up to 1.5 kHz that originates from air condition

TABLE IV. NOISE LEVELS AT VARIOUS POSITIONS

Position	Description, contributing sources in aisle	dBA
1	Servers, storage, cabinet cooling system	76,2
2	Servers, storage, cabinet cooling system	76,1
3	Servers, storage, cabinet cooling system	79,2
4	Servers, storage, cabinet cooling system	76,9
5	Servers, storage, cabinet cooling system	78,1
6	Servers, cabinet cooling system	79,1
7	Servers, storage, cabinet cooling system	74,2
8	Servers, cabinet cooling system	75,0
9	Servers, cabinet cooling system	74,7
10	Servers, cabinet cooling system	73,9
11	Servers, cabinet cooling system	70,9
12	Servers, cabinet cooling system	70,9
13	Routers	75,2
14	Routers	80,3
15	Fire central (alarm and dischargers)	70,7
16	Console room	57,8
17	Archive	55,6
18	Operators room	58,5
19	Entrance hall	58,3

and IT equipment cooling fans. These fans operate at discrete frequencies, and certain frequencies like blade passing frequency (BPF), have more power than others. *BPF* is the product of the fan rotation speed and the number of fan blades, (4), [2]:

$$BPF = \frac{N_R N_B}{60} \tag{4}$$

where *BPF* is the basic frequency of fan tonal components, N_R is the fan rotation speed (rotations per minute, RPM) and N_B is the number of fan blades. Numerous harmonics are present as multiples of *BPF*. There is also a considerable noise in a frequency range of 2-7 kHz due to the turbulent air flow produced by the fans. Above 7 kHz noise rapidly diminishes.

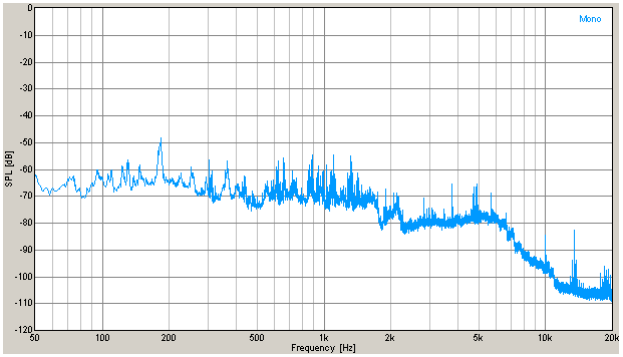


Figure 7. Spectrum of data center noise

D. *Daily and Weekly Variations*

Data centers have its temperature fluctuations that may indirectly (due to different cooling requirements) influence noise levels. This includes:

- daily variation caused by daily swings in the IT load and outdoor temperature
- day-to-day variation caused by the weather
- IT load reduction on weekends

Daily noise variations measured at position 3 in data center schematic from Fig. 6, are shown in Fig. 8.

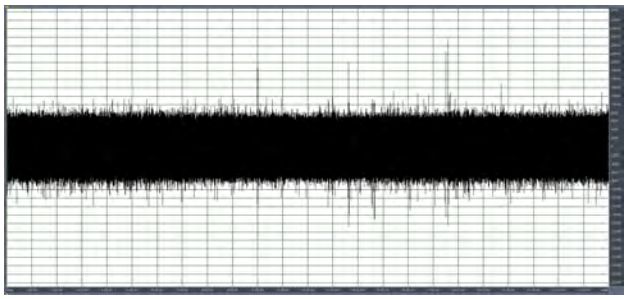


Figure 8. Daily noise variations (linear amplitude scale, working day)

Noise level is quite constant with occasional spikes due to human influence (opening data center doors, opening server cabinets, loud communication etc.). Noise levels averaged in four hour periods are shown in Table V. Summary of daily variations, but averaged over 5s period are given in Table VI. Weekly noise variations of daily averages are given in Table VII.

TABLE V. DAILY NOISE VARIATIONS

Noise level (4 hour average) dBA	0-4	4-8	8-12	12-16	16-20	20-24
	79,4	79,4	79,3	79,3	79,4	79,4

TABLE VI. DAILY NOISE VARIATIONS – SUMMARY

Noise level (5s average) dBA	min	max	aver
	78.5	88.3	79.4

TABLE VII. WEEKLY NOISE VARIATIONS – SUMMARY

Noise (daily average) dBA	Sun	Mon	Thu	Wen	Thu	Fri	Sat
	79,3	79,3	79,4	79,5	79,4	79,4	79,4

Data center where the measurements have taken place is working with low utilization of computing resources and main contributor to temperature changes is the weather alone. As can be seen from our particular case, noise variation during a period of one day and week are negligible. In some data centers temperature fluctuations may be up to 9 ºC depending on the time of day, season and the current weather. Much lower temperature excursions (just few degrees) are common in most data centers.

V. MEASURED NOISE LEVELS, SPEECH INTELIGIBILLITY AND COMMUNICATION DISTANCE

Several noise metrics have evolved for assessing the influence of noise on speech, [12].

A. Speech Interference Level (SIL)

Speech Interference Level is defined as the arithmetic average of the sound pressure levels at 500, 1000, 2000 and 4000 Hz octave bands, (5), [12, 13].

$$SIL = \frac{L_{p500} + L_{p1000} + L_{p2000} + L_{p4000}}{4} \quad (5)$$

A-weighted sound level *Lpa* correlate well with SIL for most sounds in aviation, [2]. Acceptable results of SIL values maybe derived from A-weighted noise levels by using the following approximate expression, (6), [13]:

$$SIL = L_{pA} - 10 \quad (6)$$

For data center noise spectral content shown in Fig. 7 and put into (5), more appropriate expression would be (7):

$$SIL = L_{pA} - 12{,}5 \quad (7)$$

B. Articulation Index (AI)

Articulation Index (AI) is the value, between 0 and 1.0, which describes the masking of speech by background noise. An AI of 1 means that all speech can be understood, and 0 that no speech can be understood. An AI < 0,05 means very poor speech intelligibility, and an AI > 0.80 good speech intelligibility. AI is found by evaluating the signal to noise ratio in specific frequency bands, [12], according to (8), [14].

$$AI = \frac{\sum_{i=1}^{n} I_i (SNR_i + 12)}{30} \quad (8)$$

where

AI is the Articulation Index
I_i is the frequency importance function for band i
SNR is the Signal-to-Noise Ratio for band i

AI can be calculated from the 1/3 octave band levels between 200 Hz and 6300 Hz center frequencies, using the AI calculator, shown in Fig. 9, [15].

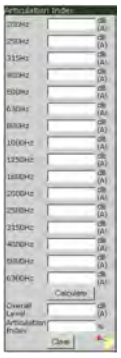


Figure 9. AI Calculator

C. Speech Intelligibility Index (SII)

The Speech Intelligibility Index (SII), [12, 16, 17], is a standardized objective measure which is correlated with the intelligibility of speech under a variety of adverse listening conditions. SII is a function of the long-term average spectrum of the speech and noise signals and is based on the Articulation Index. SII and the AI are not the same thing, but are quite similar. The SII, like the AI, is a quantification of the proportion of speech information that is both audible and usable for a listener. The SII is computed as a product of the frequency band importance function I_i, and the band audibility function A_i, (9).

$$SII = \sum_{i=1}^{n} I_i A_i \quad (9)$$

The value of the SII varies from 0 (completely unintelligible) to 1 (perfect intelligibility). SSI can be determined using the SII Calculator, Fig. 10, [17].

Speech intelligibility in a data center is determined according to mentioned intelligibility measures. Values for SIL were determined using (5) and scaling the noise values. In a similar way values for AI and SII were determined using AI and SII calculator (for SII calculator

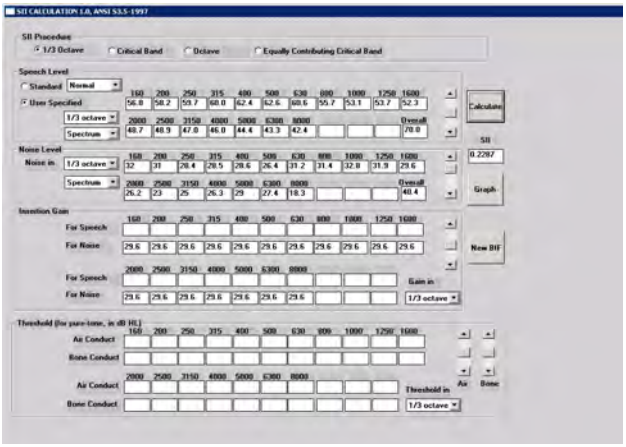


Figure 10. SII Calculator

user specified male speech spectrum normalized to 70 dB SPL was supplied). Speech intelligibility values for various noise levels are listed in Table VIII. At levels within a data center, staff have to talk loudly to be heard.

D. Communication Distance

The distance between the talker and listener (i.e. communication distance) is important when the conversation takes place. Speech levels are reduced typically by 6 dB for each distance doubling between the talker and listener. Communication distances for various values of SIL are shown in Fig. 11, [18], and for various noise levels encountered in a data center in Table VIII.

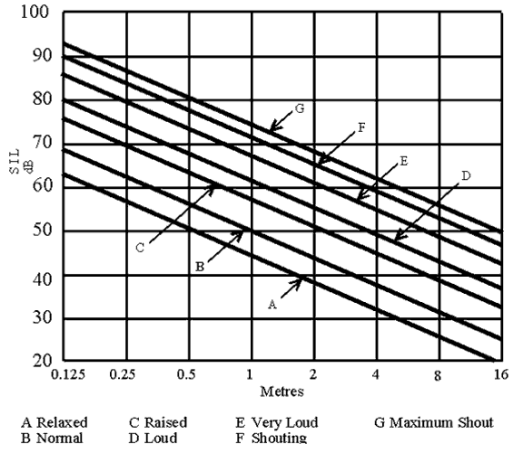


Figure 11. Communication distance for various values of SIL (adopted from [18])

TABLE VIII NOISE LEVELS, SPEECH INTELLIGIBILITY MEASURES AND COMMUNICATION DISTANCES

Noise level, dBA	Speech intelligibility[1]			Communication distance (m)	
	SIL	AI	SII	normal speech	loud speech
55[2]	42,5	0,82	0,66	2,6	9,5
60[2]	47,5	0,66	0,53	1,2	4,5
70	57,5	0,34	0,23	0,44	1,6
72	59,5	0,28	0,17	0,35	1,2
74	61,5	0,22	0,12	0,29	1,0
76	63,5	0,17	0,08	0,23	0,76
78	65,5	0,11	0,06	0,18	0,66
80	67,5	0,06	0,04	0,14	0,53
82	69,5	0,03	0,03	0,13	0,44
84	71,5	0,01	0,02	0,10	0,34

[1]normal speech, [2]noise levels at accompanying spaces

VI. POSSIBLE NOISE REDUCTION SOLUTIONS

A. Passive Noise Reduction

Data center with its walls, floors and ceiling surfaces provide little acoustic absorption. Conventional passive acoustic absorption materials are not used because they are highly flammable and may emit particulates that may trigger fire detection sensors, [19]. Materials used in a data center must be properly fire rated and fiber free. They can be mounted on walls and suspended from ceilings preventing reflections, reducing reverberation and preventing noise build up. It is also possible to use sound dampening server cabinets that use acoustic foam.

Passive noise protection headphones may be used for tasks where no communication is needed. Conventional hearing protection or noise protection headphones restrict the workers ability to communicate. If communication among workers is needed, headphones with integrated wireless communication (like Bluetooth) could be used.

B. Active Noise Control

Noise within a data center consists of tonal and broadband components. Tonal components originate from HVAC fans and equipment cooling fans with pronounced discrete spectrum. Broadband component originates from turbulent airflow. Active Noise Control (ANC) is quite successful at eliminating tonal components. Eliminating these low to mid frequency components would greatly reduce the amount of noise present in the data center.

ANC is generally not suitable for noise attenuation in large spaces, but is mostly restricted to generating local zones of silence. ANC over an extended region is not possible in the data center room in general, [20]. This is due to the random nature of the noise and the high modal density of the room itself. If trying to suppress noise within a large space better approach is to attenuate noise at the source itself, like at ends of ventilation ducts.

ANC can be successfully applied for silencing HVAC ventilation ducts, [21, 22], Fig. 12. Ventilation ducts contain low to mid frequency tonal and broadband noise. ANC systems for silencing ventilation ducts have been developed decades ago, but with reduced cost of signal processing hardware are becoming feasible for common use. In laboratory, ANC can achieve 15-25 dB additional noise reduction at frequency range of 50-350 Hz, [21], and values of 10-15 dB are reported in a real world, [22]. Specially designed add-on unit can be installed at the end of a pipe emitting or sucking air and generating noise.

ANC within servers, integrated with cooling fans, is something to be common in a future. Such system uses a magnetic field to generate minute vibrations in fan blades. A microphone samples the noise and then tunes the fan blade vibration to cancel it out, [23].

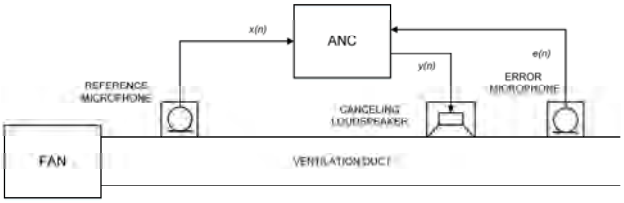


Figure 12. ANC in a ventilation duct

ANC headphones, [24], can provide greater deal of protection, however most of ANC headphone designs are optimized for aircraft use. Although ANC headphones are successful at reducing noise, they also reduce the ability to hear speech and remain aware of surroundings. Some communication headphones combine speech enhancement with noise suppression technology that elevates speech and suppresses dangerous background noise so workers can hear speech and stay protected in high noise environments, [25]. Sometimes even situational awareness with speech and noise direction recognition is preserved.

C. *Lowering Cooling Requirements*

HVAC is the main source of noise in a data center. According to [6], 20% increase in cooling air speed equates to a 4 dB increase in noise level. By reducing HVAC activity it is possible to reduce noise.

Lower cooling requirements on HVAC can be achieved by implementing higher degree of virtualization and hence achieving higher CPU usage and the lower number of necessary physical servers. Overall power consumption, cooling requirements and noise levels go down. However efficiency of power and cooling infrastructure gets worse because of underloading, [26].

Another approach is raising target temperature within a data center, [27]. The higher the target temperature in a data center, the lower is the cooling requirement. HVAC unit should not necessary run at full capacity at all times. Increasing target temperature for just few degrees during maintenance activities would lead to a quieter work environment for employees. Recommendations for air temperatures in data centers (Class 1 requirements) put them in a range 18-27 °C (allowable 15-32 °C), [28]. As an approximation it may be expected that 2°C increase in data center temperatures would cause noise decrease of 3-5 dB, [6]. Server fans will, however, try to compensate for diminished HVAC operation using the server built-in temperature-based fan speed control, adding some noise.

VII. CONCLUSION

Measured noise levels within a data center were in 70-80 dBA range (as in other data centers). These levels are below maximal acceptable eight hour exposure limit but are close enough and it may be wise to use some kind of hearing protection like noise protection headphones. Prolonged exposure to such noise levels could produce extra-aural health effects. High noise levels interfere with performing cognitive tasks, influence autonomous system and after the long-term exposure may cause behavior modifications. Noise in a data center significantly interferes with speech intelligibility (as reflected in speech intelligibility measures), making communication among personnel difficult (raising voice and reducing communication distance). Lightweight ANC headphones with build in wireless interphones would be beneficial. Staff should perform as much as possible tasks remotely from the office, using iLO console or RD connections. Passive noise reduction with certified fire resistant materials may be applied to data center surfaces. ANC is suitable for HVAC ventilation ducts. During extensive maintenance activities temperature throttling can be used.

REFERENCES

[1] TechMikeNY, Server Guide, Internet: https://techmikeny.com/server-guide/ [February 29, 2016]

[2] Fan Noise Prediction, Internet: http://www.mne.psu.edu/lamancusa/me458/11_fan.pdf [February 29, 2016]

[3] Noise Exposure - Permissible Level and Duration, Internet: http://www.engineeringtoolbox.com/noise-exposure-level-duration-d_717.html [February 29, 2016]

[4] Patricia T. Johnson, Noise Exposure: Explanation of OSHA and NIOSH Safe-Exposure Limits and the Importance of Noise Dosimetry, ER078114 –B, Etymotic Research

[5] Sharp L. F., Swiney J. F., Dansby M. R., Hyatt S. C. & Schimmel D. E.: Behavioral and physiological correlates of varying noise environments, EPA-600/1-77-038, EPA, June 1977

[6] ASHRAE Publication, 2008 ASHRAE Environmental Guidelines for Datacom Equipment, ASHRAE, 2008

[7] Pearsons K., Bennett Ricarda L. & Fidell S.: Speech levels in various noise environments, EPA-600/1-77-025, EPA, May 1997

[8] P. H. Stypulkowski, "Circuitry options, signal processing under altered signal conditions", in Tobin, H (ed.): Practical Hearing Aid Selection and Fitting, DIANE Publishing, 1997

[9] R. Jones, Factors That Affect Intelligibility in Sound Systems, Meyer Sound, Internet: http://www.meyersound.com/support/papers/speech/section2.htm [February 29, 2016]

[10] D. Alger, DC Deconstructed | Sounds of a Data Center, Cisco, Internet: https://www.youtube.com/watch?v=EQTgpE35zSM&nohtml5 [April 6 2015]

[11] National Standard for Occupational Noise [NOHSC: 1007(2000)], 2nd Ed., Canberra, Australia, July 2000

[12] Newman J. S. & Beattie K. R., Aviation Noise Effects, Report No. FAA-EE-85-2, US DoT FAA, March 1985

[13] D. Miljković, J. Ivošević and T. Bucak, "Psycho-acoustical ergonomics in a light aircraft interior", Proc. of Ergonomics 2013, pp. 111-116, 12-15 June 2013, Zadar, Croatia

[14] H. Bächler and A. Vonlanthen, "Signal processing tools to support your manner of communication", Phonak Focus 17, 1994

[15] Articulation Index, Internet: http://www.diracdelta.co.uk/science/source/a/r/articulation%20index/source.html [February 29, 2016]

[16] American National Standard ANSI S3.5-1997, Methods for Calculation of the Speech Intelligibility Index

[17] SII CALCULATION 1.0, A Software for ANSI S3.5-1997

[18] Transportation Safety Board of Canada (TSB), Railway Investigation Report R98V0148, [February 29, 2016]

[19] Y. Joshi and P. Kumar (ed.), Energy Efficient Thermal Management of Data Centers, Springer, 2012

[20] J. Daily, J. Esplin, Z. Collins and M. Shaw, Acoustical Analysis of Active Control in the Server Room of a C7 Data Centers Colocation Facility Feasibility Report, Brigham Young Univ., June 21, 2010

[21] M. Larsson, Active Noise Control in Ventilation Systems, Blekinge Institute of Technology Licentiate Dissertation Series No 2008:12, Karlskrona, Sweden, 2008

[22] L. J. Gelin, "Active noise control: a tutorial for HVAC designers", ASHRAE Journal, August 1997

[23] G. Gasior, Noise-canceling Noctua CPU Cooler Coming in 2014, The TechReport, Internet: http://techreport.com/news/24947/noise-canceling-noctua-cpu-cooler-coming-in-2014 [February 29, 2016]

[24] C. Moy, Active Noise Reduction Headphone System, Internet: http://cero-ruido.blogspot.com/2014/11/active-noise-reduction-headphone-systems.html [February 29, 2016]

[25] Sensear, Data Center Industry Case Study, 2014

[26] Suzanne Niles, "Virtualization: ooptimized power and cooling to maximize benefits", White Paper #118, Schneider Electric, 2011

[27] To Lower the Noise Levels in Data Centers, Raise the Temperature, Internet: http://www.itwatchdogs.com/environmental-monitoring-news/data-center/to-lower-the-noise-levels-in-data-centers,-raise-the-temperature-617409 [February 29, 2016]

[28] 2011 Thermal Guidelines for Data Processing Environments - Expanded Data Center Classes and Usage Guidance, ASHRAE TC 9.9, ASHRAE, 2011

Active Noise Control: From Analog to Digital – Last 80 Years

Dubravko Miljković
Hrvatska elektroprivreda, Zagreb, Croatia
dubravko.miljkovic@hep.hr

Abstract - Beginnings in the field of active noise control date from 1930s. From these pioneer attempts great deal of progress has been achieved. In this brief review paper advances from simple fixed analog systems, manually adaptive analog systems, acoustic feedback neutralization, fully adaptive digital systems with secondary path modeling and multichannel systems are briefly presented. Applications of active noise control in closed and open space, industrial and automotive applications, as well as active noise canceling headphones and new solutions with smart materials are also briefly mentioned.

I. INTRODUCTION

Active noise cancelation (ANC) is not quite a new concept. The first experiment on superposition of sound fields was supposedly made in 1878 by Lord Rayleigh using two electromechanically synchronized tuning forks, [1]. First humble beginnings that specifically targeted cancelation of unwanted noise exploiting the destructive interference of sound started in 1930s, [2, 3, 4]. Although this principle is quite simple, its practical implementations in real word encountered numerous problems that had to be solved. Early systems were analog, with no or manual adaptation that could only partially benefit from theoretical ideal. Later, with the advent of digital signal processing quite sophisticated adaptive algorithms for ANC have been developed. As with any other technical field there was a hype cycle with initial exaggerated expectations, followed by the bitter disillusionment, then by enlightenment and finally reaching the plateau of productivity. Early unrealistic expectations included possibility of silencing large spaces. Today technology is quite successful in forming smaller zones of silence around the head or even smaller spaces as in ANC headphones. Silencing of larger spaces is much more difficult to deal with, with very limited success even today.

II. BASIC PRINCIPLE

Active noise cancelation applies principle of destructive interference with the introduction of canceling acoustic signal that is of the same amplitude but opposite phase of the unwanted noise, Fig. 1. Resulting destructive interference lowers the noise levels in a protected space. It is of paramount importance that canceling signal has exactly the same amplitude and exactly the opposite phase. This is not always easy to achieve, particularly in nonstationary signals. Even small discrepancies between amplitudes and phases of original noise and canceling signal severely diminish efficiency of the method, [4].

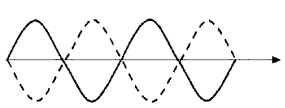


Figure 1. Principle of active noise cancelation

III. ANALOG SYSTEMS

In early years of ANC, all solutions had to be implemented using analog technology. The first patents explicitly oriented on the idea of active noise control by destructive interference with the sound wave of opposite phase have been granted to Henri Coanda, [2], in 1934, however Coanda's concept of realizing his objectives were technically incorrect. Few weeks later Paul Lueg also applied for a patent of similar goals, [2, 3].

A. Lueg's Patent

The design of acoustic ANC utilizing a microphone and an electronically driven loudspeaker was first proposed in 1936 patent by Lueg [2, 3], Fig. 2, [3]. Block diagram of his system with additional legends is shown in Fig. 3. The Lueg system is a monopole consisting of a

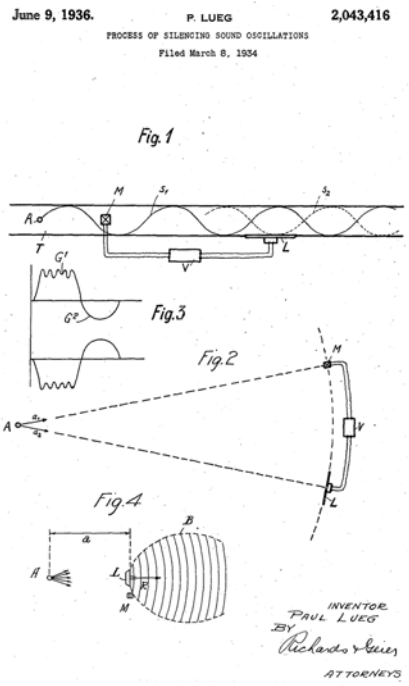


Figure 2. Lueg's patent

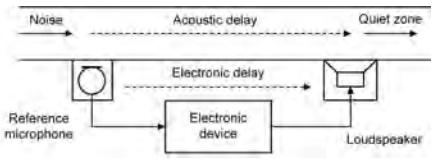


Figure 3. Lueg's system

microphone, amplifier and loudspeaker (secondary source). An acoustic noise is picked up by a microphone, whose signal is processed through an electronic device to create cancelling signal at a loudspeaker that opposes to the primary noise, [5]. Electronic device includes phase inverter with a time delay. Signal is further amplified and led to the loudspeaker disposed downstream from the microphone. At the position of the loudspeaker noise and "antinoise" cancel each other and form a quite zone. This type of system is called feedforward. It senses noise before it passes secondary source. Unfortunately Lueg was never able to demonstrate his idea successfully because it was oversimplified and field of electronics was not sufficiently advanced, [6].

B. Patent by Olson and May

In 1953 Olson and May carried their researches on developing an electronic sound absorber, which appeared to be successful over small volumes in a unidirectional sound field, [2, 4, 7, 8], Fig. 4. This type of system is called feedback system, Fig. 5. It cancels noise without the benefit of an upstream reference input.

Previously described systems were not adaptive.

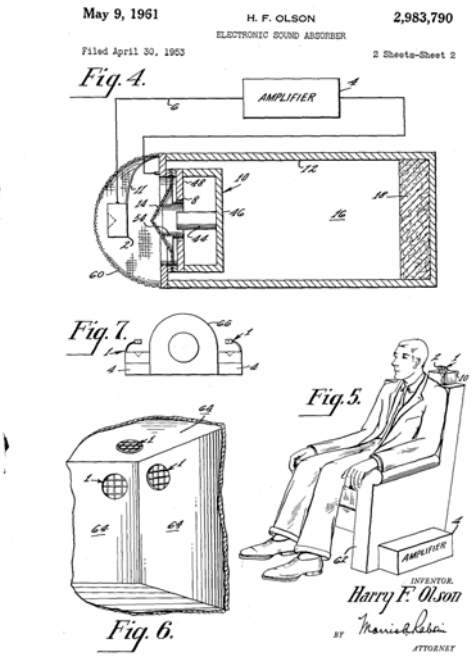


Figure 4. Olson and May active absorber

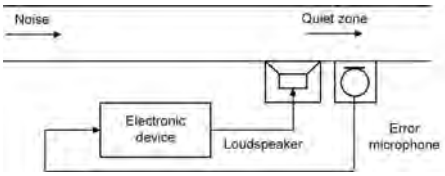


Figure 5. Feedback system (analog)

C. Conover System

In 1956, Conover described an active system for reducing noise radiated by large transformers [2, 4, 7, 9, 10], Fig. 6. It consisted of three parallel independent channels, realized with three band pass filters, as can be seen in a block diagram shown in Fig. 7.

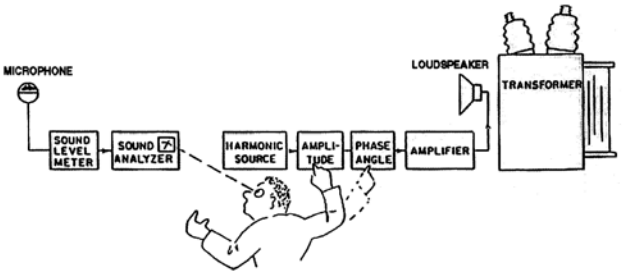


Figure 6. Manually adaptive, feedforward active noise reduction system for transformer noise proposed by Conover in 1956, from [10]

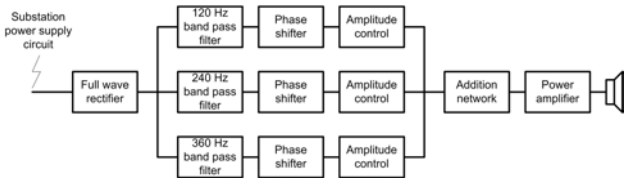


Figure 7. Block diagram of Conover's analog system

Central frequencies of these filters correspond to the main frequency emitted by a transformer and its first two harmonics. The main frequency of noise is twice the electric grid (line) frequency because transformer magnetic core contracts twice with each period of grid frequency. To achieve precise canceling signal Conover applied manual adjustment of phase and amplitude for each of three frequency components of the canceling signal. Later, in 1968 Onoda and Kido developed an automated system for transformer's noise reducing, [11].

Very detailed review of patents on ANC is given in [2].

IV. ADDRESSING FEEDBACK IN ANALOG SYSTEMS

In ANC systems there exists an acoustic feedback (hallowing) between the canceling source and reference microphone causing instability in the system. Canceling signal should be introduced in such a way that over the intended frequency range it propagates mainly in the direction of original noise. The influence of acoustic feedback could be diminished by the use of directional microphones and speakers as well as with specific loudspeaker and microphone arrangements. Most popular geometric arrangements are Chelsea dipole, Swinbanks two-element unidirectional source and Jessel-Mangiante-Canévet tripole source, [4, 5, 12].

A. Chelsa Dipole

The Chelsea dipole, developed in 1976, Fig. 8, uses a pair of loudspeakers driven out of phase and physically spaced one-half wavelength apart around the microphone situated centrally between them, [5]. The sound radiated from secondary sources cancels each other at the position of the microphone, while simultaneously providing noise attenuation in a region away from the noise source.

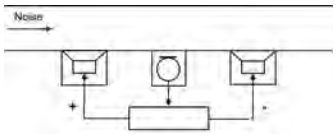


Figure 8. Chelsa dipole

B. *The Swinsbank System*

The Swinsbank system, developed in 1973, Fig. 9, uses a pair of loudspeakers that are arranged in such a way that delay and phase between the sources produces cancellation in the upstream direction and addition in the downstream direction, thus reducing the feedback from the loudspeakers to the sensing microphone, [5]. It employs electrical delay between the two out-of-phase loudspeakers to be equal to the propagation time between the speakers, [4, 5, 12].

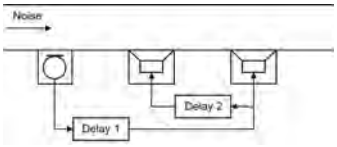


Figure 9. Swinbanks system

C. *Jessel-Mangiante-Canévet Tripole*

The Jessel-Mangiante-Canévet tripole system, from 1972, Fig. 10, uses a monopole source to cancel upstream propagation from the dipole source, eliminate acoustic feedback and produce unidirectional source, [5, 12].

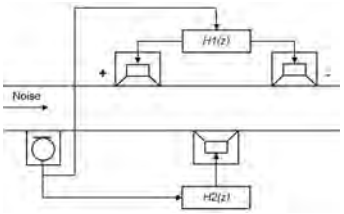


Figure 10. Jessel-Mangiante-Canévet tripole

Fixed filter systems have basic limitations. They use secondary source spacing based on the wavelength and are limited to a specific frequency range and sound speed that vary with the temperature and humidity, [5]. To solve this problem it is necessary to introduce into a system some kind of adaptation. This became possible with the advent of digital technology. Due to complexity of adaptive algorithms, practically all adaptive filters are digital.

V. DIGITAL SYSTEMS

The need for very accurate phase and amplitude of canceling signal leads toward application of the digital signal processing, (DSP) that become popular in 1980s with the introduction of inexpensive hardware (first DSP chips by Intel, AMI, NEC and TI). Use of adaptive digital filters for ANC was first introduced by Burgess [13]. An adaptive filter, Fig. 11, is a system with a linear filter that has a transfer function controlled by variable parameters and a means to adjust those according to an optimization algorithm, [14]. Output of the filter *y(n)* is given by (1), where w_i are filter coefficients and *x(n)* is the input signal.

$$y(n)=\sum_{i=0}^{N-1} w_i x(n-i) \tag{1}$$

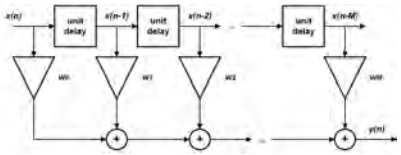


Figure 11. FIR (Finite Impulse Response) filter

The filter can be adapted by varying its weight factors by some algorithm, of which the most popular is the LMS algorithm, [14]. The LMS algorithm was introduced by Widrow and Hoff in 1959 as an adaptive algorithm, which uses a gradient-based method of steepest decent. Adaptation of filter coefficients w_i at the moment n+1 is given by (2) and (3), where μ is a step size, *x(n), y(n)*, *d(n)* and *e(n)* are signal, filter output, desired filter output and error at the moment of the sample n.

$$e(n)=d(n)-y(n) \tag{2}$$

$$w_i(n+1)=w_i(n)+\mu x(n)e(n) \tag{3}$$

Similarly to analog systems digital systems can also be feedforward or feedback, however this time used filters are not fixed but adaptive, [15, 16].

A. *Feedforward System*

In a feedforward system the signal from the reference microphone is processed by an adaptive digital filter to produce a cancelling signal, Fig. 12. Essentially it is the system identification approach where the acoustic path is modeled with an adaptive filter, [15].

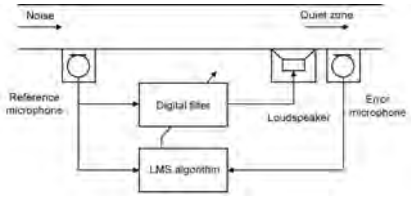


Figure 12. Feedforward system (digital, adaptive)

Real systems face two additional problems, the influence of the secondary path and an acoustic feedback.

Loudspeaker and amplifier that drives it, form the secondary path that has a transfer function *S(z)*. The influence of the secondary path *S(z)* is corrected with the introduction of the filter $\hat{S}(z)$ that filters the input signal *x(t)*, Fig. 13, hence the name Filtered-X LMS algorithm (FXLMS), 1981, [15]. $\hat{S}(z)$ provides also accurate time alignment of referent and error signals that are used by the adaptive algorithm. Accurate modeling of the secondary path is important to avoid the degradation in performance of the ANC system. The secondary path can be estimated off-line or on-line method. FXLMS adaptation is given by (4), [15]:

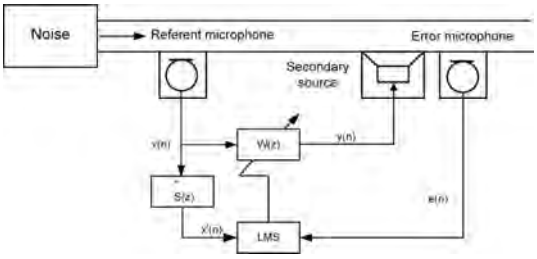


Figure 13. Feedforward system with secondary path modeling

$$w_i(n+1) = w_i(n) + \mu x'(n) e(n) \quad (4)$$

Reduction of feedback distortion is achieved by the introduction of filter $\hat{F}(z)$ for the feedback neutralization which models the feedback path *F(z)*, as shown in Fig. 14.

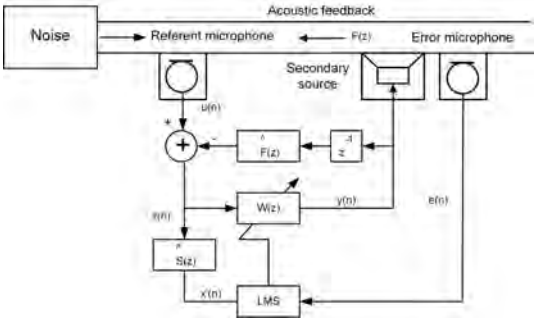


Figure 14. Feedforward system with secondary path modeling and feedback neutralization

B. Feedback Systems

Block diagram of a feedback system is shown in Fig. 15. In the feedback system the signal from the error microphone is processed by an adaptive digital filter to produce a cancelling signal, [15]. Under simplified (idealistic) assumption the secondary path *S(z)* can be approximated with a delay.

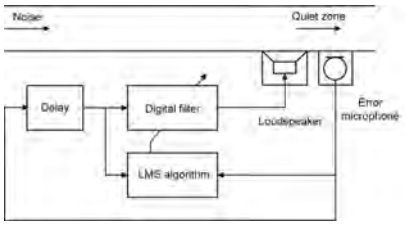


Figure 15. Feedback system as an adaptive predictor (simplified idea)

The feedback system with included secondary path modeling is shown in block diagram in Fig. 16, [15].

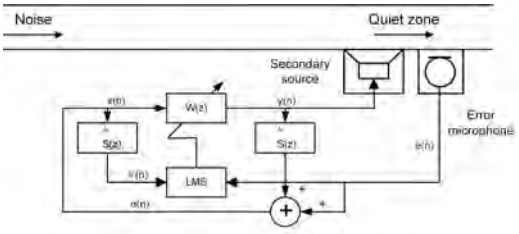


Figure 16. Feedback system with secondary path modeling

C. Synchronized Waveform Generator

Cancellation of periodic noise can be achieved by a synchronized waveform generator, Fig. 17. ANC using a waveform synthesizer was developed in early 1980s by Chaplin and coworkers, [15, 17]. Such system is suitable only for periodic (tonal) noise, present in fans, rotating machinery noise, etc. It uses a waveform generator with stored waveform samples, (5), synchronized by tacho signal (i.e. from non-acoustic sensor) and an adaptive algorithm to further modify period amplitude and phase of generated waveform. Advantage of the method is that a cancelation signal is produced by synthesis and does

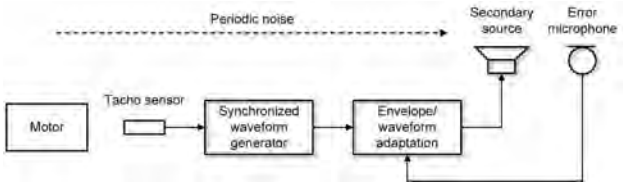


Figure 17. Narrowband feedforward ANC using the synchronized waveform generator

$$\{w_i(n), l = 0,1,\ldots,L-1\} \quad L \text{ is the period length} \quad (5)$$

not depend on a measurement of original noise, hence there is not a problem of unwanted acoustic feedback. A synchronous waveform generator can be considered as a particular case of a feedforward system.

D. Hybrid Systems

A hybrid ANC system, Fig. 18, combines feedforward and feedback control structures. Feedforward structure cancels noise correlated to the reference signal, while feedback structure cancels predictable components of the primary noise not observed by the reference sensor [15].

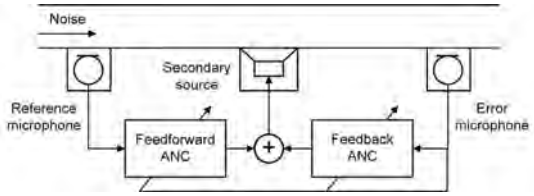


Figure 18. Principle of hybrid ANC system

E. Multichannel Systems

Noise within vehicles cabins has more complicated sound field than in ducts. Minimizing the total acoustic energy in a protected space can be achieved with a multichannel ANC system. To accommodate multiple channels (one or more reference microphones, multiple error microphones and multiple secondary sources), a multichannel system, Fig. 19, for adaptation uses extension of the single channel LMS algorithm, [15].

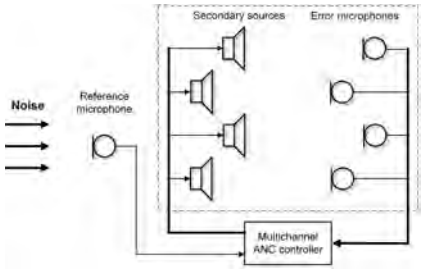


Figure 19. Multichannel system

VI. Applications

The most successful early use of ANC is control of noise in ducts, exhaust pipes, and headphones. Later came ANC in enclosed spaces such as vehicle cabins. These and other applications are listed here, [2, 16].

A. Ventilation Ducts

The active cancellation of 1-D sound fields that approximate sound in ducts has been first suggested by Lueg [3, 4]. It is particularly suitable for HVACs and

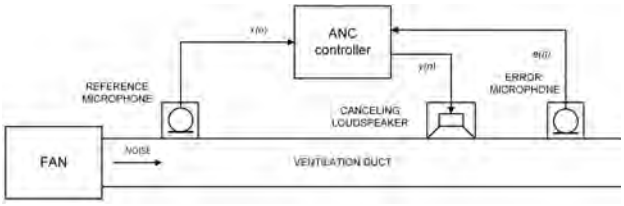


Figure 20. ANC in a duct

went channels in studio and broadcast environments as well for other cases where providing of quiet neighborhood is necessary, [18]. ANC in a duct has an advantage over passive plenum silencers because it doesn't introduce a pressure drop that must be tolerated or compensated by the more powerful fan operation, Fig. 20.

B. *Vehicle Interiors*

ANC has found its application in cars, [19], (first used by Nissan in Bluebird ARX-Z, 1992) and SUVs for the suppression of low frequency rumble in a cabin. Often these solutions use a synchronized waveform generator. Example of a car ANC system is shown in Fig. 21.

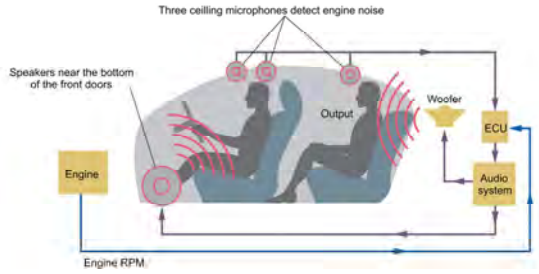


Figure 21. ANC in a vehicle [from Nikkei Electronic/Toyota]

Turboprop aircraft are notoriously noisy, particularly during the take-off and climb phase. The first ANC system has been implemented by Saab in 1994. Multichannel systems are used with numerous loudspeakers and error microphones. Saab 2000 turboprop aircraft uses 32 loudspeakers mounted on trim panels and at foot level to reduce low frequency propeller noise, Fig. 22, [20]. Nice brief review of achievement in a field can be found in [21].

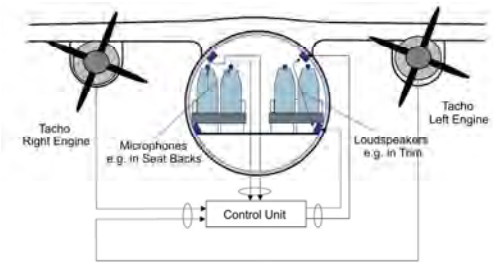


Figure 22. ANC in turboprop aircraft, adopted from [20]

C. *Interior Spaces*

ANC of large spaces is difficult to achieve. Some more success can be achieved by active modal control in rooms placing a secondary source in a corner of a room. Helmholtz resonators (HR) are sometimes used for narrowband passive reduction of noise. Semi-active approach with an adaptively tuned HR adapts to variations in noise while keeping a low energy consumption compared to fully active ANC systems, [22]. A system in Fig. 23 modifies the boundary impedance in the resonator.

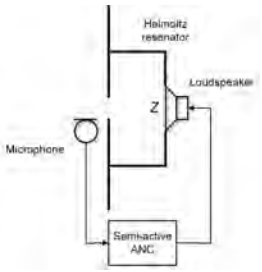


Figure 23. Semi-active ANC with tuned Helmholtz resonator

D. *Open Spaces*

Radiation of noise in an open space can be attenuated by active acoustic barriers, [4, 23], or shields placed around the source, [4]. One type of system for attenuation of large transformer noise is shown in Fig. 24, [24].

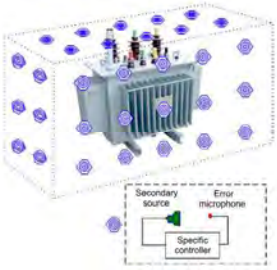


Figure 24. ANC for noise emitted by transformers, adopted from [24]

E. *ANC Headphones*

Earliest ideas about active ear defenders originated in the mid-1950s. Most notable advances in a field of noise cancelling headphones have been made in late 1970s and 1980s by Amar Bose. In 1986, the first wearing prototype headphones that used active noise cancellation has been developed. Few years later, in 1989, the first ever commercially available noise-cancelling headphone has been produced. The cancelling signal is emitted by the loudspeaker within the earcup and the quiet zone is supposed to be located around the ear drum, Fig. 25, [25].

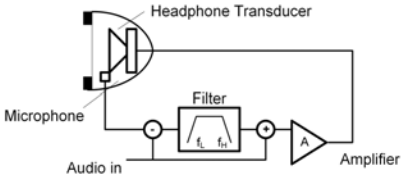


Figure 25. Closed loop ANC headphones

ANC headphones are mainly implemented as analog devices driven by simplicity, small size and cost, [7]. Analog ANC headphones typically achieve an active attenuation of about 20 dB in 100 to 200 Hz frequency range, and no active attenuation bellow about 30 Hz and above about 1 kHz, [26]. Outside this frequency range attenuation depends on used passive methods.

F. *Household Appliances*

Application of ANC in domestic appliances is a promising future field. Experiments have been done with application of ANC in a vacuum cleaner that is known as particularly noisy appliance. Passive means of noise attenuations are not quite suitable as they would disrupt airflow and deteriorate cleaner's performance. ANC setup for a vacuum cleaner is shown in Fig. 26, [27]. Great

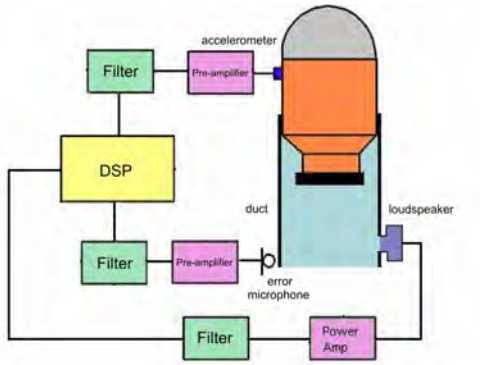


Figure 26. ANC setup for a vacuum cleaner, adopted from [27]

attenuation has been achieved at tonal components (order of 10 dB), however it translated to a much lower value (about 2 dB) over the whole frequency spectrum. Similar approaches are proposed for washing machines.

G. *Smart Materials*

By integrating sensors, electronics and piezoelectric actuators in smart material it is possible to produce ANC surfaces for the reduction of sound radiated from structures. These modern architectures (starting from mid 1990s) use low cost digital signal processors that are integrated directly as distributed architecture into a smart material. Individual elements can be interconnected in an array (or grid). The smart foam as a kind of hybrid active-passive absorbing material for enhancing acoustic absorption at low frequencies is show in Fig. 27, [28].

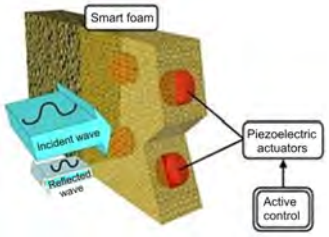


Figure 27. Smart foams, adopted from [28]

VII. CONCLUSION

In this short paper brief review of progress has been covered, from early ideas, analog and later digital implementations. Great progress has been made and with low cost digital hardware active noise control has become a reality for many applications. Main success has been achieved at attenuating low frequencies, particularly tonal components and forming local zones of silence. More modest results have been achieved for silencing larger interiors and open spaces. Early analog systems were nonadaptive or only manually adaptive. Clever geometric arrangements of microphones and loudspeakers were used to neutralize the acoustic feedback. Digital systems with adaptive filters adapt to changing environment conditions, incorporating feedback neutralization and secondary path modeling and achieving much better attenuations. Multichannel systems proved themselves useful for silencing interior spaces of vehicles. Most successful applications today are ventilation ducts, ANC headphones and reduction of vehicle interior noise (turboprop aircrafts, luxury cars and SUVs). In future we may expect to find ANC in some household appliances, like vacuum cleaners. Also, introduction of smart materials (e.g. smart foam) can enable silencing of low frequency noises emitted from larger structural surfaces.

REFERENCES

[1] J. W. S. Rayleigh. The Theory of Sound, Vol. 2, Macmillan, 1896.

[2] D. Guicking, Patents on Active Control of Sound and Vibration - an Overview, May 2009.

[3] P. Lueg, “Process of silencing sound oscillations”, United States Patent 2043416, 8 March 1934.

[4] H. G. Leventhall and L. Wong, A Review of Active Attenuation And Development of an Active Attenuator 'Open Refuge', HSE Contract Research Report No. 4/1988, W S Atkins Engineering Sciences, 1988.

[5] L. L. Beranek, Noise and Vibration Control Engineering – Principles and Applications, John Wiley, 1992.

[6] C. Hansen, S. Snyder, X. Qiu, Laura Brooks and Danielle Moreau, Active Control of Noise and Vibration, 2nd ed., CRC Press, 2012.

[7] L. J. Eriksson, Waves of Silence, Quarter Section Press, 2015, USA

[8] H. F. Olson and E. G. May, “Electronic sound absorber”, J. Acoust. Soc. Am., 25, pp. 1130–1136, November 1953.

[9] W. B. Conover and W. F. M. Gray, “Noise reducing system for transformers”, U. S. Patent 2,776,020, January 1, 1957.

[10] S. J. Elliott and P. A. Nelson, “Active noise control”, IEEE Signal Processing Magazine, Oct 1993, pp. 12-35

[11] S. Onoda, K. Kido, “Automatic control of stationary noise by means of directivity synthesis”, 6th International Congress on Acoustics, Tokyo, Japan, 21—28 August 1968.

[12] A. Barber, Handbook of Noise and Vibration Control, 6th ed., Elsevier 1992.

[13] J. C. Burgess, “Active adaptive sound control in a duct: a computer simulation”, J. Acoust. Soc. Am., 70 (3), pp. 715–725, 1981.

[14] B. Widrow and S. D. Stearns, Adaptive Signal Processing, Prentice Hall, Englewood Cliffs, 1985.

[15] S. M. Kuo and D. R. Morgan, “Active Noise Control: A Tutorial Review”, Proceedings of the IEEE, Vol. 87, No. 6, June 1999.

[16] D. Miljković, “Active noise control”, Proceedings MIPRO 2007 - Vol. III, CTS & CIS, MIPRO, 21-25 May 2007, Opatija, Croatia

[17] G. B. B. Chaplin and R. A. Smith, “Waveform synthesis-the Essex solution to repetitive noise and vibration”, Proc. Inter-Noise 83, pp. 399–402, 1983.

[18] M. Larsson, Active Noise Control in Ventilation Systems, Blekinge Institute of Technology Licentiate Dissertation Series No 2008:12, Karlskrona, Sweden, 2008.

[19] S. J. Elliott, I. M. Stothers, P. A. Nelson, A. M. McDonald, D. C. Quinn, T. Saunders, “The active control of engine noise inside cars”, Proc. Inter-Noise, pp. 987–990, Avignon, France, 1988.

[20] S. Johansson, Active Control of Propeller Induced Noise in Aircraft, PhD Thesis, Blekinge Institute of Technology, Ronneby, Sweden, December 2000.

[21] G. Billoud, Active Control at Lord Corporation – A Reality, Technical Report LL-6508, Lord Corporation, 2001.

[22] C. J. Radcliffe and C. Birdsong, “An electronically tunable resonator for noise control”, SAE Technical Paper 2001-01-1615

[23] J. Guo and J. Pan, “Application of active noise control to noise barriers”, Acoustic Australia, Vo 25, No. 1, pp. 11-16, 1997.

[24] X. Qiu, J. Lu and J. Pan, “A new era for applications of active noise control”, Proc. Inter-Noise 2014, Melburne, Australia, 16-19 November 2014.

[25] C. Moy, Active Noise Reduction Headphone System, http://cero-ruido.blogspot.com/2014/11/active-noise-reduction-headphone-systems.html [February 29, 2016]

[26] L. Håkansson, S. Johansson, M. Dahl, P. Sjösten and Ingvar Claesson, Noise Canceling Headsets for Speech Communication, in, Noise Reductions for Speech Applications, CRC Press, 2002.

[27] R. Paurobally, “Active control of noise from a vacuum cleaner”, Proc. of ICSV13, Viena, Austria, July 2-6, 2006.

[28] P. Leroy “Smart foams for enhancing acoustic absorption”, Acoustics'08, Paris, June 29 – July 4, 2008.

Responding to Stakeholders' Resistance to Change in Software Projects – A Literature Review

S. L. R. Vrhovec*,**
* University of Maribor, Faculty of Criminal Justice and Security, Ljubljana, Slovenia
** University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
simon.vrhovec@fvv.uni-mb.si

Abstract - Software project success rates are gradually improving over the last few years. However, this does not apply to large software projects as only 10 percent are successful. One of the most prominent reasons for high failure rates is resistance to change. Resistance to change has been studied for almost seven decades however a simple answer to this issue simply does not exist. In this paper, we present a review of research on resistance to change in software projects. More precisely, we first establish the need for a change in resistance to change terminology and define the term stakeholder resistance. Next, we deal with diagnosing stakeholder resistance as its root causes are often hidden in the complex inter-human relationships and cannot be identified easily. Finally, we present how to deal with stakeholder resistance. Project managers often use a generic method for dealing with stakeholder resistance, such as user participation. When stakeholder resistance actually manifests there is most often no response to it at all from the management. Blindly following a generic method for lowering stakeholder resistance is not efficient as methods need to be adapted to specific situations and address the right root causes. Also, the situation needs to be monitored and responses prepared in case stakeholder resistance manifests itself, persists or increases during the project.

I. INTRODUCTION

Organizations have always been facing resistance to change as it is a natural reaction to change. With the emergence of computers and introduction of information and communication technology into the workplace resistance to change became even more noticeable and problematic. One could talk about a technological shock at first introductions of information and communication technology as it significantly affected the way of doing things in organizations. Also, fast pace of technological development contributed to frequent changes to keep up with the latest technology and therefore gain or keep the competitive advantage of organizations.

Today, after decades since first introductions of information and communication technology into the workplace, little has changed. Information and communication technology is still advancing at a high pace and organizations need to quickly adapt to the dynamic market environment to gain or keep their competitive advantage. Consequently, changes are very frequent and occasionally significantly affect both the way of doing things and the relations between employees. Despite decades of research on resistance to change and experiences gained in organizations facing it, resistance to change remains a hard nut to crack. Resistance to change manifests in various levels of intensity from apathy to sabotages in practically all software projects.

The share of unsuccessful software projects according to the iron triangle (time, cost, functionality) remains alarmingly high [1]–[5]. According to the CHAOS Manifesto 2013 [6], only 39 percent of software projects are deemed successful, i.e., completed on time, on budget and with all intended functionalities. The next 43 percent of software projects are deemed challenged, i.e., finished and operative however over budget, over time or with less functionality than initially planned. The remaining 18 percent of software projects fail, i.e., aborted or finished but the software is never used in practice. When looking only at larger software projects, the numbers drop dramatically as only 10 percent are successful, 52 percent are challenged and a massive 38 percent are failed. These numbers are even more problematic as larger software projects usually change the way of doing things, the business processes, and may negatively affect the whole implementing enterprise if they fail [5], [7]. Resistance to change is commonly recognized as one of the critical reasons for such high failure rates of software projects [5], [8]–[12].

Unfortunately, inaction is the most common managers' response to resistance when it actually occurs and even when they respond to it, responses are often ineffective [11], [13]. In order to develop effective responses, managers need first to gain a deep insight into the resistance situation [14]. However, this is hard to achieve without stakeholder involvement due to the complex and dynamic nature of the resistance phenomenon [14]–[17]. Furthermore, even when managers involve stakeholders in information gathering, the developed responses can still be myopic since the managers may intentionally or unintentionally disregard important insights and interests of other key stakeholders.

In this paper, we review the literature on resistance to change and how managers can respond to it. First, we establish the need to change the established terminology and introduce the term stakeholder resistance as an

adequate alternative. Next, we review the available literature on diagnosing resistance as a key step in managing it that can only be effective if different stakeholders' perspectives are considered. Finally, we review how managers anticipate stakeholder resistance and respond to it after it occurs. Based on the review, we lay the foundations for further work in this area by promoting the idea of stakeholder involvement in formulation of responses to stakeholder resistance.

The paper is structured as follows. First, we define stakeholder resistance to new software. Next, we review the literature focusing on in-depth understanding and diagnosing stakeholder resistance. Finally, we present how development of responses to stakeholder resistance has been investigated in the literature.

II. DEFINING STAKEHOLDER RESISTANCE

Resistance to change has been first defined by Lewin [18] as the force against change in organizations. Since then, resistance has been thoroughly studied in various fields. Managerial psychology mostly focuses on the individual level while information systems research addresses primarily the users of newly implemented software on individual and group levels [17], [19], [20].

In information systems research, *user resistance* is commonly used instead as the users of the information system are considered the main resisting stakeholders. It is defined as opposition of an individual user or a group of users to change associated with a software project [21]. It is consistently considered as a key issue in software projects and may be present before, during and after the software project [10], [21], [22]. Research shows that user resistance can be destructive and consume excessive resources on one hand [5], [9] however on the other hand there may also be justified reasons for user resistance, such as flawed software [11], [21], [23]. Proper management thus needs to both mitigate the negative effects of resistance and develop constructive solutions.

TABLE I. RESISTANCE LEVELS AND CHARACTERISTIC RESISTANCE BEHAVIOR

Resistance Level	Resistance Behavior
1	Being inactive Lacking interest in project activities
2	Gossiping about implementers Avoiding involvement in the project Voicing feelings of fear and concern toward the new system Doing as little as possible in the project context Non-key actors refusing to accept responsibility in the project
3	Complaining about the project Denying the need for the new system Voicing opposing points of view regarding the project Gossiping about the project
4	Attacking the credibility of the project Not utilizing the system when use is mandatory Leaking information about the project Forming coalitions
5	Attacking the credibility of the implementers Boycotting the project activities or the system Posing ultimatums Making threats
6	Deliberately committing errors Going on strike Sabotage Resorting to violence

Despite its widespread use, the term user resistance seems to be misleading [24]. Information system research literature focuses heavily on software users paying little attention to other stakeholders [25] even though they may be significantly contributing to resistance or even resisting themselves [11], [16], [17], [26]–[28]. Therefore, we find it more appropriate to use the term *stakeholder resistance* throughout this paper. Stakeholder resistance manifests itself in various ways [14], [15], [29]. Based on its severity, we can identify six levels of resistance behavior as presented in Table I [11].

Another branch of research needs to be mentioned at this point because it is closely related to stakeholder resistance, namely *information systems adoption*. Adoption is a widespread topic in information systems research and stems from the diffusion of innovations theory. New software is considered as an innovation that is adopted by its users. A debate is still going on in the research community whether resistance and adoption are antonyms. The intersection of both branches is undoubtedly significant however we can identify some key differences between them. First, the diffusion of innovations theory is based on the open market therefore its applicability to internal organizational environment is questionable. Second, some situations, such as mandatory use, pose a challenge. In these situations the innovation can be used however not as much or as efficiently as it could be. Such use is commonly referred to as resistant use [30]. Third, the diffusion of innovations theory does consider the social aspects of adoption however they are vastly related only to the innovation attributes. Early resistance research already showed that resistance may be affected by the meaning of new software rather than the software itself. The software project can be a tool that a certain interest group may take advantage of to improve their influence in the organization triggering strong resistance from other stakeholders [16]. Research also showed that the introduction of the same software into different environments may end up with very different outcomes [15]. Resistance may also significantly intensify when converging from the individual level to the group level, e.g., when individuals start forming coalitions to provide additional protection to their interests [15]. Also in these cases, the subject of resistance may switch from the software to its meaning or even its advocates [15]. The adoption branch of research does not consider any of these cases.

III. DIAGNOSING STAKEHOLDER RESISTANCE

Stakeholder resistance is a complex phenomenon that has been compared to the symptoms of a bodily disease therefore its' manifestation form needs to be distinguished from the underlying causes [11], [13], [14], [17], [26]. The causes of stakeholder resistance may not be straightforward and trivial to identify as the same manifestation form could have very different causes in different organizational contexts.

The literature provides theoretical explanations of stakeholder resistance and resistance checklists [11]. Theoretical explanations of stakeholder resistance provide insight into the resistance black-box [11]. In contrast, comprehensive resistance checklists attempt to provide an overview of all potential causes of resistance [11], [13]. A meta-analysis of published checklists and theoretical explanations identified the following causes of resistance [17]: lack of top management commitment, past outcomes, perceived threats, organizational politics, direct costs, capabilities gaps, collective action problems, myopia, conservatism, reactive mindset, incommensurable beliefs, groupthink, speed and complexity, and lack of perceived value.

TABLE II. GENERIC METHODS FOR MANAGING STAKEHOLDER RESISTANCE

Method	Method (*continued*)
Align change and its organizational context [16]	Stakeholder participation [8], [13], [16], [25], [28], [32]
Examine the impact of change [8]	Persuade opinion leaders [28], [32]
Employ an appropriate pace [8]	Persuade stakeholders [8], [11], [16], [25], [28]
Change *per partes* [8]	Restructure incentives [16], [25]
Manage interests [28]	Reward ideas [8], [32]
Fix organizational problems [16]	Acknowledgement of employee's effort [32]
Restructure stakeholder relationships [11], [16]	Employee training [8], [11], [16], [25], [32]
Clarify, promote and share vision [8], [11], [28], [32]	Morale boosting activities [8], [32]
Document new procedures and software [8], [25], [32]	Verify facts [14]
Clarify job definitions [8], [25], [32]	Acknowledge feelings [14]
Plan job transfers [8], [25]	Clarify feelings, beliefs, values and behaviour [14]
Limit hiring for a fixed period of time [8]	Challenge ineffective behaviours, unviable beliefs and values [14]
Provide change information to users [8], [11], [28], [32]	Suggest effective behaviours, viable beliefs and values [14]
Open communication [8], [25], [32]	Show sympathy [8]
Discussion of issues with stakeholders [11], [32]	Job counseling [8], [25], [32]
Redesign software [11], [16]	Group therapy [8]
Improve requirements gathering and software design [16]	Coercion [11], [16], [25], [28], [32]
Improve collaboration between software designers and stakeholders [16]	Threat with job loss [32]
Involve stakeholders in software design [25]	Replace non-supporting opinion leaders [16], [32]
Establish support services [25]	*Laissez-faire* approach [11]

Causes of resistance may help managers understand resistance however more in-depth information on the particular situation is needed to respond to it adequately [14], [16], [26]. Simply assessing causes of resistance may prove ineffective as individual stakeholders may not be able to evaluate them adequately by themselves [17]. Only by consolidating different stakeholders' perspectives, a comprehensive picture of the resistance manifestation and its underlying root causes can be attained [17].

IV. ANTICIPATING AND RESPONDING TO STAKEHOLDER RESISTANCE

Managing stakeholder resistance consists of two main phases – *anticipation* of stakeholder resistance and *responding* to manifested stakeholder resistance [11]. Nevertheless, most studies do not distinguishing between the two phases [11]. Early studies promoted various ways of stakeholder participation in the change process as an adequate method for managing resistance [13], [18]. Such methods that are appropriate for managing resistance regardless of the situation are considered as *generic methods* for managing resistance. A comprehensive list of generic methods for managing resistance is presented in Table II.

All generic methods for managing resistance may not be applicable to every situation [11], [14]–[16]. This was first pointed out more than six decades ago by Lawrence [13] who argued that stakeholder participation cannot be an universal solution to stakeholder resistance. In some cases, such as when participation is perceived as only *de iure* but not *de facto*, it can as well intensify resistance [13], [16]. This phenomenon is described by the contingency theory which assumes that some situations make certain generic methods more effective than others [28], [31]. Managers should choose and develop *situation-specific methods* that fit specific organizational settings based on extensive information gathering [11], [14], [16], [28]. In determining the adequate methods, there is a major emphasis on information gathering because methods for managing resistance can be effective only when they address the right root causes of resistance [11], [14], [17].

However, even situation-specific methods for managing resistance may not be always effective [11]. It may be impossible for managers to predict stakeholder resistance with certainty due to the complexity of software projects therefore management needs to monitor it and respond adequately if resistance actually occurs or does not diminish as expected [11], [13], [14], [16], [26], [33], [34]. There is a major gap in literature on this issue as little attention has been paid to responses to resistance after it actually occurs and only recently authors started devoting themselves to studying it thoroughly [11].

Responses to stakeholder resistance can be categorized into four categories: inaction, acknowledgement, rectification, and dissuasion [11]. *Inaction* can be due to unawareness of the extent of resistance or its causes, deliberate ignorance of resistance, or impotence to respond to encountered resistance. *Acknowledgement* refers to responses that are limited solely to acknowledgement of resistance. *Rectification* describes responses intended to tackle issues. Rectification can be congruent when responses address the root causes of resistance or non-congruent when they do not. *Dissuasion* denotes efforts to divert stakeholders from resisting. It includes coercion, authoritative persuasion and supportive persuasion [11].

Research shows that inaction and acknowledgement increase resistance in vast majority of cases while rectification and dissuasion might lower it [11]. There is strong support that congruent rectification lowers resistance and that non-congruent rectification increases it [11]. Dissuasion is successful in lowering resistance only through authoritative persuasion while the results on coercion and supportive persuasion are non-conclusive [11]. Authoritative persuasion seems to be the most successful response to resistance however it may not solve the underlying root causes. For example, requiring mandatory use can result in the so-called *resistant use* which may lead to decreasing job satisfaction, negative feelings towards supervisors and decreasing loyalty to the organization [30]. Therefore, the most desirable response to stakeholder resistance is congruent rectification, i.e., responses that address the adequately diagnosed root causes of stakeholder resistance.

In existing literature, it is entirely up to the management to diagnose and prepare responses to stakeholder resistance based on gathered information [26], [34]. This is a challenging task due to the complexity of stakeholder inter-relations and uncertainty regarding their reaction to the software project [8], [13], [14], [26], [30]. A collaboration of multiple stakeholders is thus required in order to enable adequate diagnosing of root causes and improve the quality of its conclusions [17]. There is a similar issue when developing responses to resistance. Once managers understand the root causes of resistance they can prepare adequate responses addressing them. However, it remains a challenge to predict the reaction of different stakeholders to management-imposed responses and careful monitoring is required to detect whether there is a negative reaction to them [14]. Including stakeholders in the development of responses to resistance should lower the uncertainty on their reaction as they have the opportunity to learn about responses beforehand and negotiate about them with other stakeholders.

V. CONLUSION

In this paper, we reviewed the existing literature on resistance to change. We first established the need to promote the term stakeholder resistance instead of the widely established term user resistance as the latter inadequately emphasizes users even though any stakeholder could resist the change. Next, we recognized the need for an in-depth understanding of stakeholder resistance and uncovered the difficulties of resistance diagnosis. Then, we presented generic and situation-specific methods for managing resistance as found in the literature and discussed their limitations. Finally, we laid foundations for further work in this area by promoting the idea of stakeholder involvement in formulation of responses to stakeholder resistance. Further research on categorizing resistance according to the resisting stakeholders would also be beneficial.

REFERENCES

[1] M. M. Altuwaijri and M. S. Khorsheed, "InnoDiff: A project-based model for successful IT innovation diffusion," Int. J. Proj. Manag., vol. 30, no. 1, pp. 37–47, 2011.

[2] D. Baccarini, G. Salm, and P. E. D. Love, "Management of risks in information technology projects," Ind. Manag. Data Syst., vol. 104, no. 4, pp. 286–295, 2004.

[3] P. L. Bannerman, "Risk and risk management in software projects: A reassessment," J. Syst. Softw., vol. 81, no. 12, pp. 2118–2133, 2008.

[4] K. El Emam and A. G. Koru, "A replicated survey of IT software project failures," IEEE Softw., vol. 25, no. 5, pp. 84–90, 2008.

[5] K.-K. Hong and Y.-G. Kim, "The critical success factors for ERP implementation: an organizational fit perspective," Inf. Manag., vol. 40, no. 1, pp. 25–40, 2002.

[6] The Standish Group, "CHAOS Manifesto 2013: Think Big, Act Small," 2013.

[7] D. Lavbič, O. Vasilecas, and R. Rupnik, "Ontology-based multi-agent system to support business users and management," Technol. Econ. Dev. Econ., vol. 16, no. 2, pp. 327–347, 2010.

[8] J. J. Jiang, W. A. Muhanna, and G. Klein, "User resistance and strategies for promoting acceptance across system types," Inf. Manag., vol. 37, no. 1, pp. 25–36, 2000.

[9] H.-W. Kim, "The Effects of Switching Costs on User Resistance to Enterprise Systems Implementation," IEEE Trans. Eng. Manag., vol. 58, no. 3, pp. 471–482, 2011.

[10] R. Meissonier and E. Houzé, "Toward an 'IT Conflict-Resistance Theory': action research during IT pre-implementation," Eur. J. Inf. Syst., vol. 19, no. 5, pp. 540–561, 2010.

[11] S. Rivard and L. Lapointe, "Information Technology Implementers' Responses to User Resistance: Nature and Effects," MIS Q., vol. 36, no. 3, pp. 897–920, 2012.

[12] B. Žvanut, P. Pucer, S. Ličen, I. Trobec, N. Plazar, and D. Vavpotič, "The effect of voluntariness on the acceptance of e-learning by nursing students.," Nurse Educ. Today, vol. 31, no. 4, pp. 350–355, 2011.

[13] P. R. Lawrence, "How to Deal with Resistance to Change," Harv. Bus. Rev., vol. 32, no. 3, pp. 49–57, 1954.

[14] K. Hultman, "Resistance to Change, Managing," Encyclopedia of Information Systems, vol. 3. Academic Press, San Diego, pp. 693–705, 2003.

[15] L. Lapointe and S. Rivard, "A multilevel model of resistance to information technology implementation," MIS Q., vol. 29, no. 3, pp. 461–491, 2005.

[16] M. L. Markus, "Power, politics, and MIS implementation," Commun. ACM, vol. 26, no. 6, pp. 430–444, 1983.

[17] S. L. R. Vrhovec, T. Hovelja, D. Vavpotič, and M. Krisper, "Diagnosing organizational risks in software projects: Stakeholder resistance," Int. J. Proj. Manag., vol. 33, no. 6, pp. 1262–1273, 2015.

[18] K. Lewin, "Frontiers in Group Dynamics: Concept, Method and Reality in Social Science; Social Equilibria and Social Change," Hum. Relations, vol. 1, no. 1, pp. 5–41, Jun. 1947.

[19] S. Laumer, "Why Do People Reject Technologies – A Literature-Based Discussion of the Phenomena 'Resistance to Change' In Information Systems and Managerial Psychology Research," in ECIS 2011 Proceedings, 2011.

[20] S. Oreg, M. Vakola, and A. A. Armenakis, "Change Recipients' Reactions to Organizational Change: A 60-Year Review of Quantitative Studies," J. Appl. Behav. Sci., vol. 47, no. 4, pp. 461–524, Feb. 2011.

[21] H.-W. Kim and A. Kankanhalli, "Investigating user resistance to information systems implementation: A status quo bias perspective," MIS Q., vol. 33, no. 3, pp. 567–582, 2009.

[22] M. Pardo del Val and C. Martínez Fuentes, "Resistance to change: a literature review and empirical study," Manag. Decis., vol. 41, no. 2, pp. 148–155, 2003.

[23] E. H. Ferneley and P. Sobreperez, "Resist, comply or workaround? An examination of different facets of user engagement with information systems," Eur. J. Inf. Syst., vol. 15, no. 4, pp. 345–356, Aug. 2006.

[24] P. Eskerod, M. Huemann, and G. Savage, "Project Stakeholder Management—Past and Present," Proj. Manag. J., vol. 46, no. 6, pp. 6–14, 2015.

[25] M. Ali, L. Zhou, L. Miller, and P. Ieromonachou, "User resistance in IT: A literature review," Int. J. Inf. Manage., vol. 36, no. 1, pp. 35–43, 2016.

[26] S. Fiedler, "Managing resistance in an organizational transformation: A case study from a mobile operator company," Int. J. Proj. Manag., vol. 28, no. 4, pp. 370–383, 2010.

[27] J. D. Ford, L. W. Ford, and A. D'Amelio, "Resistance to change: the rest of the story," Acad. Manag. Rev., vol. 33, no. 2, pp. 362–377, 2008.

[28] P. C. Nutt, "Leverage, resistance and the success of implementation approaches," J. Manag. Stud., vol. 35, no. 2, pp. 213–240, 1998.

[29] L. Coetsee, "From resistance to commitment," Public Adm. Q., vol. 23, no. 2, pp. 204–222, 1999.

[30] M. van Offenbeek, A. Boonstra, and D. Seo, "Towards integrating acceptance and resistance research: evidence from a telecare case study," Eur. J. Inf. Syst., vol. 22, no. 4, pp. 434–454, 2013.

[31] J. Woodward, Management and technology. London: Her Majesty's Stationery Office, 1958.

[32] S. S. C. Shang, "Dual strategy for managing user resistance with business integration systems," Behav. Inf. Technol., vol. 31, no. 9, pp. 909–925, Sep. 2012.

[33] D. Seo, A. Boonstra, and M. van Offenbeek, "Managing IS adoption in ambivalent groups," Commun. ACM, vol. 54, no. 11, p. 68, Nov. 2011.

[34] S. Vrhovec and R. Rupnik, "A model for resistance management in IT projects and programs," Electrotech. Rev., vol. 78, no. 1–2, pp. 73–78, 2011.

Object-Oriented Programming Model for Synthesis of Domain-Specific Application Development Environment

Tomislav Lugarić*, Zvonimir Pavlić**, and Dejan Škvorc**,
*Creative Fields d.o.o., Zagreb, Croatia
**University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia
zvonimir.pavlic@fer.hr

***Abstract*—Domain-specific languages are becoming increasingly popular allowing experts to focus on solving their specific problems instead of focusing on the means needed. Various domain-specific tools have been developed that allow experts to design systems and workflows by connecting graphical elements together. These elements have to either be predefined by a developer, or created by the expert by combining simpler objects together. The issue of bridging the gap between the semantics of the problem and the technology used to implement the solution therefore stays with the software developer.**

In this article, we propose an enhancement of the object oriented programming paradigm that is aimed at easier conversion of the technology (objects, classes) into their visual representation. We propose a high-level implementation domain to expert domain compiler that is able to automatically generate domain-specific representations of available technical components. The compiler is designed as a general-purpose domain-agnostic agent, which generates appropriate domain-specific representations of the technology using technical data obtained from the system and domain-specific data provided as a set of rules. This paper discusses the requirements placed on the architecture of the input system, the design requirements for the domain-agnostic compiler and the data needed for a complete description of the domain specifics.

I. Introduction

Requirements present in modern software development are becoming increasingly diverse, and have to account for a growing number of requirements that are unique for a certain domain of expertise. In order to satisfy these requirements, it is essential to integrate the knowledge of the experts into the developed software. In context of the software that is massively used by consumers, such as various office productivity suites, graphical editing or communication applications, the level of needed expertise is sufficiently low to integrate it into the solution. Furthermore, the required number of functionalities is low enough to allow one version of the software solution to satisfy needs of a wide consumer group.

In case of more complex applications, such as industrial simulations, banking institutions, and even scientific or research and development institutions, it becomes significantly more difficult to develop a software solution that can be delivered to the customer as-is, providing a satisfactory level of functionality and flexibility. One possible solution to this is development of the needed software in-house. This requires teams consisting of programming experts and domain experts. Another solution lies in the simplification of the software development process, so it can be tailored to the mindset of the domain experts. In this approach, the various components the system can be built of are raised to a higher level of abstraction than in the traditional object-oriented paradigm. Instead of thinking about objects, classes and loops, typical of a software developers mindset, the components are represented as something closer to the domain experts mindset, e.g. gauges, oscillators and signal generators (electrical engineering), clients, accounts and offices (banking) or volume meshes, boundary conditions and equation solvers settings (computational fluid dynamics). The data flow visual programming model [1] has proven to be successful, as it focuses primarily on the data that is exchanged between nodes that communicate with one another. There is quite a number of such tools available for various purposes [2], ranging from simple tools aimed at consumers with basic knowledge, e.g. Yahoo Pipes [3], to highly specialised tools for experts, such as Simulink (mathematics) [4] or LabVIEW (control engineering) [5]. Although these tools provide a more intuitive method of application development, the connection between the underlying source code and the visual representations still needs to be coded by software developers.

In this paper, we propose a novel programming paradigm, which places a number of additional requirements on the object-oriented paradigm in order to facilitate automatic conversion of the source code written by the developers into a visual representation that is suited for the visual programming paradigm. We wish to build upon the ideas of automatic GUI generation based on annotated source code [6]. Instead of automatically generating a GUI, we wish to generate blocks which represent objects or operations that best represent the tools for solving a specific problem in a domain experts mindset. These blocks can then be turned into GUIs if necessary, or they can be turned into visual blocks that the expert

can then connect into a working system using visual programming paradigm. The system will not be restricted to a specific programming language, rather it will support any object-oriented programming language. Furthermore, the generator of the blocks will be domain agnostic, and it will be able to generate various sets of building blocks from the same codebase, depending on the domain specific knowledge it is given. This way we wish to create a compiler that will be able to bridge the gap between software developers and domain experts by automatically adapting code written by the developers for further customization by the domain experts.

II. Related work

Visual programming describes any system that lets the user specify a program using two dimensional notation, instead editing the one dimensional stream of characters [7]. It allows the programmer to create a software artifact by manipulating graphical objects instead of specifying the artifact textually. Basic programming elements are often represented as blocks, with visual cues about using and linking these blocks. This approach has been successfully used in teaching by reducing the difficulties beginner programmers are facing when they start programming. Programming by assembling blocks together reduces the possibility of syntax errors and helps the novice programmer to understand the logic of the software artifact. It has been used to teach students modern object-oriented languages [8] and basic programming concepts using computer games [9]. There has also been success in programmer teachers education [10][11], facilitation of childrens' logical and computational reasoning [12] as well as introducing programming concepts to non-CS students [9]. Visual programming is also used by the experts, for GUI development [8], computer graphics [13], simulation [10], multi-agents systems [14][15] and mobile app generation [16].

There are several different visual programming paradigms in use. *Grammar building blocks* paradigm uses the grammar of a traditional imperative programming language transformed in blocks, and graphical cues help the programmer combine this blocks together. This format is very close to source code, and it can be perceived as a syntax guide for an imperative programming language. The most popular examples of these languages are Scratch [17] and Squeak [18]. *Data flow programming* is commonly used in professional applications [1], such as Simulink [4] and Labview [5]. Another example is the "Computational Model Builder" [19] developed for the Visualisation Toolkit [20], that allows non-programmers to build engineering workflows that involve simulations and experiments [21]. Blocks represent functions, and they are linked with each other along with the flow of data. Programming is performed by defining the flow of execution through the flow of data by linking outputs of blocks into the inputs of other blocks. Another visual programming paradigm is using *behavior trees* for defining the program flow of execution by connecting block in the tree-like structure. Each block is evaluated and returns a status back to the previous block along a tree structure, allowing the dependency of each block behavior to the behaviour of its children. *Event-based rules* is relatively simple paradigm where the programmer defines rules based on events in the system. A rule is triggered whenever the condition is met, and launches the specific intruction [22]. *Programming-by-demonstration* is a paradigm easily understandable for the novice programmers, because they specify the program by showing the system what to do [7]. This paradigm can be very expressive, depending on the system that uses it [23].

Creating user interfaces is a common task in application development, and it can become time and money consuming when developing applications for diverse platforms. Several user interface description languages are used for describing the GUI at a more abstract level and leave the platform-dependent GUI rendering to an automated GUI generation system [24][25]. Model-based GUI design uses various models to generate the user interface: task models, domain models, presentation models, dialog models [26][27][29]. Graphical user interfaces are generated automatically from the source code using code annotated with abstract commands of user interaction [6]. Automated GUI generation has also being used in GUI testing[30] and teaching students the basics of object-oriented design[28].

III. System description

The proposed programming model aims to facilitate automatic transition from the domain of the software developer into the domain of the expert using and extending the software. This transition is achieved using a translator, but instead of translating between various programming languages (e.g. from a high level language into machine code), this translator will translate between various domains of expertise. The input of the translator can be either a set of already developed classes with no connections between them, it can be a partially or completely functional system with the connections already in place. The translator will analyse the existing system and create appropriate blocks for a visual programming system depending on the domain of expertise it is configured for. The expert will then be able to define interactions of the provided blocks in order to implement needed functionalities. Since the blocks that are presented to the expert are already adapted to the experts domain, this approach allows the expert to achieve a level of expressiveness similar to that of a software expert, while staying within his or her domain of expertise.

The proposed system consists of three key components:

1) input system
2) domain-specific knowledge
3) domain-agnostic compiler

The input system can be written in any object-oriented language, and must be able to provide a predefined set

of functionalities in order to be analysed. The domain specific knowledge is a database of all necessary instructions and logic needed to perform the transformation of the input system into the expert's domain. The domain-agnostic compiler performs the actual transformation by first analysing the system, then applying the necessary level of abstraction and adding programming logic as instructed by the domain-specific knowledge.

IV. Input system requirements

In an object oriented environment, every class can be observed as a black-box, and is fully defined by its programming interface, while the specifics of how the underlying functionalities are implemented are invisible to the developer using the class. Similarly, in a data flow model, every node is fully defined by the data it accepts and the result it produces. Expanding on the idea of automatically generating GUI elements based on code annotations as described in reference [6], our proposed system will impose a requirement on all examined nodes to conform to a specific programming interface. Instead of using annotations, which are an expansion of a programming language, we propose utilisation of the already existing object-oriented concept of abstract classes, which is guaranteed to be supported by any modern object-oriented programming language.

Every node in the examined system must therefore be able to provide these sets of functionalities:

1) Identification/addressability (generic and specific)
2) Catalogue of available functionalities
3) Catalogue of available information
4) Catalogue of observeable events

The first functionality ensures the node can be extracted from the system, i.e. the compiler can request the system to provide a specific node. For example, when examining a banking system, a node representing an account will be able to identify that it is a bank account (as opposed to a bank client), and will be able to discern itself from other bank accounts by giving its account number. Similarly, in a system used for electrical circuit design, a resistor will be able to identify itself from a capacitor.

The second functionality ensures that every node can provide a list of functionalities that are exposed for usage in the visual blocks. In essence, this is an equivalent to the list of mangled method names that are located in a compiled library (.dll or .lib). The third functionality ensures that every node can provide a list of data it can be asked about.

The fourth functionality allows each node to be used as a trigger for the system. This can be regarded as a signal in a signals-and-slots paradigm or as a status change in the observer pattern. We propose to use these functionalities to enable each node to tell the system about itself, similarly to how a developer can learn about objects by looking at their interfaces (header files in C++). These functionalities can be implemented using virtual functions, and their presence can be mandated by inheriting from an abstract class. In essence, each object in the input system that needs to be exposed to the domain-agnostic compiler has to inherit from an abstract class that adds an interface through which the compiler can learn about the object. This methodology ensures a minimal overhead on the development process of new systems while making retrofitting of existing systems feasible.

V. Domain description

The database of the domain-specific knowledge is the component that gives the system its expressiveness. It contains all the instructions necessary to perform the transformation from the technical implementation into the domain-specific representation and to shift the focus from the implementation units into the individual operations. The process of generating the domain description requires a specification of all desired functionalities and a thorough analysis of the experts process. The experts process needs to be dissected into logical blocks for which a description will be made. We consider using UML use-case and activity diagrams as a description language for the domain-specific knowledge specification. The use-case diagram can be used to specify the system functionality and components (building blocks), while the activity diagram can represent the interaction and interconnection logic between objects. These diagrams have been used for modelling application behavior [30], as well as specifying GUI components and automatic GUI generation [29], and we state that it can be used in this environment for generating visual language building blocks.

In a hypothetical CAD application, these blocks can be various operations on 3D objects, with implementations of individual objects in the background. In a banking application example, they can be transactions, with account operations as a background.

Each operation that is defined in the domain specific knowledge will be defined by 4 things:

1) Types of blocks requested from the system
2) Functionalities needed from each block
3) Data needed from the user (input data)
4) Interconnection logic (layer of abstraction)

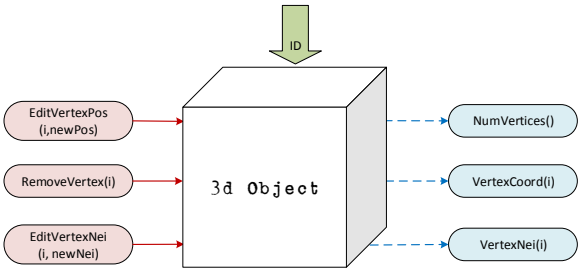


Fig. 1. Representation generated without domain-specific knowledge

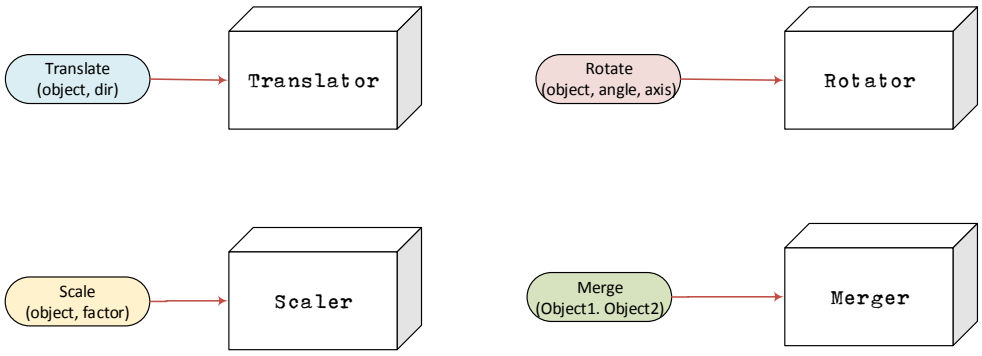


Fig. 2. Representation generated with CAD domain knowledge

This data can then be parsed by the domain-agnostic compiler and the system can be analysed and assembled based on it. Each functionality described in the domain description is a set of instructions used by the domain-agnostic compiler in order to create a block. If a block cannot be successfully generated, the compiler can simply skip it and continue to attempt to generate the next block. For example, if the domain-specific data is describing a meteorological application, one of the blocks might require the presence of show depth sensors in the system in order to generate a block for predicting future snowfall. If the system fails to provide this basic block, the generated representation will simply lack the snowfall predictor block, but will provide all the other blocks. This way, we will allow analysis of various systems using the same domain-specific knowledge.

VI. Domain-agnostic compiler

In its basic mode, when no domain-specific data is provided, the domain-agostic compiler makes a direct mapping of everything enumerated in the three catalogues into sockets on the graphic representation of the node. In this mode, the result is essentially an interactive class diagram of the analysed system. When domain-specific data is applied, the domain-agnostic compiler becomes a domain-aware compiler, and adds the engineering knowledge needed to transition the system into the domain of expertise required. By replacing the database of the domain specific knowledge the same input system will be transformed into a different representation appropriate for a different domain of expertise.

As an example, let us examine a system which has support for manipulating 3D objects. Each object is defined by a set of vertices and their connectivity. Each object has a method for providing its name, satisfying requirement 1. Each object allows changing coordinates of the vertices, changing neighbour list for a vertex, adding new or deleting old vertices and deletion of the object itself, satisfying requirement 2. For requirement 3, the object provides info about number of vertices and their coordinates. For requirement 4, the object provides information when it is clicked by the mouse. The system is then analysed and transformed by the domain-agnostic compiler without any domain-specific data. The result will have only a single block, representing a 3D object, as shown in Figure 1.

Then two databases of domain-specific knowledge are considered. The first database is designed by a CAD expert, and the second is designed for a computer animator. The CAD expert wishes to think of the process in terms of object rotations, translations, merging and scaling, while the animator wishes to think about making an object do specific movements or actions (e.g. walk, follow a trajectory, fall apart, explode...).

Firstly, a database of CAD domain-specific knowledge is added into the process. This database is written for a simplistic CAD package, offering object rotations, translations, merging and scaling. The resulting representation will therefore generate 4 blocks: scaler, translator, rotator and merger. The domain-specific knowledge will contain the information about the input that is requested from the user, and which data and functionality needs to be available in the system.

For the scaler, a pseudocode describing the analysis and generation of the representation may be as follows:

1) Find all objects which are 3D objects (from generic data in ID)
2) Ensure they offer functionality to edit vertices (from list of offered functionalities)
3) Ensure they offer data about existing vertex positions and neighbours (from list of offered data)
4) Generate a block which asks user for object ID and scaling factor
5) Add a layer of mathematical calculations that compute edge lengths based of existing positions and neighbours of vertices
6) Add a layer of mathematical calculations that compute new vertex coordinates and passes them to the 3D object
7) Embed the 3D object block into the newly generated block

8) Offer to the user only blocks which satisfy requirements in 2) and 3)

Similarly, we generate the translator, rotator and merger by polling the existing blocks for their functionalities, inserting the required mathematical calculations and generating the blocks. The resulting system now offers blocks as shown in Figure 2, which offer operations which are more adapted to a CAD expert.

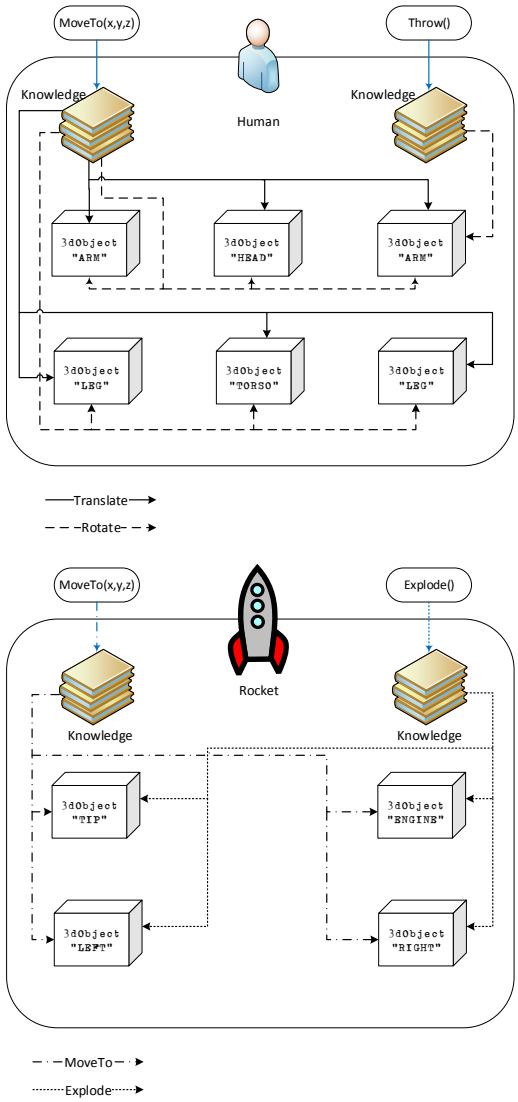


Fig. 3. Representation generated with 3D animation domain knowledge

For a second example the same input system is now paired with an animation expert's domain-specific knowledge. The basic building blocks are the same 3D objects used in the previous example, and they offer the same basic functionalities. The animator's tool is designed to allow him to animate a human figure and have it interact with various objects, as shown in Figure 3. For the "human" object, the database has the knowledge that it needs 6 3D objects (2 legs, 2 arms, a head and a torso). It also knows how to compute their relative positions from the position of the human. The actions the human figure executes can also be translated to absolute and relative movements of the objects (e.g. "Throw" is executed as a rotation of the right arm, "Move" can be executed as a translation of the torso and head combined with a set of rotations and translations for legs and arms). There is also knowledge on how to animate a fireworks rocket. The rocket has 4 3D objects (tip, engine, left and right half, for example). There is a move action for the rocket, that is executed as a coordinated translation of the 4 parts, and there is an explode action which is executed as a set of random translations of the 4 pieces. The requirements for the basic functionalities of the 3D objects are the same, but, depending on the domain specific knowledge, these basic functionalities are used differently. Using the same input system as for the CAD expert, we have managed to create an entirely different tool suited for a computer animator.

After this step of transforming the system from the implementation domain into the expert domain, the primary building blocks exposed to the user are no longer the implementation units (classes), but rather operations, which may include one or more classes as required. Using the same basic input system, and only replacing the domain-specific engineering knowledge, we can create tools suited for different experts. The CAD expert has a tool that has operations as building blocks, and can develop his process focusing on operations he wishes to do with various objects. For example, his process can be expressed as "Load a cube and a sphere, enlarge the sphere 2 times, and merge it with the cube to form a pillar. Rotate the pillar so the round part is on top." The animator can focus on having objects of specific types, and develop his process only telling them what to do, instead of focusing on how they achieve it. His process can be "Make the human move 100 meters forward, throw a rocket in the air and walk back, while the rocket flies 100 meters up and explodes". Neither of them needs to go through the tedious process of telling the underlying 3D objects what to do. Those operations are already embedded in the domain-specific knowledge. The technological implementation of the underlying logic can be developed in the developers domain of expertise, validated, and then automatically transformed into the experts domain for further refinement and customization.

VII. Conclusion and Future Work

The model proposed in the article aims to provide an automatic means of generating specialised building blocks for a domain specific language based on automatic analysis of existing systems. Instead of using annotations to expose functionalities, as proposed in [6], and as is seen in various software development platforms such as Qt [31] or VTK [20], we propose using only basic object-oriented concepts of abstract classes and inheritance, in order to make this model applicable to any object-oriented language. The model aims to reduce the need for time-consuming communication between the experts who use the software and the developers by formalising the knowledge needed

to transform the system from the technical implementation domain into the expert domain, and allow the experts themselves to customize the system according to their needs, while the developers can focus on optimizing, improving and debugging the underlying technology. More research will be needed to assess the best method of generating the domain specific knowledge, as well as determining the information that is needed from the experts in order to create the knowledge database. Furthermore, we believe there is also possibility of automating at least a part of the process of adapting the input system to provide all data needed by the domain-agnostic compiler. There is also a possibility to auto-generate a GUI directly as an output of the domain-agnostic compiler, or to make the compiler auto-generate output other than blocks for a visual domain-specific language. This customization of the compiler output is also a possible research topic for the future.

ACKNOWLEDGMENT

The authors acknowledge the support of the Croatian Science Foundation through the Recommender System for Service-oriented Architecture (IP-2014-09-9606) research project. The authors also acknowledge the support of Creative Fields d.o.o., Zagreb, as well as Consumer Computing Laboratory from Faculty of Electrical Engineering and Computing at University of Zagreb.

REFERENCES

[1] W. M. Johnston, J. R. P. Hanna, and R. J. Millar, "Advances in dataflow programming languages," *ACM Comput. Surv.*, vol. 36, no. 1, pp. 1–34, Mar. 2004. [Online]. Available: http://doi.acm.org/10.1145/1013208.1013209

[2] craft.ai, "The maturity of visual programming," 2016. [Online]. Available: http://www.craft.ai/blog/the-maturity-of-visual-programming/

[3] Yahoo, "Yahoo! pipes home page," 2012. [Online]. Available: http://pipes.yahoo.com/pipes/

[4] The MathWorks, "Simulink - simulation and model-based design," 2016. [Online]. Available: http://www.mathworks.com/products/simulink/

[5] National Instruments, "Labview system design software," 2016. [Online]. Available: http://www.ni.com/labview/

[6] J. Jelinek and P. Slavik, "Gui generation from annotated source code," in *Proceedings of the 3rd Annual Conference on Task Models and Diagrams*, ser. TAMODIA '04. New York, NY, USA: ACM, 2004, pp. 129–136. [Online]. Available: http://doi.acm.org/10.1145/1045446.1045470

[7] B. A. Myers, "Visual programming, programming by example, and program visualization: a taxonomy," in *ACM SIGCHI Bulletin*, vol. 17, no. 4. ACM, 1986, pp. 59–66.

[8] P. G. Kluit, M. Sint, and F. Wester, "Visual programming with java: evaluation of an introductory programming course," in *ACM SIGCSE Bulletin*, vol. 30, no. 3. ACM, 1998, pp. 143–147.

[9] K. Bromwich, M. Masoodian, and B. Rogers, "Crossing the game threshold: A system for teaching basic programming constructs," in *Proceedings of the 13th International Conference of the NZ Chapter of the ACM's Special Interest Group on Human-Computer Interaction*. ACM, 2012, pp. 56–63.

[10] M. B. Rosson and C. D. Seals, "Teachers as simulation programmers: minimalist learning and reuse," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2001, pp. 237–244.

[11] C. Seals, "Visual programming for novice programmer teachers," in *Proceedings of the 2005 conference on Diversity in computing*. ACM, 2005, pp. 26–27.

[12] A. Han, J. Kim, and K. Wohn, "Entry: visual programming to enhance children's computational thinking," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. ACM, 2015, pp. 73–76.

[13] P. E. Haeberli, "Conman: a visual programming language for interactive graphics," in *ACM SigGraph Computer Graphics*, vol. 22, no. 4. ACM, 1988, pp. 103–111.

[14] P. Bottoni, P. Mussio, B. Olivieri, and M. Protti, "A completely visual environment for agent-based computing," in *Proceedings of the working conference on Advanced visual interfaces*. ACM, 1998, pp. 261–263.

[15] F. Klügl, R. Herrler, and M. Fehler, "Sesam: implementation of agent-based simulation using visual programming," in *AAMAS*, vol. 6, 2006, pp. 1439–1440.

[16] D. Wolber, "App inventor and real-world motivation," in *Proceedings of the 42nd ACM technical symposium on Computer science education*. ACM, 2011, pp. 601–606.

[17] J. Maloney, L. Burd, Y. Kafai, N. Rusk, B. Silverman, and M. Resnick, "Scratch: a sneak preview [education]," in *Creating, Connecting and Collaborating through Computing, 2004. Proceedings. Second International Conference on*. IEEE, 2004, pp. 104–109.

[18] D. Ingalls, T. Kaehler, J. Maloney, S. Wallace, and A. Kay, "Back to the future: the story of squeak, a practical smalltalk written in itself," in *ACM SIGPLAN Notices*, vol. 32, no. 10. ACM, 1997, pp. 318–326.

[19] Kitware Inc., "Introducing the computational model builder, cmb," 2016. [Online]. Available: www.computationalmodelbuilder.org

[20] Kitware, "Vtk - the visualization toolkit," 2016. [Online]. Available: www.vtk.org

[21] C. Zdanski, S. Davis, Y. Hong, D. Miao, C. Quammen, S. Mitran, B. Davis, M. Niethammer, J. Kimbell, E. Pitkin *et al.*, "Quantitative assessment of the upper airway in infants and children with subglottic stenosis," *The Laryngoscope*, 2015.

[22] M. MacLaurin, "Kodu: end-user programming and design for games," in *Proceedings of the 4th international conference on foundations of digital games*. ACM, 2009, p. 2.

[23] S. Srbljic, D. Skvorc, and D. Skrobo, "Widget-oriented consumer programming," *AUTOMATIKA: Journal for Control, Measurement, Electronics, Computing and Communications*, vol. 50, no. 3-4, pp. 252–264, 2009.

[24] A. Puerta and J. Eisenstein, "Ximl: a common representation for interaction data," in *Proceedings of the 7th international conference on Intelligent user interfaces*. ACM, 2002, pp. 214–215.

[25] K. A. Schneider and J. R. Cordy, "Aui: A programming language for developing plastic interactive software," in *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference on*. IEEE, 2002, pp. 3656–3665.

[26] T. Clerckx and K. Coninx, *Integrating task models in automatic user interface generation*. TUL, 2003.

[27] G. Mori, F. Paternò, and C. Santoro, "Tool support for designing nomadic applications," in *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 2003, pp. 141–148.

[28] P. Dewan, "How a language-based gui generator can influence the teaching of object-oriented programming," in *Proceedings of the 43rd ACM technical symposium on Computer Science Education*. ACM, 2012, pp. 69–74.

[29] J. M. Almendros-Jimenez and L. Iribarne, "Designing gui components for uml use cases," in *null*. IEEE, 2005, pp. 210–217.

[30] M. Vieira, J. Leduc, B. Hasling, R. Subramanyan, and J. Kazmeier, "Automation of gui testing using a model-driven approach," in *Proceedings of the 2006 International Workshop on Automation of Software Test*, ser. AST '06. New York, NY, USA: ACM, 2006, pp. 9–14. [Online]. Available: http://doi.acm.org/10.1145/1138929.1138932

[31] The Qt Company, "Build your world with qt," 2016. [Online]. Available: http://www.qt.io/

Logistic and production computer systems in small-medium enterprises

M. Pighin
Department of Mathematics, Computer Sciences and Physics,
University of Udine, Italy
maurizio.pighin@uniud.it

Abstract - This paper presents a survey which describes how small-medium enterprises (SMEs) implement and use their information system with respect to their logistic and production processes.

The first part of the study identifies the characteristics of the companies and detects their general attitude towards information technology (IT). The second part propose a list of detailed logistic and production processes to analyze the structure and workflow of companies. For each process, if it is present, the company indicates whether it is supported by electronic means and, if so, the information system adopted, according to three categories: ERP system, office automation, specific developed system. The processes are grouped in ten macro-areas, five for the logistics and five for the production, and the aggregate results of the survey are presented in the paper.

The novelty of the study and its interest, both in academic and institutional context as in the real world, resides in the opportunity to evaluate the attitude of SMEs towards information technology in defining, organizing, planning and controlling their logistic and production processes.

I. INTRODUCTION

The present study is aimed to analyze, with special attention to the SME sector, the level of coverage that the information system provides with respect to logistic and production processes, examining how they are developed and implemented in company organization [1], [2].

We believe that the novelty of the study and its interest, both in academic and institutional context as in the business world, resides in the opportunity to appreciate the attitude of SMEs towards information technology in defining, organizing, planning and controlling their processes [3], [4].

The main function of information technology was initially recognized in process automation; only later, other factors, such as knowledge representation and issues related to the activity planning, contributed to the evolution of Enterprise Information Systems (EIS). Clearly, the evolutionary path taken by IT, from a mere support tool able to replace the human operator, became a vehicle for change essentially in two directions: the integration between areas and innovation of processes [5], [6].

The principal approach is with Enterprise Resource Planning (ERP) systems whose main prerogatives are control and optimal management of resource flows within the organization [7]. They are composed of subsystems natively integrated in terms of computer architecture and logic design; the sharing of a common database allows a consistent representation of reality; they permit procedures to enable the interaction and cooperation of management dependencies [8], [9], [10].

In many cases also other less structured approaches are followed, both with "Office" tools, such as Excel or Project, both with specialized tools for logistic and/or production management, and, to a lesser extent, with totally manual procedures.

II. THE SURVEY AND THE SAMPLE

The Friuli Venezia Giulia region (North-East Italy), especially in the last years, shows lively and active attention to the promotion and support of innovation, shared commitment across the board in universities, businesses and institutions.

The economic system of the Friuli Venezia Giulia is characterized by traditional industries: furniture, wood, metallurgy and manufacture of metal products and machinery. However, the growing weight taken over the years by the field of electrical and electronic equipment stands out positively. Mirror of Friuli Venezia Giulia is undoubtedly the province of Udine, not only for the importance in terms of the number of local units (almost 47%) and employees (44%), but also for a production system very similar to that regional. These reasons justify the choice of these SMEs as a reference universe for our survey. We built a sample of all firms reporting a turnover between 5 and 50 million euro, based on the information available at the Industrial Association of the Province of Udine, eliminating previously non-homogeneous ones from the point of view of processes or as part of specific areas of the service sector. It was so granted the membership of the sample in the categories SMEs and at the same time, company size sufficient to allow investments in IT potentially significant.

The aim of the survey is to assess the approach of SMEs towards IT in organizing their logistic and production processes verifying the type of information system adopted [11], [12]. It is well known that the diffusion of information technologies and their extensive use is higher in large companies; thanks to their size, they can justify higher fixed costs, benefit from a higher attraction of skilled labor and enjoy easier access to financial capital. Equally obvious is also the diversity of the role and weight that innovation and investment in IT-

intensive sectors, play in skilled labor and technology. In this sense, then, it is clear the interest and utility to monitor the reality of SMEs and their approach to IT [13].

The final sample is composed of 37 companies, half of which, approximately, has been active for more than 30 years (only 9.5% started within less than 10 years). 40% of companies have not important relationships with foreign countries (share of turnover of less than 20%); also 40% are involved in the exports with more than 60% of turnover. Almost the entire latter group is composed of metalworking and furniture firms, the sectors most represented in the sample [TABLE I].

TABLE I. SAMPLE COMPOSITION BY ACTIVITY

Activity	*%*
Metalworking	42.9
Furniture	28.5
Electrical/Electronic	20.0
Construction	8.6
Total	100.0

Both in terms of staff employed [TABLE II] and turnover [TABLE III] the companies fall mainly into the category of SMEs (staff employed between 50 and 249 people, annual turnover between 11 and 50 million Euro)

TABLE II. SAMPLE COMPOSITION BY EMPLOYEES

Employees	*%*
<50	28.6
50-99	40.5
100-150	21.4
>150	9.5
Total	100.0

TABLE III. SAMPLE COMPOSITION BY TURNOVER

Turnover	*%*
<10	26.8
10-19	41.5
20-39	24.4
40+	7.3
Total	100.0

The questionnaire is divided into three sections. The first and second are intended to identify the characteristics of the company and to detect the general attitude towards information technology [14], [15], [16]. The third part consists in a list of detailed logistic and production processes to analyze the structure and workflow of companies [17], [18]..

Production processes are grouped in five macro-areas

- Management of structural data: Bill of materials, Operations definition, Factory layout, Production calendars.
- Definition and Analysis of Production Costs: Material and operation cost definition, Roll-up costing.
- Planning and scheduling: Master production Schedule (MPS), Material Requirement Planning (MRP), Capacity Resource Planning (CRP), Detailed production scheduling.
- Internal operations definition and control: Internal Production Orders, Operation definition and accounting.
- External operations definition and control: External production orders, Operations definition and accounting.

Also Logistic processes are grouped in five macro-areas

- Management of structural data: Coding plane, Products data, Factory layout.
- Transactional data: Warehouse handling.
- Specialized requests: Batch management, Serial number management, Storage sectors, Automatic warehouse.
- Inventory: Warehouse costs, Rotation inventory.
- Analysis and control: Stock, Product availability, Handling control.

For each process, if it is present, the company must indicate whether it is supported by electronic means and, if so, the information system adopted, according to three categories: ERP system, office automation, specific ad hoc developed system.

III. THE GENERAL ASPECTS

About the question of the fundamental reasons of the decision to acquire information technology, the companies are in agreement especially on two factors: the performance improvement in their processes and reduction of costs.

There is, however, who points out other causes, such as compliance with external legislative constraints (for instance detailed tracing of logistic movements, different models of inventory evaluation), alignment with corporate strategy and adjusting its position relative to competitors [TABLE IV].

It is significant the “zero” assigned to the need for strengthening the core business, probable sign of a widespread difficulty in recognizing in the management of information systems as a source of competitive advantage. This is consistent with the well-established tendency to concentrate the critical knowledge for the SMEs in the hands of a small group of people (if not the only entrepreneur). This approach inevitably limits the innovative potential of information technologies.

TABLE IV: MOTIVATION OF IT INVESTMENT

Motivation	*Points*
Effectiveness	59
Efficiency	49
Law constraint	35
Competition	12
Strategy	11
Process advantage	7
Innovation	7
External production chain	4
Strengthening core business	0

There is not the same convergence, however, to recognize the obstacles on the way of the adoption of information systems. However, a few inconveniences

prove more serious than others: definition of objectives and functional specifications, identification of problems and causes, staff training and internal resistance. Reading data from the opposite perspective, the only item on which there is a broad agreement is the easy availability of outside expertise [TABLE V].

TABLE V: DIFFICULTIES IN IT ADOPTION

Difficulties	*Points*
Target definition	33
Problems origin identification	25
Training	24
Internal resistance to change	20
Technical definition	17
Cost/Benefits	15
Internal competence	15
Software/Hardware	13
External competence	5

IV. THE LOGISTIC AND PRODUCTION COMPUTER SYSTEMS

As already mentioned in the previous chapter, the main section of the questionnaire consists of a list of detailed activities, divided by macro-areas and, at an upper level, by functional area (logistics and production). For each of them, it was required to specify, if practiced, the type of information system that supports the process.

The variety of economic sectors and the particular areas in which each company competes, require, first of all, a verification of the presence of the mentioned processes. With our sample the companies declares in the average the 67.5% presence of the proposed process. The absence of processes respectively 32.6% (logistic) and 32.5% (production), usually is referred to specialized requests, as rotation inventory or detailed production schedule. The global data is satisfactory and ensures comparability among analyzed companies and representativeness of results.

TABLE VI analyzes in detail the absence of processes in the logistic macro areas.

TABLE VI: LOGISTIC - % ABSENCE OF PROCESS

Logistic	*%*
Structural data	19.0
Handling	14.3
Inventory	28.6
Analysis/Control	31.0
Specialized requests	70.0

Definition and management of structural data and warehouse handling are the areas more covered and managed; inventory and other warehouse analysis follows with a coverage of about 70%. Only specialized requests, like rotation inventory, are very poorly treated used only in particular conditions.

TABLE VII analyzes in detail the absence of processes in the production macro areas. Definition and management of structural data and Costing are the areas more covered and managed; planning and management of Internal operation follows with a coverage of about 65%. The External operation are the lesser managed, used only in particular conditions.

TABLE VII: PRODUCTION - % ABSENCE OF DEFINED PROCESS

Production	*%*
Structural data	26.2
Costing	25.4
Planning	36.7
Internal operation	34.5
External operation	39.7

SMEs show a high level of computerization: personal utility tools are used, in fact, from 95.2% of them, confirming it as the basic support that organizations use. However, more sophisticated technologies have a large diffusion: specialized systems (83.3%) and ERP (78.6%).

It is possible to appreciate the average degree of coverage that each kind of technology provides to the company logistic and production processes [TABLE VIII].

TABLE VIII: % OF COVERAGE OF LOGISTIC AND PRODUCTION PROCESS BY TECHNOLOGY

Technology	*%*
ERP	62.6
Specialized Systems	17.8
Office Automation	15.9
No System	3.7

Immediately visible is the positive data represented by the remaining part of the processes (3.7%) performed without the aid of computer. The ERP is the system that provides the largest support (62.6%) among SMEs.

Reasoning on the value of ERP coverage, despite being the highest, it is remarkable that about than 37.4% of the processes, on average, do not use a management system that by its nature is intended to provide a complete and integrated support to the operations. It has already been pointed out to the flexibility that allows partial implementation in independent functional modules, but it does not seem to be the only explanation that justifies the data set out above.

The result appear more clear when considering only organizations with ERP; the share of SMEs in which this information system supports more than half of the process is 66.7%; of these, however, only 36.4% have coverage by more than 70% of the process, in a quite articulate and extended mode to be consistent with the fundamental characteristics an integrated information system. In TABLE IX and X we analyze in detail the different logistic and production macro-areas.

In logistics undoubtedly ERP systems are the most used and offer greater levels of coverage, with a peak, as indeed predictable given the volumes involved, in the procedures for handling and especially in inventory. The Office automation is relatively widely used in the management of the structural data, for its simplicity in associating to individual items sheets and/or design and/or images. It is little used in the procedures that treats large data sets as the handling and inventories.

TABLE IX: LOGISTIC - % OF COVERAGE OF LOGISTIC PROCESS BY TECHNOLOGY AND BY MACRO-AREAS

Logistic	*% ERP*	*% Spec. Systems*	*% Office Autom.*	*% No System*
Structural data	67.0	12.7	16.4	3.9
Handling	76.9	14.4	6.0	2.8
Inventory	80.6	14.7	4.7	0.0
Analysis Control	68.4	17.6	10.5	3.4
Specialized Requests	58.4	19.8	18.5	3.2

Symmetrically specialized systems are less used in procedures with greater sets of data and more in the processes of analysis and control

The specialized requests, that are relatively not very present, have the least use of ERP systems (often just they are not present in ERP systems) and the remain percentage is divided between the specialized systems and office automation.

The manual operation is very low everywhere, with a "zero" in the inventory procedures.

TABLE X: PRODUCTION - % OF COVERAGE OF PRODUCTION PROCESS BY TECHNOLOGY AND BY MACRO-AREAS

Production	*% ERP*	*% Spec. Systems*	*% Office Autom.*	*% No System*
Structural data	69.4	16.8	10.6	3.2
Costing	55.4	15.5	24.9	4.3
Planning	40.6	16.0	32.7	10.8
Internal Operations	42.6	31.7	22.1	3.6
External Operations	66.8	19.0	12.9	1.3

In the area of production systems, ERP systems are still the most common, but certainly to a lesser extent than in Logistic. They are widely used in the definition and management of the structural data such as Bill of Material (BOM), and Costing (that are related to structural definition). Nevertheless, there is a relatively low use in the procedures of Planning and management of Internal operations. Regarding the Planning area, analyzing the raw data we observe the MRP is widespread while other planning procedures, such as CRP or MPS, are lesser used.

The high percentage in External operations should be evaluated considering that they are relatively rarely used and, when present, require a strong link with purchases and logistics and therefore the ERP system is very useful.

Office automation is used in support of Costing and Planning operation, while Specialized system are more used in Internal and External operation.

Manual operation is almost absent, with the exception of planning that is, in low percentage, always done without computerized support.

Analysis of data shows some fragmentation of the information system for SMEs where we find the parallel use of different technologies in support of the same functional area. The most likely causes, and not necessarily alternative, seem to be two: widespread provision of basic management packages (less restrictive in how to use and less invasive in terms of organization than sophisticated ERP), or implementation and use of the management system often inadequate.

V. CONLUSION

The sample of SMEs showed a good level of computerization in logistic and production systems guaranteed, in particular, from the ERP.

The analysis by functional macro-area has also emphasized that the use of computer technology is often inadequate, resulting in a messy support to enterprise management.

A partial coverage of the processes and, not least, the difficulties encountered in completing the questionnaire help to reveal, therefore, a mentality unwilling to look beyond the purely technical aspects of the application of IT and difficult to recognize in them an opportunity growth and innovation.

Undoubtedly, the challenge for the future goes through a better and more coherent use of information systems, in particular ERP systems that are aimed at integrating the various business functions.

Not least it is the need to consider the information system not a simple tool, often replaced by spreadsheets or similar application, but the heart of the organization and control of business processes.

REFERENCES

[1] S. Alter,"Information systems: a management perspective", Prentice Hall Ed., 2002

[2] K. Laudon, J. Laudon, "Management Information Systems", Pearson Ed., 2014

[3] T. Davenport, "Enterprise 2.0: The New, New Knowledge Management?", Harvard Business Online, 2008

[4] M.E. Porter, "Competitive Advantage: Creating and Sustaining Superior Performance", The Free Press, 2004

[5] R. Galbraith, "Designing the Global Corporation", Jossey-Bass Ed., 2000

[6] J.A. O'Brien, "Management information systems", Irwin Ed., 2010

[7] G. De Witt, "Informatica moderna e produttività di impresa", F.Angeli Ed., 2001

[8] M. Golfarelli, D. Maio, S. Rizzi, "Ingegneria dei sistemi informativi", Esculapio Ed., 2000

[9] E. Turban, E. McLean, J. Wetherbe, "Information technology for management", Wiley Publ., 2002

[10] R K. Rainer, B. Prince, "Management Information Systems", John Wiley and Sons, 2015

[11] O. Wight, "Manufacturing Resource Planning: MRP II: Unlocking Productivity Potential", John Wiley and Sons, 1995

[12] C. Poirier, M. Bauer,"e-Supply chain: using the Internet to revolutionize your business", Berret-Koehler Publ., 2000

[13] G. Lo Martire, "Il controllo di gestione", F.Angeli Ed., 2003

[14] M. Christopher "Logistic and Supply Chain Management", FT Publishing, 2015

[15] N. Slack, A.Brandon-Jones, R. Johnstone, A Betts, "Operation and Process Mangement", Pearson, 2015

[16] M. Pighin, A. Marzona,"Sistemi Informativi Aziendali, strutture e processi", Pearson Ed., 2011

[17] M. Pighin,"Business Information Systems in Small Medium Enterprise: Why and How", in Proceedings of 23rd International Business Information Management Association Conference - Sustainable Growth, Economic Development, and Global Competitiveness, IBIMA Press, Mineola (USA), 2014, pp. 474-480

[18] M. Pighin, A.Marzona, P. Vatri "La struttura dell'organizzazione – Indagine sull'uso dei sistemi informativi aziendali nella piccola e media impresa", Aracne Ed., 2009

The implications of employing component based software design in non-commercial applications

Bruno Zorić*, Goran Martinović* and Ivica Crnkovic**
* Faculty of Electrical Engineering, Osijek, Croatia
** Chalmers University, Department of Computer Science and Engineering, Sweden
goran.martinovic@etfos.hr

Abstract – Software development today is riddled with various challenges, and hangs on the balance of cost and profit. The industry is yearning for shorter times to market without any reduction in quality. Since component based software engineering emerged as a prominent and solid approach to alleviate these challenges, the question remains are there any benefits of employing this paradigm in other areas, excluding commercial applications, such as the area of academic research. This paper considers a different aspect of component oriented design application. It tries to demonstrate the implications of its use in a case study of scientific research on a pattern classification use case, where the majority of component advantages such as cost reduction or profit maximization do not apply. The use case presented in this case study utilizes component based software engineering to build components of a classification system. Feature extraction, classifiers and evaluation procedures are interchangeable components which enable the discussion regarding the benefits of component based development in non-commercial areas of application.

Keywords: Classification, component based software engineering, components application

I. INTRODUCTION

Software development at present represents a significant challenge considering numerous available platforms, programming languages, paradigms, the sheer importance of software but also the constantly growing race between cost and profit. It is becoming increasingly important to push the product to the market as soon as possible, while maintaining a high level of quality. In classic software development, the process begins with the separation of system elements into subsystems, and after that into lower level blocks, such as interfaces, classes, methods and similar [1]. Through the implementation and interaction of these blocks, systems with the specified functionality are realized that are more or less replaceable, expandable and scalable. If any of these features is aspired towards, special attention needs to be given to the design stage where the solution elements are abstracted in various ways. Component based development (CBD) is a relatively novel paradigm which, although in some ways resembling object oriented programming (OOP), surpasses it and improves upon it [2]. The primary goal is no longer to implement parts of the solution but to design and create solution segments – carefully designed, precisely modeled and defined, well documented and interoperable components. The main benefits of such an approach are cost reduction and shorter time to the shelf stemming from the fact that components are readily available. However, these benefits are mainly aimed at the commercial segment, and this paradigm is most often found in the industry. Other areas of potential use remain scarce, since this approach alongside its advantages brings several drawbacks mainly concerning the time needed to create good, scalable components, an agreement is needed regarding interfaces, but also domain and development knowledge required in order to achieve it is quite substantial.

Development of various software solutions in academic research is common and necessary. It is often not enough to utilize specific tools, and relying on tool creators to provide the software required by the ongoing research could prove problematic. This is clearly an area where domain knowledge is not lacking, so the question could be posed why is CBD underutilized in this area? In order to shed some light on this topic, the advantages and shortcomings of CBD should perhaps be viewed from another direction, and this is hopefully what this paper tries to achieve. Given a case study of a classification framework for the comparison of various classifiers, feature selection and other involved procedures (which could be developed with CBD in mind), benefits and drawbacks of this paradigm concerning areas other than commercial application are discussed.

The rest of this paper is organized as follows. Section II describes the key terms and concepts in CBD. Pattern classification, which is to be used as a use case in order to demonstrate the benefits of CBD in research is presented in section III. A scenario of a CBD based environment for pattern classification is considered in section IV. Section V provides a discussion regarding the main benefits and drawbacks of CBD application in research environments. Finally, section VI concludes the paper.

II. COMPONENT BASED DEVELOPMENT

The goal of component based development is to employ clearly defined, finished, standardized components to create new software solutions [3]. Only if the needed components are nonexistent, time and effort should be invested to create them and add them to the catalogue. The reuse of finished components is not novel in human operation, and it is not novel in software engineering. The OOP paradigm is created with this in mind, but CBD raises this concept on a far higher level by solely relying on already created components with far greater reusability and exchangeability. This approach demands significantly larger investments of time and effort in the design stage with the hope of return of investment (ROI) through repeated reuse.

A component is a piece of software created from a potentially large number of classes, attributes and methods offering the designated functionality described and enforced through the use of specified interface. Definitions available in the literature describe components in a similar manner, for example in [2]: „A software component is a unit of composition with contractually specified interface and explicit context dependencies only. A software component can be deployed independently and is subject to composition by third parts". A point of major importance for any component is its standardization and documentation. Were this not the case, significant issues could arise from the use of such a component in a complex system or during component communication. In order to enable the utilization of a component and communication with other components, functionality is defined for each component through the provided interfaces while requirements are formed via required interfaces.

Interfaces are used to define component interaction with its environment. They can be considered partially similar to interfaces in OOP, defining contracts that classes need to respect in order to maintain compatibility and provide polymorphism. As with described interfaces, CBD interfaces need to be simple, since complex interfaces lead to maintenance issues or component replacement/upgrading related problems. The interface itself defines a series of operations that are either provided to other components inside a system or are required from them in order to utilize the component's functionality itself. In other words, these are used for components to provide or request services from one another. Most commonly these operations are described with method signatures (names and parameter lists). According to [1], the interface should be strictly separated from the component, both physically and logically. This means that the interface should be contained in a separate assembly to be employed by any system segment that relies on it (e.g. adapters, UI components and other glue code). Since components are completely separated from their interfaces, the only point of interaction is precisely the interface. It in turn enables easy component replacement and upgrade, since everything is implemented against an abstraction. It is important to note that interfaces can be upgraded as well, but cannot lose backwards compatibility in the process. Should an interface be redefined in a way that removes or replaces a part of the functionality, all old components would be rendered useless and would require redefinitions and recompilations to acquire new interface elements [4].

In order to accept any paradigm, norms and standards of its use need to be defined. Considering that CBD is still relatively young, the standards are not compliant so there exist several component models aimed at different languages or applications that try to define the rules of component definition and development. According to [5], a component model defines a set of standards to create components, from naming, interaction, adaptation, evolution to implementation. Some of the better known component models are .Net, MS COM, Enterprise JavaBeans, CORBA, OSGi etc. An exhaustive list and comparison are available in [6], where beside the description of models, a framework enabling individual model evaluation is given.

III. PATTERN CLASSIFICATION

Classification belongs to the area of machine learning, and it represents the assignment of patterns to sets so that all set elements have similar properties [7]. Each pattern consists of a series of values (possibly of different type, e.g. numeric, categorical etc.) describing it. This value series is usually represented by a vector and is called a feature vector. Besides the feature vector, each pattern is given a designation of the set it belongs to. This designation is usually called a class or a class label. All class labels are known in advance so classification is also called supervised learning. Its applications are numerous, including product quality assessment, biometrics, medical data recognition / categorization and so on (see e.g. [8, 9]).

In order to start the classification procedure it is necessary to represent the observed objects with data adequately describing them. It is often not prudent to include all the data collected (e.g. all pixels of an image), but to transform the original data into a different information space in various ways. Since the use case in this paper concerns textured images, and the input data comes in image form, the most common way to transform the input data to feature vectors is to employ transformations such as wavelet or Fourier, to use statistical descriptions such as histograms, gray level co-occurrence matrices and other approaches, as the ones presented in [10].

The procedures for performing classification are called classifiers. According to the "no free lunch theorem", no classifier can be deemed more effective than another or from random guessing in advance [7]. The superiority of an individual classifier arises from the type of the problem, data distribution and other information about the data. It is therefore of paramount importance to study the problem to be solved and to adapt. Still, the literature demonstrates the overwhelming presence of several well performing classifiers such as distance based classifiers (e.g. k-Nearest Neighbors (kNN) [11], and its variants), neural networks (e.g. multi-layer perceptron (MLP) or radial basis networks [12]) or support vector machines. Besides those, classifier ensembles are utilized where multiple models are combined through procedures such as bagging, boosting or voting in order to reach a decision.

Classification result analysis can be performed in several different manners [7]. The most common parameter considered during result comparison is the misclassification rate (MCR). This is the ratio of incorrectly classified patterns and the total number of classified patterns. The larger this value gets, the poorer the performance of the classifier. Sometimes, this value alone is not enough and other specialized quality measures obtained through the confusion matrix are introduced, such as recall, precision, kappa statistic, f-measure and others. The confusion matrix is a square matrix displaying the count test patterns of a class are classified as each of the classes. Other approaches of result presentation are available, such as the Receiver operating characteristic (ROC) or the area underneath it, called Area under curve (AUC). However, the most common problem present in

performance assessment is not the quality measure, but test set selection and the reliability of the result based on the available data. The main issue is data availability, since the number of data with known class labels is most commonly limited. Upon this data, a model is designed, validated and tested which poses a considerable problem [13]. To ensure unbiased classifier performance assessment, several approaches are proposed, such as holdout testing, bootstrap testing and k-fold cross validation testing. The outline of operation for each of these tests is presented in Fig. 1.

Several tools for comparing machine learning approaches are already available, such as Weka, Orange etc. They can create classification models, analyze the performance and additional approaches can also be created and added since they are based on Java and Python, respectively. Novel experimental approaches available in the literature that you should want to compare yours with are rarely available. More often than not researchers develop their own testing frameworks. As mentioned earlier, in order to enforce the interfaces enabling a fair comparison, a regulating body is needed. That way, all testing could be abstracted for example to the automated system run by the publisher. Should any of these tools become the standard, then the points made in the discussion of this will be fulfilled.

IV. A COMPONENT BASED CLASSIFICATION SYSTEM

In order to demonstrate the application and the upsides of CBD, an application is developed that is based on components and enables feature extraction from texture images, the application of classifiers on this data and the performance analysis of the classifier employed. Each listed task is performed by a separate component, and all interaction is specified in functional interfaces. In contrast to the development of components, where significant time is dedicated to the design stage, some other development approach could be utilized, such as the agile development. That approach would not be without its benefits, but in this use case the specification of the system should be clear upfront, and the interfaces should not change and evolve (as they would in agile development) if they are to be utilized for the comparison of different methods found in the literature over a significant amount of time.

The application is developed in the C# programming language and uses the .Net framework. Both components and interfaces are given as separate assemblies (.dll files), and the application itself is deployed as an executable. A high level outline of the proposed system is given in Fig. 2. The goal of the proposed solution is to enable the classification of textures through the utilization of different feature vectors. Besides that, the application should enable performance assessment and verification of the classification procedure.

The entire system is composed of three key components and the code serving as glue enabling the

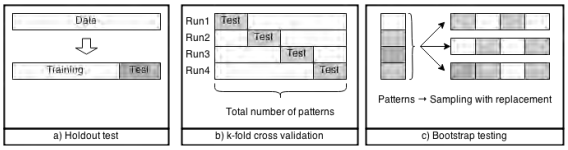


Figure 1 Modes of data splitting and utilization during analysis

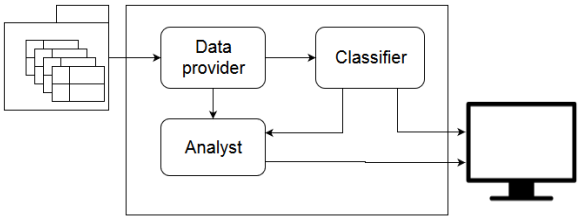


Figure 2. High-level outline of the developed system

instantiation and use of the defined components. The components used are:

- **Dataprovider** – A component to extract features from raw data and provide patterns
- **Classifier** – A component to determine the class of a pattern
- **Analyst** – A component offering classifier performance analysis on a given data set

Since the proposed system performs texture classification, its inputs are digital images of various textures. They are represented with 2D arrays of integers, each denoting the intensity. Greyscale images are used for simplicity, but RGB images can also be considered. Since pixel values used directly usually do not represent descriptive features, a transformation is performed in order to increase vectors' descriptive properties. In this paper two types of feature extraction procedures are used (each approach given as a component), histogram calculation and local binary pattern (LBP) histogram calculation. This component is not limited to using graphical images, it can also use text files, sound, standard .arff files and others. The implementation handles the details while the component only cares about the abstractions.

A histogram is a graphical representation of data distribution. When applied to digital images it represents the distribution of a certain color intensity or the gray levels intensity, depending on the image. The data being utilized consists of positive integers representing individual image pixels. These values can be directly applied as features describing a pattern [14]. Fig. 3b shows an example of a histogram obtained from an image used (Fig. 3a) in the experiment.

Local binary patterns features are commonly histograms described in [15]. An LBP operator is utilized on image windows by observing the central pixel (g_c), while the region around it (g_i, i=0,…,x) is modified to binary form depending on the central value, as given in (1). The whole region is then replaced with a calculated value, most often using some kind of weighted sum such as demonstrated in Fig. 4. A histogram is calculated based on the obtained values.

$$L_{x,y} = \sum_{i=0}^{x} sign(g_i - g_c) * 2^i \qquad (1)$$

The components based on the aforementioned are presented in Fig. 5. The data provider enables obtaining of feature vectors based on the list of strings representing file paths. The data provider does not care what kind of files are in the given paths, although the concrete implementations of this component provided in this paper

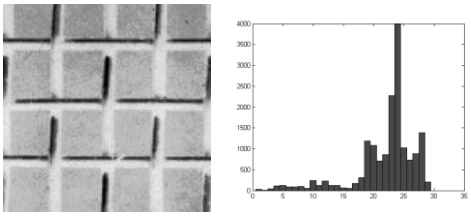


Figure 3. a) Texture sample b) Histogram of the texture

use image files to extract features. Two concrete implementations of this component are developed for the needs of the experiment, one giving a histogram of gray levels and another providing a histogram of uniform LBPs.

The second component is the classifier, having the task of providing the classification model based on data given by the provider component. After the model is created, the classifier enables pattern classification, i.e. it determines and returns the class label based on the input patterns feature vector. Similarly to the first component, two concrete implementations of this component are provided in the developed system. The first one is the k-NN, and the second one is the neural network, i.e. MLP.

The final component is the analytic component utilizing testing methodologies mentioned in section III. This component provides the option of classifier evaluation on a given dataset provided by the specified data provider. In this paper only one concrete implementation of this component is given, based on holdout testing methodology. Should the need arise, other components could easily be added, since the procedure of introducing new components is the same as for the first two components.

Although each of these components is a separate assembly and can be used independently in other projects, it is necessary for components to communicate with each other. This is ensured through compatible interfaces. As stated in [1], interfaces should be separated from the components both logically and physically. The reason resides in the fact that different components may rely on these interfaces, and components with the same intent should share interfaces. The client application (glue code) also has to have means of communication with each of the components. This paper utilizes several interfaces necessary in order to provide the required functionality. Physically, interfaces are separated into three separate assemblies (.dll): DataFramework, ClassifierFramework, AnalysisFramework. Each of the components utilizing the interfaces is provided as a separate assembly as well. The interfaces are shown in Fig. 5.

The concrete implementation of the system presented in Fig. 5 is shown in Fig. 6. The right side of the UI contains the control panel. Components that are utilized are selected from the drop-down menus, and it is possible to introduce new, previously unknown component implementations from .dll-s, instantiate them and use them

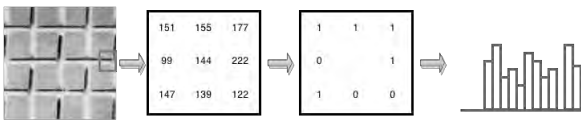


Figure4. Example of LBP histogram calculation

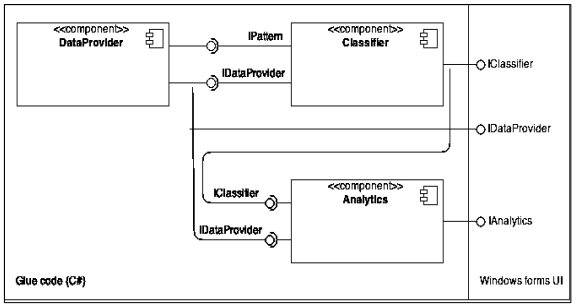


Figure5. Components and interfaces of the system

instead of the currently utilized ones on the fly, without any need for recompilation. This functionality conforms to the premise that components need to be replaceable. It is enabled through the use of interfaces in the client application, so that the application itself does not care which concrete component is in use, as long as it subjects to the specified interface contract. The instances of these components are created by the use of a factory design pattern where the factory class keeps a list of all components in the form of a dictionary. This dictionary is filled through reflection, and it is possible to find a .dll file containing a concrete component. The application will check the assembly contents identifying the type implementing the appropriate interface. Besides the replacement of a component in the way earlier described, it is possible to replace a component with a newer version just by exchanging the current .dll. The new version can for example perform certain tasks more efficiently (using different data structures, parallelization etc.). Since interfaces are separated, the components and client applications rely on them to define the communication contracts. Other than defining behavior, classes for data exchange can be defined. Such a class is defined in the analysis interface. This class provides data about the classification performance. If necessary, it can be inherited and extended to include additional information without breaking functionality, while various adapters can also easily be written to conform the objects passed around to desired interfaces.

The operation of the application is presented in Fig. 6. Data is loaded from a directory containing images of three

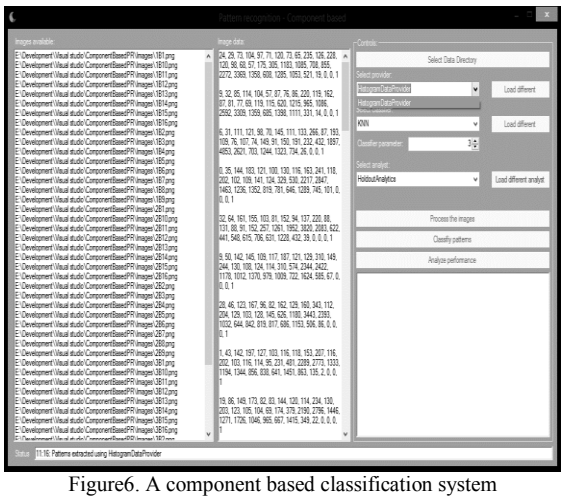


Figure6. A component based classification system

different types of textures. Each texture is represented with 16 images. After the directory is selected and path to files retrieved, images can be processed with the data provider component. Images are processed in a way that the component obtains a feature vector from the image using one of the methods described in section III. As can be seen, the system initially offers only one data provider component. This component is added statically, and if necessary additional alternative components can be dynamically loaded. The classifier component can be exchanged equivalently. At the beginning, the system only knows about the kNN classifier, but the MLP can be loaded from a .dll file. Concerning the analytics component, only one is present and available. Others can be loaded, but were not implemented for the needs of this paper. The only thing necessary in order to provide another analytics component is to implement IAnalytic interface defined in the AnalyticFramework .dll file. The final result of the analysis is presented in Fig. 7. It is notable that the classification accuracy is presented and also the confusion matrix is given. Based on this matrix it is possible to calculate different statistical quality measures, described in section III. If a component is to be developed that would calculate these measures, it would implement the same interface (IAnalytic) and inherit in some form the base class for result storage (AnalysisResult). It would expand the base class to incorporate other measures, and the system utilizing such a component would be able to downcast the resulting AnalysisResult object, fully exploiting the available performance data without breaking compatibility and functionality of other clients unaware of these benefits.

V. DISCUSSION

The proposed system enables a CBD approach to classification. In order to achieve this, careful consideration was imminent in the design stage for the system and its components to facilitate reuse and replacement. Even though the presented use case is relatively simple, the proportion of the design stage is quite larger than the proportion of the implementation stage in the overall development time (approximately 1:4 in favor of design). This is obviously a very important consideration to take into account when approaching research problems of this kind. Fast paced research that requires frequent change could prove problematic for this approach. However, the time invested in design pays dividends later when the opportunity for reuse or replacement comes in sight. Through this approach replacement was rendered easy due to the fact that all components conform to specific required and provided interfaces. During result analysis, classifiers were replaced on the fly, as well as data preprocessing (feature extraction) components. The built classifiers were later used in a different project without problems. Problems in this area stem from the fact that these interfaces are quite specific, and it is often quite hard to conform to the ever growing requirements appearing during research progress and novel ideas occurring on daily basis.

The knowledge required for the creation of a system such as this one is two-fold. Both domain and development knowledge and experience are needed and on somewhat high levels. Although this could prove difficult at times, CBD could be a middle ground between standard development and the usage of finished tools. It could enable

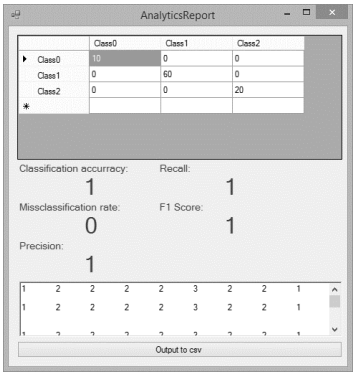


Figure 7. System results display

a faster pace of tool and methods creation and implementation, at the same time providing finished solutions (components), from which tools could be created according to individual needs.

The interfaces for this use case are created to facilitate its needs. That is not to say that they could not be either specified differently or simply better defined. For the listed use case the goal was to enable replacement in order to facilitate easier testing and analysis of various approaches to feature extraction, classification and results analysis. If further approaches are to be developed, this framework should enable a simple comparison with currently available methods. All that is needed is for the new component to conform to the given interface. Furthermore, other researchers could use our components and easily demonstrate the benefits and differences of their own approaches. Today, this is enabled through the presentation of algorithms in the literature or through source code exchange. The latter is something that is not always possible or that the authors are not always inclined to do. Publishers and reviewers could use a database of different components to verify the results and so on (even now some require the source code). It is apparent from the aforementioned that these interfaces are quite important so their definition should be somewhat centralized. This could perhaps be achieved through the open source community, through publishers, through academic institutions which could maintain a component specification or even standardizing bodies or professional associations. In the latter cases the components could be adapted when the appropriate publication is selected. The problem could arise from competing interface standards. If too many of them should be present, it would lead to a state of confusion.

As stated earlier, this approach would enable the sharing of components between researchers. Approaches in this research area are sometimes rather complex and rely heavily on various other domains. Often researches only try to improve a single aspect of a classification system (e.g. a feature extraction method) and have no need for the implementation of a classifier or analysis classes and methods. This development approach could enable a simple method of knowledge and implementation exchange all the while providing support to comparison and result analysis. Also, it could enable rapid prototyping of novel methods. If of the shelf components could be utilized, time would not be spent on the creation, compilation and adaptation of code but could be directly invested in research and the development of the new approach. Potential drawbacks are

possibly the component framework selected, or the interfaces chosen but this could be alleviated through adapters or through recompilation.

In the area of classification both the result analysis and the presentation of results are not universal. There exists a plethora of options to use when performing the analysis (such as hold-out testing, cross-validation etc.) but also when presenting these result (such as using confusion matrices, parameters derived from them or utilizing ROC curves and other). Sometimes this selection is enforced by the data or the problem, and sometimes it is freely selected. The use case presented in this paper utilizes a confusion matrix and holdout testing, but enables the replacement of this component if the need should arise (for example reviewer comments or another method of analysis is required). A component such as this enables a more uniform result analysis which could lead to greater understanding of the approaches and could give more insight into the achieved performance. It would enable a fairer comparison of different approaches scattered across the literature regarding result gathering and presentation.

VI. CONCLUSIONS

The goal of this paper was to consider the concepts of component based software development and to study the extent in which it is applicable in the area of pattern classification. At first, the definition and the descriptions of key terms are given. Main differences between the traditional object oriented programming paradigm and the component based development are demonstrated, while some prominent advantages and shortcomings of this, relatively novel, approach are listed. In order to employ this development method in pattern classification, key concepts of the pattern recognition area are explained. Feature extraction, classification and results analysis are the suggested areas in which component based development could be utilized. Based on the stated, an application is developed that offers the aforementioned segments of the classification process in the form of independent components. These components are proven to be upgradeable, exchangeable and replaceable. The mode of operation is presented for the provided solution and it demonstrated an easy component switching during operation. Adding components into the system is done through the use of reflection so the system does not have to know anything about the given component in advance, other than it implements the appropriate interfaces, whether required or provided ones.

Component based software development has proven to be an interesting concept in the area of pattern classification which is demonstrated in the proposed use case. The testing and analysis of the classification procedures or feature extraction methods is simplified considering the ease of replacement. For research purposes, if one was to develop a novel method, its implementation as a component inside this system enables fast and accurate comparison with other procedures and methods. Additionally, anyone can create such a component enabling very fast evaluation of results, and this is another concept present in CBD – third party components. Considering the advantages brought by CBD in the described application, it is easy to note the benefits it would bring to large enterprise systems or any system created under one of the agile development methodologies. Main disadvantages of this approach are the lack of standardization, long design phase time and very high demands on developer knowledge and experience. Since components have to be general, and yet offer specific functionalities, it takes a considerable amount of effort in the design stage to reach high quality standards. This could present a significant challenge for fast paced research, where new ideas are implemented daily. As this methodology matures and standard practices emerge, it is to be expected for the downsides to fade and CBD to become an unavoidable part of software development in the area of research, drawing from the ideas presented in this paper in the area of pattern classification.

[1] I. Crnkovic, “Component-based software engineering—new challenges in software development,” Software Focus, vol. 2, no. 4, 2001, pp.127–133.

[2] C. Szyperski, Component Software: Beyond Object-Oriented Programming, 2nd ed., Boston, MA, USA, Addison-Wesley Longman Publishing Co. Inc., 2002.

[3] I. Crnkovic, M. Chaudron, and S. Larsson, “Component-based development process and component lifecycle,” in Proceedings of International Conference on Software Engineering Advances, Tahiti, French Polynesia, October 29 – November 3, 2006, pp. 44–44.

[4] J. Lowy, Programming .NET Components, 2nd Edition. O’Reilly Media, Inc., 2005.

[5] G. T. Heineman and W. T. Councill, “Component-based software engineering,” Putting the Pieces Together, Addison-Westley, 2001.

[6] I. Crnkovic, S. Sentilles, A. Vulgarakis, and M. R. Chaudron, “A classification framework for software component models,” IEEE Transactions on Software Engineering, vol. 37, no. 5, 2011, pp. 593–615.

[7] R. Duda, P. Hart, and D. Stork, Pattern Classification. Wiley, 2001.

[8] H. Hu, C. Ottoc L. Xiaoming, A.K. Jain, "Demographic Estimation from Face Images: Human vs. Machine Performance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.37, no.6, 2015, pp.1148-1161.

[9] M. Weingant, H. M. Reynolds, A. Haworth, C. Mitchell, S. Williams and M.D. DiFranco, "Ensemble Prostate Tumor Classification in H&E Whole Slide Imaging via Stain Normalization and Cell Density Estimation," Machine Learning in Medical Imaging, 2015, pp. 280-287.

[10] X. Xie and M. Mirmehdi, “A Galaxy of Texture Features”, Imperial College Press, December 2008, pp. 375–406. [Online]. Available: http://www.cs.bris.ac.uk/Publications/Papers/2000853.pdf

[11] Martinović G., Bajer D. and Zorić B., “A Differential Evolution Approach to Dimensionality Reduction for Classification Needs”, International Journal of Applied Mathematics and Computer Science, Vol 24, No. 1, 2014, pp. 111-122.

[12] Bajer D., Zorić B. and Martinović, G., “Automatic Design of Radial Basis Function Networks Through Enhanced Differential Evolution”, Proceedings of the 10th International Conference on Hybrid Artificial Intelligence Systems, Bilbao, Spain, 22-24 June,.2015, pp.244-256.

[13] N. Japkowicz, M. Shah, Evaluating learning algorithms: a classification perspective, Cambridge University Press, 2011.

[14] M. N. Dan Popescu, Radu Dobrescu, “Texture classification and defect detection by statistical features,” International Journal of Circuits, Systems and Signal Processing, vol. 1, no. 1, 2007, pp. 79–84.

[15] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 7, 2002, pp. 971–987.

Extended Approach to Selecting a project-specific Reliability Growth Model

Jamal Krini, Abderrahim Krini, Ossmane Krini and Josef Börcsök

Department of Computer Architecture and System Programming
University of Kassel, Wilhelmshöher Allee 71 – 73, 34121 Kassel, GERMANY

Abstract: - The software reliability represents one of the most crucial aspects of functional safety. Software is used in almost any electronic systems. When a safety-critical system is considered, it has to be ensured that the software is able to fulfil the minimum safety requirements. If this is not the case, the software will have to be improved until the minimum requirements are fulfilled. Reliability growth models are used to determine the reliability of software. Depending on the area of application one specific model can deliver better prognosis results than another. This paper serves to introduce a methodology, which is provided for the purpose of selecting the most appropriate model.

Key-Words: - Functional safety, error-related hazard rate, mean value function, forecasting, Poisson model, program hazard rate, Sigmoid function, systematic errors, Shooman model, unknown errors, prediction, failure rate, reliability growth models, time-dependent models, software engineering, safety-related systems, prognosis, bathtub curve, early failures, normal failures, late failures, probability of default, failure density, MTTR, MTTF, MTBF, maximum likelihood estimation, estimation procedures, model parameters, effectiveness investigation model, quality criterion, absolute error.

1 Introduction

A safe and at the same time reliable way of software engineering is no easy matter, especially when it concerns safety-related systems. Therefore, the question arises of how to ensure the functional safety of software. In this context, for example, it is important to know where the aforesaid system is used. Companies that create safety-critical software have an enormous responsibility. The incidents of the recent years show how dangerous faulty software can be. On 3rd July 1988 an US warship shot down a rocket and according to the evaluation of the computer it should have hit an Iranian fighter jet. However, the plane that was hit was an Airbus A300.There were 290 passengers on board and all of them have lost their live. This and many other examples go to show that software errors can cost human lives. To counter this, the software has to be thoroughly tested. A variety of methods does exist that all serve to make software more secure. Here, the reliability growth models are to be used.

2 Prognosis

The term “prognosis” derives from the ancient Greek and essentially means “the prediction of”. A prognosis serves to predict the future development of a particular phenomenon or specific situation, taking account of any relevant and known factors. It is only an assessment. The reliability prediction represents a very important type of prediction. Any reliability prediction can be described through the help of statistical models. The reliability growth models were initially used for the hardware sector .Only later was the basic idea transferred from the hardware to the software. There do, however, a number of differences exist here. While hardware can age, for example, software does not imply an aging process. Restrictions regarding the reliability in the software area have to be seen as a result of errors that have occurred within the software creation. The consequences of those errors are - as well as failures due to wear or aging - not predictable. The failure behavior can be illustrated by the use of the bathtub curve (see Figure 1 and Figure 2).

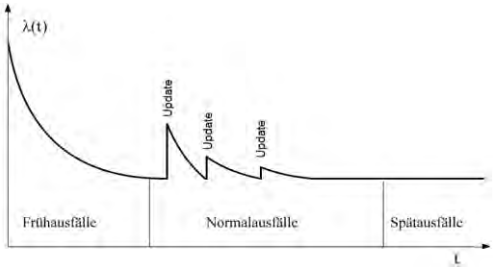


Figure 1: Failure behavior of software

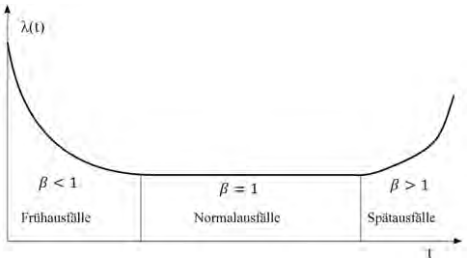


Figure 2: Failure behavior of hardware

The curve depicts the following three phases:

- Early phase (early failures),
- Use phase (normal failures),
- Wear phase (late failures).

Early failures:

In the course of the life-cycle in the early phase the number of failures decreases due to the testing. The failures of the hardware can result from manufacturing defects, material defects and operating errors. The failures of the software can, for example, result from programming errors.

Normal failures:

The use phase of the hardware results in virtually no wear. The failure rate is almost constant. The causes of the failures are purely coincidental. This is not true for the software. Due to the fact that in the course of the use phase continuously new updates are provided as time goes by the failure rate increases or decreases.

Late failures:

In the wear phase, the failures of the hardware are increasing over the lifespan of the hardware. The failures do mainly result from ageing, wear and signs of fatigue. The failure rate increases. On the contrary, the failure rate of the software remains constant. The reason for this is that software does not show any signs of wear due to ageing or signs of fatigue.

The state of a technical system can rather be that it is operational or that it is not:

$$x(t) = \begin{cases} 1, \text{if system is at time t operational} \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

The life-cycle of a technical system comprises the period of time T from the initial commissioning (t=0) to the failure (t=T). Consequently, the following applies to the state of a system:

$$x(t) = \begin{cases} 1 \text{ for } t < T \\ 0 \text{ for } t \geq T \end{cases} \quad (2)$$

Reliability represents an important figure of functional safety. It is a measure for the probability of a provision of a function within a certain defined time period. The following applies to reliability:

$$R(t) = e^{-\lambda t} \quad (3)$$

The probability of default represents another important figure of the functional safety. It indicates the probability of failure or the probability of the destruction of a system. This applies to the failure probability:

$$F(t) = 1 - e^{-\lambda t} \quad (4)$$

The density of failure is represented by the time derivation of the probability of default.

$$f(t) = \frac{dF(t)}{dt} \quad (5)$$

The reliability and the probability of default do always result in the sum 1.

$$F(t) + R(t) = 1 \quad (6)$$

Figure 3 serves to illustrate this fact.

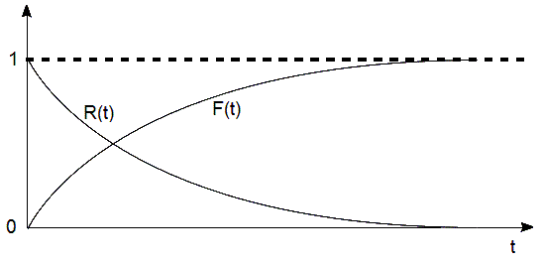


Figure 3: Reliability and probability of default

The MTTF represents the average operating time until a failure.

$$MTTF = \int_0^\infty t \cdot f(t)dt = \int_0^\infty R(t)dt \quad (7)$$

The MTBF, which is also known as "Mean Time Between Failure", represents the mean time between two failures. The MTTR stands for "mean time to repair" issues. This is precisely in line with the time that is needed to detect and fix an error. Figure 4 illustrates the relationship between MTTF, MTTR and MTBF.

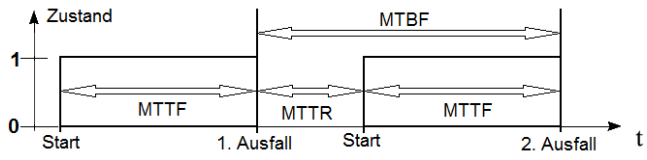


Figure 4: Relationship between MTTF, MTTR and MTBF

3 New Research Approach

Due to the fact that software has become so complex that it is no longer possible to control all kinds of influences on the reliability of the software. However, software could be tested up to a certain predefined point of safety through an investment of much time and money but this effort would quickly reach the economically

acceptable limits. A great number of reliability growth models has been published over the last few years. The models vary according to their underlying driving resource, their type of influence and the estimated lifetime distribution. Most of the reliability growth models are simply extensions of other models. The main categories of the reliability growth models are:

- Time-dependent models (data or test results with a temporal reference)
    - Shooman Model
    - Jelinski Moranda Model
    - Poisson Model
    - Schick Wolverton Model
- Time-independent models (data or test results with no temporal reference)
    - Mill Model
    - Lipow Model
    - Rudner Model
    - Nelson Model

An important aspect of the reliability prediction is the estimation procedure. The estimation procedure is used for a determination of the model parameters. The most popular methods are:

- Moment Method (method of moments)
- Least Squares Method (method of least squares)
- Maximum Likelihood Estimation (MLE)

The maximum likelihood estimation (MLE) is considered as the most effective and most used method. It is defined as follows:

$$L = L(\phi) = \prod_{i=1}^{n} P(X_i = x_i|\phi)$$

In order to carry out an estimation of the model parameters according to MLE, the following steps have to be taken:

- Establishing the likelihood function
- Forming the natural logarithm of the likelihood function
- Deriving the likelihood function with regard to any parameter that is to be estimated
- Calculating the maximum expectation of the distribution parameter or parameters

The reliability models are used in many areas of application. Depending on the area of application it may occur that one model may deliver better prognosis results than the other model. Selecting the most appropriate model has the highest priority. In order to be able to choose an appropriate model it has solely to be resorted to empirical values. However, it is a matter of fact that none of the models is able to forecast good predictions for every single data record. Moreover, it cannot be decided in advance which of the models fits best with the corresponding prognosis. The more appropriate the selected model, the lower is the forecast error. The goal is to create an approach that makes it possible to select the most proven model so that the forecast error can be kept to a minimum. A new model is presented here, which is called “Probatermittlermodell” (Effectiveness Investigation Model).

4 Application of the New Research Approach

It is of crucial importance that a quality criterion of a model has to be defined for an application. Thereby, the exact absolute error is determined. The structure of the research approach consists of the following 4 channels (see figure 5):

- Channel 1
    - Splitting of data in
        - 1. data set
        - 2. data set
- Channel 2
    - Reliability prediction by means of model 1
- Channel 3
    - Reliability prediction by means of model 2
- Channel 4
    - Data evaluation

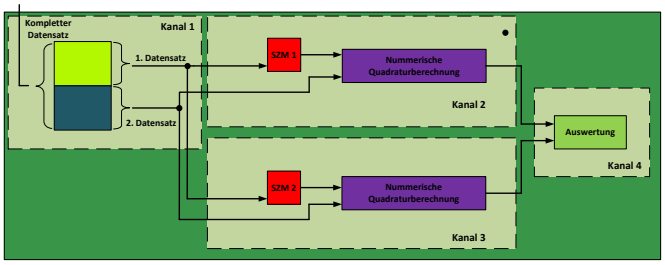


Figure 5: overall concept

The task of channel 1 is to split the incoming default data into two data sets (1. data set, 2. data set). The first data set serves as a basis for forecasting and the second data set is used for a determination of the quality criterion. Both data sets are fed to the channel 2 and channel 3. Channel 2 and channel 3 have the same structure. The main difference is that there are different reliability models in the channels. First of all a prediction is made by means of the first default data (1. data set) of the entire default data package (see table 2). The points in time of the next failures are estimated here. In the ideal case, the result would correspond to the points in time of the 2. data package (see table 3). However, in reality, there is some variation. The extent of the deviation corresponds to the quality criterion of the model. Furthermore, the extent of the deviation is defined by the absolute error. For this purpose, the deviation of any point between the actual value a_t and the forecast value a_p is determined.

Subsequently, all the deviations are added up. For the deviation of a point applies:

$$A_x = |a_{x_t} - a_{x_p}| \qquad \text{Eq. (7.16)}$$

The index x corresponds to the number of the present default data. For a calculation of the absolute error the total of the individual deviations has to be made up:

$$A = \sum_{i=1}^{x} A_i \qquad \text{Eq. (7.17)}$$

For simulation purposes, 280 default data are considered. The used default data are shown in the table 1.

Table 1

36	44	59	74	85	94	95	99	113	117	137
140	146	149	156	160	164	179	190	191	201	215
240	241	249	250	262	286	298	307	308	336	339
355	364	377	381	409	430	448	474	481	496	499
513	529	532	545	549	554	582	583	603	652	664
665	724	728	739	744	759	785	807	834	873	886
909	913	914	921	930	933	935	943	946	959	970
989	994	1031	1033	1043	1049	1075	1078	1091	1105	1116
1120	1122	1123	1135	1141	1167	1179	1181	1183	1188	1192
1204	1208	1210	1234	1244	1245	1316	1318	1389	1390	1411
1415	1417	1428	1441	1447	1448	1460	1461	1467	1493	1499
1531	1532	1544	1547	1565	1586	1642	1643	1686	1696	1744
1762	1781	1792	1814	1819	1821	1826	1837	1840	1857	1861
1903	1909	1935	1941	1942	1949	1963	1971	1982	1992	2004
2005	2008	2012	2019	2023	2030	2035	2049	2060	2067	2068
2079	2092	2107	2138	2151	2153	2155	2173	2174	2183	2201
2219	2220	2233	2246	2255	2265	2270	2290	2292	2304	2320
2356	2363	2364	2367	2372	2420	2422	2427	2453	2475	2493
2496	2503	2522	2530	2539	2551	2571	2577	2580	2582	2586
2592	2629	2634	2641	2659	2663	2718	2769	2770	2783	2793
2803	2826	2827	2845	2850	2866	2878	2882	2907	2952	2961
2994	3001	3003	3012	3014	3033	3038	3040	3053	3062	3064
3073	3074	3093	3113	3122	3125	3162	3169	3179	3180	3210
3216	3221	3226	3240	3246	3247	3257	3269	3286	3304	3350
3365	3389	3392	3401	3456	3469	3482	3489	3517	3524	3538
3540	3549	3553	3560	3585						

The splitting into two data sets has to take place before the default data are made available to the reliability models. The first data set as well as the second data set consists of an amount of 140 default data. It has to be taken into account that at the second data set the MTTF value has to be reset because the second data set has to be independent of the first data set. The table 2 and table 3 serve to illustrate the first and the second data set.

Table 2

36	44	59	74	85	94	95	99	113	117	137
140	146	149	156	160	164	179	190	191	201	215
240	241	249	250	262	286	298	307	308	336	339
355	364	377	381	409	430	448	474	481	496	499
513	529	532	545	549	554	582	583	603	652	664
665	724	728	739	744	759	785	807	834	873	886
909	913	914	921	930	933	935	943	946	959	970
989	994	1031	1033	1043	1049	1075	1078	1091	1105	1116
1120	1122	1123	1135	1141	1167	1179	1181	1183	1188	1192
1204	1208	1210	1234	1244	1245	1316	1318	1389	1390	1411
1415	1417	1428	1441	1447	1448	1460	1461	1467	1493	1499
1531	1532	1544	1547	1565	1586	1642	1643	1686	1696	1744
1762	1781	1792	1814	1819	1821	1826	1837			

Table 3

3	20	24	66	72	98	104	105	112	126	134
145	155	167	168	171	175	182	186	193	198	212
223	230	231	242	255	270	301	314	316	318	336
337	346	364	382	383	396	409	418	428	433	453
455	467	483	519	526	527	530	535	583	585	590
616	638	656	659	666	685	693	702	714	734	740
743	745	749	755	792	797	804	822	826	881	932
933	946	956	966	989	990	1008	1013	1029	1041	1045
1070	1115	1124	1157	1164	1166	1175	1177	1196	1201	1203
1216	1225	1227	1236	1237	1256	1276	1285	1288	1325	1332
1342	1343	1373	1379	1384	1389	1403	1409	1410	1420	1432
1449	1467	1513	1528	1552	1555	1564	1619	1632	1645	1652
1680	1687	1701	1703	1712	1716	1723	1748			

Only the first data set is made available to the reliability models. A prognosis is made through the use of the first data set. For the prognosis the following values result for the reliability model 1:

$$\tilde{\mathbf{u}}_0 = \mathbf{803,0111}$$

$$\tilde{\boldsymbol{\phi}} = \mathbf{1,0445 * 10^{-4}}$$

Therefore, what follows for the mean value function is:

$$\mu(t) = \tilde{u}_0 \cdot \left[1 - e^{-\tilde{\phi} \cdot t}\right]$$

$$\mu(t) = 803{,}0111 \cdot \left[1 - e^{-1{,}0445 * 10^{-4} \cdot t}\right]$$

The estimated mean value function is illustrated in figure 6.

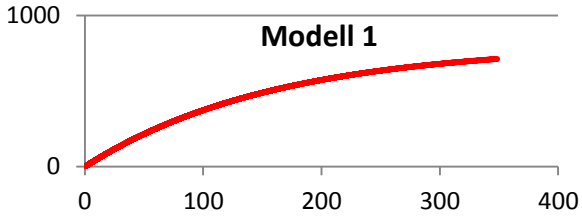


Figure 6: average function

The prognosis default data are shown in table 4.

Table 4

0,251	1,675	2,010	5,516	6,016	8,177	8,675	8,758
9,339	10,498	11,160	12,070	12,895	13,885	13,968	14,215
14,544	15,120	15,450	16,025	16,436	17,585	18,487	19,061
19,143	20,043	21,105	22,329	24,853	25,909	26,071	26,233
27,693	27,774	28,502	29,957	31,409	31,489	32,536	33,582
34,305	35,107	35,508	37,110	37,270	38,229	39,506	42,372
42,927	43,007	43,245	43,642	47,439	47,597	47,992	50,039
51,767	53,178	53,413	53,961	55,446	56,071	56,773	57,707
59,263	59,729	59,961	60,117	60,427	60,892	63,755	64,141
64,681	66,068	66,375	70,595	74,486	74,562	75,551	76,310
77,069	78,811	78,886	80,246	80,624	81,830	82,733	83,034
84,912	88,279	88,951	91,408	91,928	92,076	92,744	92,893
94,301	94,671	94,819	95,780	96,444	96,592	97,255	97,329
98,728	100,19	100,85	101,07	103,78	104,29	105,02	105,09
107,28	107,71	108,08	108,44	109,45	109,89	109,96	110,69
111,55	112,78	114,08	117,38	118,45	120,17	120,38	121,02
124,93	125,85	126,77	127,26	129,23	129,73	130,71	130,85
131,48	131,76	132,2	134,00				

For the prognosis of reliability model 2, however, the following values do result:

$$\tilde{\mathbf{u}}_0 = \mathbf{377,5373}$$

$$\tilde{\boldsymbol{\alpha}} = \mathbf{0,0883}$$

$$\tilde{\boldsymbol{\beta}} = \mathbf{74,8982}$$

This applies to the mean value function:

$$\mu(t) = \tilde{u}_0 \cdot \left[1 - \left(\frac{\tilde{\beta}}{\tilde{\beta} + t}\right)^{\tilde{\alpha}}\right]$$

$$\mu(t) = 377{,}5373 \cdot \left[1 - \left(\frac{74{,}8982}{74{,}8982 + t}\right)^{0{,}0883}\right]$$

The estimated mean value function for the reliability model 2 is depicted in fig. 7.

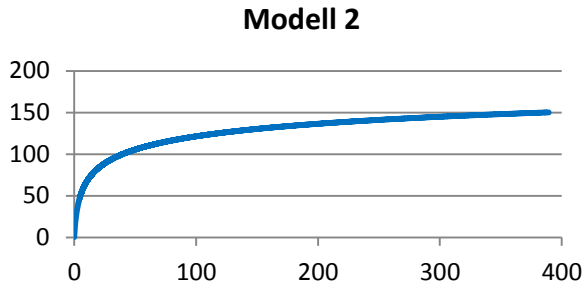


Figure 7: average function

The prognosis default data are illustrated in table 5.

Table 5

1,306	7,808	9,153	20,488	21,801	26,883	27,937	28,110
29,285	31,500	32,691	34,250	35,595	37,128	37,252	37,620
38,104	38,931	39,393	40,183	40,733	42,218	43,330	44,015
44,111	45,149	46,327	47,625	50,123	51,104	51,252	51,399
52,687	52,756	53,376	54,572	55,716	55,778	56,573	57,344
57,864	58,431	58,709	59,795	59,901	60,528	61,342	63,083
63,408	63,454	63,592	63,820	65,912	65,995	66,203	67,256
68,113	68,794	68,905	69,164	69,853	70,137	70,453	70,869
71,546	71,745	71,845	71,911	72,042	72,238	73,411	73,566
73,780	74,323	74,442	76,024	77,405	77,431	77,771	78,029
78,284	78,861	78,886	79,327	79,448	79,833	80,117	80,211
80,790	81,798	81,995	82,703	82,850	82,892	83,080	83,122
83,513	83,615	83,655	83,918	84,098	84,138	84,316	84,336
84,708	85,093	85,265	85,322	86,012	86,140	86,323	86,341
86,879	86,985	87,073	87,161	87,405	87,508	87,526	87,697
87,902	88,189	88,489	89,238	89,477	89,855	89,902	90,042
90,878	91,072	91,264	91,366	91,772	91,873	92,072	92,101
92,228	92,284	92,382	92,730				

After both prognosis functions and the corresponding default data have been determined, the quality criterion can be defined by means of the absolute error and therefore the actual default data are required. The data are obtained from the second data set (see table 3).

The following equation is used to calculate the absolute error for model 1:

$$A_{Modell\ 1} = \sum_{i=1}^{x} \left| a_{x_t} - a_{x_p} \right|$$

$$A_{Modell\ 1} = 710{,}458$$

This equation is used to calculate the absolute error for model 2:

$$A_{Modell\ 2} = \sum_{i=1}^{x} \left| a_{x_t} - a_{x_p} \right|$$

$$A_{Modell\ 2} = 2549{,}02$$

Due to the fact that the absolute error of model 1 is smaller than the absolute error of model 2, model 1 is considered as the most appropriate. Using the numerical squaring method leads to the same result.

5 Conclusion

Software reliability models are counted among the principal aspects of functional safety. It is therefore essential to continue researching in that field so that further and better prediction models can be created. This paper introduces a procedure model, which should lead to a better prognosis. The default data that are available have to be split so that a prognosis for the first data package can be made and subsequently can be compared with the original data. This enables statements about the quality. The more suitable the chosen reliability model, the more precise will be the prognosis.

6 Sources

[1] Hou YZ. & Yu BY., 2011. A novel software reliability model with the decrease-rate of removing erros. Proceesings of the 2011 International Conference on Machine Learning and Cybernetics.

[2] Shigeru Yamada, 2013. Software Reliability Modeling: Fundamentals and Applications 2013.

[3] J. Börcsök, 2004. Electronic Safety Systems, Hardware Concepts, Models, and Calculation. Heidelberg, Germany 2004.

[4] J.D. Musa, Software Reliability Engineering 2004. More Reliable Software Faster and Cheaper. United States of America 2004.

[5] J. Börcsök, O. Krini, 2009. Principle Software Reliability Analysis with different Failure Rate Models. IEEE Int. Symp. ICAT 2009

[6] B. Lindqvist, A. Doksum, 2003. Mathematical and Statistical Method in Reliability. Covent Garden, London 2003.

[7] O.Krini, J.Börcsök: ”Different Approaches to Predict Software Reliability”, Proceedings of 24th international congress, 2011

[8] Huan-Jyh Shyur:” A stochastic software reliability model with imperfect-debugging and change-point”, The Journal of systems and software, 2003.

[9] IEEE Recommended Practice on Software Reliability, Standards Committee of the IEEE Reliability Society, 2008.

[10] X.Zhang, X.Teng, H.Pham: “Considering Fault Removal Efficiency in Software Reliability Assessment”, IEEE vol.33, 2003.

[11] A. Birolini, “Reliability Engineering: Theory and Practice”, Springer Verlag, Heidelberg, 2007.

[12] Jose Maria Martinez and Michael Gonzalez Harbour, "RT-EP: A Fixed-Priority Real Time Communication Protocol over Standard Ethernet". Proceedings of the 9th International- Conference on Reliable Software Technologies, Ada-Europe, York, Springer, LNCS-3555, June, 2005.

[13] J.L. Medina Pasaje, M. Gonzalez Harbour, J.M. Drake Moyano, "MAST Real-Time View: Graphic UML Tool for Modeling Object-Oriented Real-Time Systems". Proceedings of the 22th IEEE Real-Time Systems Symposium, London, UK, December, 2001, ISBN:0-7695-1420-0, pp. 245,256.

[14] J. Börcsök, 2004. Electronic Safety Systems, Hardware Concepts, Models, and Calculation. Heidelberg, Germany 2004.

[15] J.D. Musa, Software Reliability Engineering 2004. More Reliable Software Faster and Cheaper. United States of America 2004.

[16] N. F. Schneidewind, Software Reliability Engineering 1992. Minimizing risk in Applying Metrics on Multiple Projects, Proc. IEEE Int. Symp, pp. 173-179.

[17] J. Börcsök, O. Krini, 2009. Principle Software Reliability Analysis with different Failure Rate Models. IEEE Int. Symp. ICAT 2009

[18] B. Lindqvist, A. Doksum, 2003. Mathematical and Statistical Method in Reliability. Covent Garden, London 2003.

[19] M. Takahashi and Y. Kamayachi, IEEE Computer Society, 1991. An Empirical Study of a Model for Program Error Prediction, in Software Reliability Models. pp. 71-77.

[20] Musa, J.D.: “Software Reliability Engineering”, McGraw-Hill, New York, 1998.

Embedded Linux Controlled Sensor Network

M. Saari *, A.M. Baharudin **, P. Sillberg *, P. Rantanen * and J. Soini*
* Tampere University of Technology/Pori Department, Pori, Finland
** Keio University, Tokyo, Japan
mika.saari@tut.fi

Abstract - This study utilizes a simple model for constructing sensor nodes – master controller combinations in the Internet of Things. The model combines hardware and software for embedded systems which measure a predefined set of parameters. The master controller manages several sensor nodes, collects data from them and provides data for clients. The paper introduces a proof-of-concept implementation based on the model. The implementation uses an embedded Linux based small computer and microcontroller based sensor nodes in the context of condition measurement, and represents a way to use wireless data transfer between controller and nodes. The target of this study was to test the model, to determine how well a cost-efficient single-board computer could be used to gather sensory data from several sensor nodes, and how this data can be provided for clients over the public Internet.

I. INTRODUCTION

Wireless sensor networks have developed at a fast pace in recent years and have also been one of the major focuses of research in wireless technology. This rapid development has been facilitated by the evolution of electronics miniaturization, growth in performance and energy efficiency, and the development of protocols. Through the fast performance, optimization and miniaturization technology of hardware, the sensors that collect environmental information from the surroundings have been miniaturized. The developments of ever smaller processors and falling prices have enabled brand-new uses for electronics.

Embedded systems are typically designed for a specific application or purpose and come in a variety of shapes and sizes. Moreover, embedded systems are suitable for application systems with strict requirements for functionality, reliability, cost, size, and power consumption. One application of embedded systems that is in constant demand and under continuous development is telecommunications components. The price of wireless components in particular has fallen so much that experts can see clear potential there. In regards to rising energy costs and energy savings, regulation and control technology for buildings would seem to offer considerable potential for embedded solutions. In addition, the property, safety and surveillance technology sector offers growth in the area of embedded systems.

Because of the numerous economic and technical benefits, embedded systems usually adopt embedded Linux as the operating system. This paper introduces a proof-of-concept implementation that uses a small embedded Linux based computer and microcontroller based sensor nodes in the context of condition measurement, and represents a way to use wireless data transfer between controller and nodes. The paper introduces the implementation of a sensor network solution for sensor data collection and transmission. This study was performed in intensive collaboration between Tampere University of Technology (TUT) in Finland and Keio University in Japan. The implementation was created in the TUT Pori department in 2015.

The construction of this paper is as follows. In section II, we review related work. The model of our system is introduced in Section III. In section IV we describe the proof-of-concept implementation. Section V includes a discussion and suggestions for future research on the topic and finally, section VI summarizes the study.

II. BACKGROUND

The Internet of Things (IoT) is the expansion of Internet services, which connects everyday physical objects to the network. This connection between network and physical objects makes it possible to access remote sensor data and to control the physical world from a distance. The first mention of the term IoT is said to have come from Kevin Ashton in 1999. There are also several books about IoT. The Amazon web store found 3114 instances with the phrase “Internet of Things” in January 2016. In this paper we introduce one implementation of the IoT-world.

In our earlier research [1], the focus was on collecting data from one sensor packet which was connected by wire directly to a single-board computer. This was done by using embedded Linux and BeagleBone Black hardware, which is a credit card–sized single-board computer similar to Raspberry Pi. The focus of the research was to collect data and to deliver it over the network. The plan was to build several of these data collector service computers and use them in a specific alarm system.

In this research the focus has been redirected toward the Wireless Sensor Network (WSN) type of solution. A survey conducted in 2002 compiled the basic features of sensor networks [2]. The aim was to collect data from several points to one master node. The collected data are provided to the network. This data could be used in smart house type construction. For example [3] presented a small smart house system, based on one Arduino development board. In that research, the proposed system monitored the environment and also controlled lights, temperature, alarms, and other household appliances. In Finland, where heating is necessary most of the year, a

low cost sensor network could be used for monitoring and controlling the heating and air ventilation systems.

The research started with modeling the construction of a system with one master and several nodes. We introduced an abstract model of the sensor network. There are several studies about more complex models designed for IoT. For example, one research study [4] introduced the Physical Service model, where they described a device model, resource model, and service model. Another piece of research [5] introduced a Wireless Sensor network abstraction model that has three levels: node abstraction, network abstraction and infrastructure abstraction. If this is compared to our research, we focused on the simpler infrastructure abstraction and we used low-cost off-the-shelf equipment available from a local store for our example implementation. The choice of hardware supports rapid prototyping and the test configurations are easily repeatable. For example, [6] and [7] use the Raspberry Pi kind of approach in their wireless sensor network gateway prototyping. The model itself does no set any limitations but one target was to use low-cost and easy-to-use hardware. Also, open hardware and open source software were the selection criteria for the components of the implementation.

The security issues are an important part of wireless sensor networks – How to prevent information leakage or whether the transferred data is vulnerable. The research about security issues of wireless data communication was made by [8]. They focused on ZigBee [9], which we also used in the proof-of-concept implementation.

III. Model of the Sensor Network

The model of the sensor network is shown in Fig. 1. The s*ensor nodes* are used to gather measurements such as temperature, humidity and air pressure from the environment. The number of sensor nodes is not limited to any particular amount and in theory there can be an unlimited number of nodes. The nodes are meant for simple tasks which consist only of passive data collection without the need for advanced data analysis or preprocessing. The *Master node*, in general, has the capability to run a full-feature operating system, and its job is to control the sensor nodes and manage the data collection process. The collected data is stored on the master and provided for client devices over the public Internet or using a more restricted local network.



Figure 1 Model of sensor network.

In an earlier research study, we concluded that when using cost-effective single-board computers, such as the BeagleBone, the limited computing performance could cause issues [1]. Problems arise when a single computer (i.e. the master node) runs the web service and database, and also measures the sensor data, in which case the requirements for the hardware can simply be too much. To mitigate this issue, separate sensor nodes are added, which manage the actual process of collecting the data. This way the number of sensors can be increased without additional strain on the limited processing capabilities of the master node.

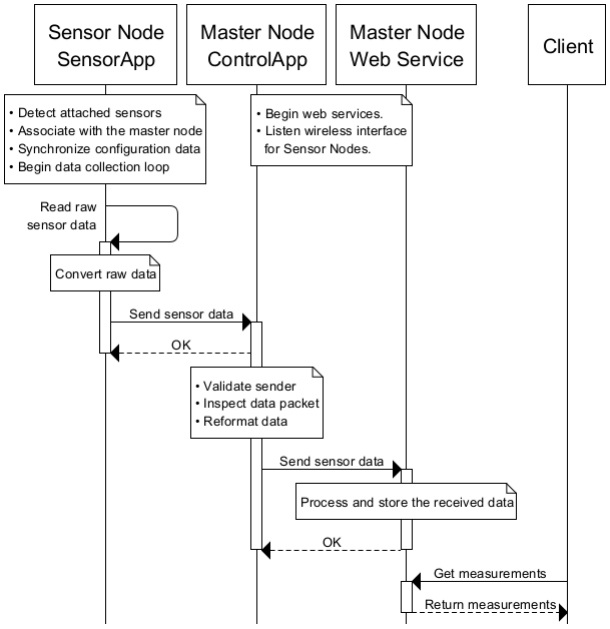


Figure 2 Sequence of operations.

The collection, storing, providing and usage of the sensor data are carried out by four components – or roles. Fig. 2 shows the sequence of operations of each of the components. The *Client* is the consumer of the collected data. The client retrieves the data by accessing the *Master Node*, which consists of two software components to facilitate the collection, the storage and delivery of the sensor data. The master node is a centralized gateway for the *Sensor Nodes*. They include the software that collects and relays the sensor data to the master node. In our prototype system, the client is simply a web browser accessing the data for visualization purposes. The rest of this section describes the details of the model and the three other components – SensorApp, ControlApp, and Web Service.

A. Sensor Nodes

The sensor nodes are devices that provide the raw data. They should be inexpensive, easy to deploy and replaceable. Each sensor node may have a different sensor configuration, and they should operate independently from other sensor nodes.

It was determined that the sensor node should implement the following features (steps denoted in *italics* are not part of the prototype implementation):

1) *Detect attached sensors.*
2) *Associate with the master node and synchronize configuration data.*
3) Read raw data from the attached sensors.
4) Convert the raw data to a transferable format.

5) Send the data over a wireless network to the master node.

6) Sleep and start over from the third step.

The choice of data format used between the sensor nodes and the master node depends on: the available network, bandwidth, and computational resources; developer preference; and use case. Especially for testing and debugging, a human readable format is recommended, but not strictly required.

B. *Master Node*

The requirements for the master node are higher than for a sensor node. The master node is specified to be a gateway between data consumers (clients) and data producers (sensor nodes). The master node should be powerful enough to execute multiple processes, such as:

- communications with sensor nodes,
- database operations and
- web services.

As the master node is powered on, it will run the code to listen for sensor nodes to associate and send the data. It will also start the *Web Service* for processing the data and delivering it to the outside world over the Internet.

When the master node receives the data packet, it will inspect the address and details of the sender, compare it to the associated sensor nodes, and parse the data. As a security measure, the master node could be instructed to discard all data sent by an unknown sensor node or by a sensor node that does not provide the correct configuration.

If the data received from sensor nodes is not in a structured format it should be transformed to allow easier use. Commonly used formats for Internet applications are JSON and XML, and both are good choices for delivery of data to clients.

IV. EXAMPLE SYSTEM IMPLEMENTATION

The example works as a proof-of-concept implementation, and we do not have a particular use case for it as such. In our earlier publication [1], we presented a system for environment sensing, which utilizes various sensors for collecting measurements (e.g. temperature, humidity), and the case is still equally valid. In fact, the model presented in this paper can be seen as an improvement on the system described, and the model has also been developed based on the findings of the studies performed earlier. The primary purpose of the prototype solution was to show that a feasible system for remotely collecting sensor data can be constructed based on the model.

For the master node we chose an Intel Galileo Gen 2 Development Board [10], which is based on Intel x86 architecture. The Galileo is a single-board computer similar to Beagle Bone or Raspberry Pi. Each of the three boards has the features required to implement our example use case, and in principle, either of the other boards could have been chosen instead. We chose Arduino Uno for the sensor nodes. Galileo, BeagleBone and Raspberry Pi belong to a higher price category than Arduino Uno. Because of the lower price, Arduino Uno makes a more feasible platform for numerous sensor nodes.

Galileo includes a 10/100 Ethernet connection, which we use to connect to the public Internet for the purpose of delivering the collected data to the clients. The web services are provided by a server built on Node.js [11]. The Node.js instance can be somewhat resource intensive, but it seemed to work acceptably in our tests with a small number of concurrent users.

The internal memory of the Galileo is quite limited, so we installed the Yocto Linux operation system [12] on a microSD card to provide a larger storage space. This is especially important, as the database is also located on the master node, and the built-in memory may not be capable of holding all of the collected data. Additionally, the embedded Linux based operation systems usually include common Linux software such as Secure Shell (SSH) server for easier configuration of the node and they are also capable of running other applications primarily targeted for full-feature desktop or server computers (e.g. Apache Tomcat, Java virtual machine).

One advantage of the Galileo board is the support for ready-made hardware expansions shields designed for Arduino. In our example implementation we used the less powerful Arduino Uno [13] for the sensor nodes, which allows us to use the same expansion components for both the master node and the sensor nodes. In this construction we used the Arduino Wireless Proto Shield with XBee modules [14] to implement the communication between the master node and the sensor nodes. The XBee modules are based on ZigBee and are designed for low-power wireless networks. The expansion modules make it possible to add and remove components easily reducing the need for soldering.

Communication with XBee can be done using the basic AT or the more advanced API mode [15]. In our implementation, the XBee API mode was utilized. Using the API mode with packet communication allows the transport layer to handle collision situations and possible data corruption.

Communication between sensor nodes and the master was tested indoors for finding out the average operation range. At a range of one to two rooms (about 15 to 20 meters) the communication did not drop any packets. At larger distances (about 30 to 40 meters) with a few walls between the sensor node and the master node, some packages were dropped, but the communication still worked at an acceptable level. When the distance was about 50 meters and there were several walls no packets were received. The most significant factor of operating range was the thickness and amount of walls. In the worst case, no packets were received when the sensor was on another floor and the distance was less than 30 meters. Our test environment was an old factory building converted to office use. The walls and floors are somewhat thicker than in an average building, which could affect the results.

We also ran stress tests against the master node's web service interface. The concurrent communications between master node and sensor nodes had only a minor effect on the overall performance of the web service. Based on our observations the Galileo board reserved about half of the CPU time for the Arduino process that performed the communication between the nodes. For use cases with low amount of clients, the limited CPU capabilities were not a problem. With a larger amount of concurrent clients (more than 50) the service experienced a noticeable increase in response times. Even higher amount of clients (more than 150) the latency became excessive and the server occasionally dropped the client connections.

Each of our sensor nodes contains three types of sensors: an air pressure sensor [16]; a photoconductive cell luminosity sensor (NSL-19M51) [17]; and a combined humidity and temperature sensor (AM2302) [18]. There are plenty of community created libraries and drivers for interacting with the attached hardware on Arduino, though in our case, we had some difficulties in finding useful code examples of how to utilize the sensor components we had chosen.

For the visualization of the data we utilized the methods presented in [19]. Importantly, the visualization is performed by the client's web browser displaying the JavaScript-based web page hosted on the master node. This approach is useful as it conserves the limited resources of the master node by off-loading the visualization work-load to the client side.

V. DISCUSSION

The model presented in this paper does not define the requirements for data encryption. The reason is that whether encryption is required or not largely depends on the use case. In our case, we chose ZigBee for data transmission between the master and the sensor nodes. ZigBee provides encryption by default, which is strong enough for our case. In fact, in our example use case, no encryption would be required simply because the data is not sensitive - for example, temperature and humidity can be measured by anyone, simply by entering the sensor location (room). If the sensor nodes autonomously send the details to the master node, and the master does not actively control the sensors, the requirements for encryption are lower than in the case where the master node actually controls the slaves. If any control data – orders on if, what, and how often – are sent by the master node, more thought should be put into the encryption, as well as into the authentication methods to reduce the risk of malicious use of the nodes. Often the chosen transmission method (e.g. ZigBee, Bluetooth and WLAN) can offer hardware based authentication and encryption options fit for most cases.

Another equally important issue not discussed in this paper is energy efficiency. The components chosen for the example system have low power consumption, but we have not performed extensive measurements for the power usage of the implemented system. Similarly to encryption, the model defines the requirements. In practice, the master node can usually be placed in a location, which has access to a constant power source, but the remote nodes may need to be run on battery power. For this reason one should carefully choose which components to use in the remote sensors to minimize the power consumption. In addition to component choices, the energy efficiency can be improved by the device software. The measurements performed by the sensors do not necessarily use much power, and most of the energy is spent on the wireless transmission of data. In our case, we control the power consumption by limiting the frequency and number of data transfers. In this case, there is no need to gather the measurements strictly in real-time. This makes it possible either to take measurements at a more relaxed pace (for example, every few minutes) or take measurements more often, but send the results infrequently in larger result sets. If control data is sent between the master and the sensors, the master can also control the rate of transmission. Also, if the data is preprocessed on the sensor nodes, the nodes themselves can make simple decisions on the frequency of communication. For example, if the nodes detect that there is a larger change in the measured values, there might be a need for more frequent transmission of data, but if the values stay the same, the data can be sent less frequently. In principle, there is no need to transfer any data as long as there is no change in the measurements, although in practice, status checks between the master and the sensor nodes should be performed to ensure that the sensors are alive and well.

One possible future research topic this paper does not analyze in depth is the matter of scalability. From the model's point of view, there can be an unlimited amount of sensor nodes for each master node, but often the chosen technologies pose limitations on the actual number of devices. In our tests the networks have been relatively small with a single master and only a few (less than ten) sensor nodes. With these kinds of small configurations the model has been proven to work, but it would be interesting to see what problems would arise in larger networks. Of course, reaching the technical device limits for the communication methods for testing purposes may be difficult simply because the required number of devices can be so high that it would cost too much to ever acquire enough sensor nodes. In fact, it is possible that the master node's capacity to process and store the received data would run out sooner simply because of the limited computing performance of the master's hardware.

VI. SUMMARY

This paper introduced a model for sensor networks used for gathering and distributing sensor measurements. The model consists of several sensor nodes and of a master node. The sensor nodes collect the raw sensor data, and the master node gathers the data from each of the sensor nodes, and provides the data for clients in a structured format. A proof-of-concept implementation based on the model was also introduced.

REFERENCES

[1] M. Saari, P. Sillberg, P. Rantanen, J. Soini, and H. Fukai, "Data collector service - practical approach with embedded linux," in 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015, pp. 1037–1041.

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: a survey,” Comput. Networks, vol. 38, no. 4, Mar. 2002, pp. 393–422.

[3] A. Adriansyah and A. W. Dani, “Design of Small Smart Home system based on Arduino,” in 2014 Electrical Power, Electronics, Communicatons, Control and Informatics Seminar (EECCIS), 2014, pp. 121–125.

[4] X. Jin, S. Chun, J. Jung, and K.-H. Lee, “IoT Service Selection Based on Physical Service Model and Absolute Dominance Relationship,” 2014 IEEE 7th Int. Conf. Serv. Comput. Appl., 2014, pp. 65–72.

[5] T. Laukkarinen, Abstracting Application Development for Resource Constrained Wireless Sensor Networks. Tampere, 2015.

[6] C. P. Kruger, A. M. Abu-Mahfouz, and G. P. Hancke, “Rapid prototyping of a wireless sensor network gateway for the internet of things using off-the-shelf components,” Proc. IEEE Int. Conf. Ind. Technol., vol. 2015-June, no. June, 2015, pp. 1926–1931.

[7] Lukas, W. A. Tanumihardja, and E. Gunawan, “On the application of IoT: Monitoring of troughs water level using WSN,” in 2015 IEEE Conference on Wireless Sensors (ICWiSe), 2015, pp. 58–62.

[8] J. Dos Santos, C. Hennebert, and C. Lauradoux, “Preserving privacy in secured ZigBee wireless sensor networks,” in 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), 2015, pp. 715–720.

[9] Zigbee Alliance, “ZigBee Specification, Document 053474r20,” 2012, Retrieved February 19, 2016 from http://www.zigbee.org/download/standards-zigbee-specification/

[10] Intel Corporation, “Intel Galileo Gen 2 Development Board,” 2014. Retrieved February 19, 2016 from http://www.intel.com/content/www/us/en/embedded/products/galileo/galileo-g2-datasheet.html

[11] Node.js Foundation, “Node.js v5.6.0 Documentation.” Retrieved February 19, 2016 from https://nodejs.org/api/

[12] Yocto Project and Linux Foundation, “Yocto Project.” Retrieved February 19, 2016 from https://www.yoctoproject.org/

[13] M. Banzi and Shiloh, “Getting Started with Arduino,” 3rd ed., vol. 11. Maker Media, Inc, 2014.

[14] MaxStream, “XBee TM Series 2 OEM RF Modules,”. 2007, pp. 1–60.

[15] M. Kooijman, Building Wireless Sensor Networks Using Arduino. Packt Publishing Limited, 2015.

[16] Freescale Semiconductor, “Miniature I2C Digital Barometer MPL115A2,” 2013. Retrieved February 19, 2016 from http://cache.freescale.com/files/sensors/doc/data_sheet/MPL115A2.pdf.

[17] L. Curvan, “Data Sheet, NSL-19M51, TO-18 Open Plastic Encapsulated,” 2008. Retrieved February 19, 2016 from http://docs-europe.electrocomponents.com/webdocs/002e/0900766b8002e0d5.pdf.

[18] Aosong Electronics Co Ltd, “Digital-output relative humidity & temperature sensor/module DHT22/AM2302,” 2015. Retrieved February 19, 2016 from https://www.sparkfun.com/datasheets/Sensors/Temperature/DHT22.pdf.

[19] J. Soini, P. Sillberg and P. Rantanen, “Prototype System for Improving Manually Collected Data Quality,” 2014 Proceedings of the 3rd Workshop on Software Quality Analysis, Monitoring, Improvement, and Applications, SQAMIA 2014, September 19-22, 2014, pp. 99-106.

Portable Sensor System for Reliable Condition Measurement

J. Soini *, P. Sillberg *, P. Rantanen * and J. Nummela **
* Tampere University of Technology/Pori Department, Pori, Finland
** Riffid Oy, Rauma, Finland
jari.o.soini@tut.fi

Abstract – Regarding sustainable development, there is a growing need to gather more and more various kinds of measurement, space, and consumption information about property. The necessity for property condition measurement is apparent and the appropriate circumstances, such as indoor air quality and suitable temperature, have an essential influence on comfort and welfare at work and, at the same time, have significance in terms of energy efficiency. This paper presents a portable prototype system for property condition measurement. The objective was to generate a reliable system that improves the quality and also the visual presentation of the collected data. The paper presents the components of the system and the technology utilized to implement the system. The results of piloting in a real-life environment, where particular focus was placed on both controlling energy efficiency and well-being at work, are also presented.

I. INTRODUCTION

Nowadays there is a growing interest in the indoor air condition of buildings, the operability of structure-technical equipment and systems, as well as in energy efficiency. Likewise, concerning sustainable development, there is a growing need to receive more diverse measuring, status, and consumption information on properties. Indoor air quality and temperature have an essential influence on working and living conditions. With condition measurement, fulfillment of condition objectives can be better ensured and possible anomalies can be detected early on. Companies in charge of the construction and maintenance of properties (and an increasing amount of private households and housing companies as well) profit from condition measurement in their activity. Property condition measurement offers the following benefits:

- energy conservation
- real-time consumption and status control
- improvement of working conditions
- malfunction management
- reliable history information of property behavior
- verification of the functionality of property automation
- alert notification.

In future, the need to receive more diverse measuring and status/freeze frame data on properties will grow. Measurement produces data that can be used, for example, to monitor consumption and circumstances for predictive maintenance purposes, or for the purposes of property automation. It is crucial to ensure that measurement systems work continuously and flawlessly, or if problems arise, the problem areas can be easily identified so that the necessary adjustments and possible corrective actions can be carried out. This paper presents one solution for this purpose - a portable prototype system for reliable condition measurement. The system developed facilitates the maintenance staff's work in detecting possible anomalous condition situations in properties.

Related research in this area has been conducted, for example, by Leong et al. [1] who have studied a Near Field Communication (NFC) and Bluetooth bridge system for connecting Bluetooth-enabled mobile devices to NFC-enabled consumer services. Ihasalo [2] has presented a tool for the construction and evaluation of continuous building performance measurement. Banerjee et al. [3] concentrate on the security aspects of sensors that collect sensitive or private data. Saari et al. [4] studied on collecting data from a sensor packet which was connected by wire directly to a single board computer. This was made by using embedded Linux and BeagleBone Black hardware, which is similar credit card–sized single-board computer as Raspberry Pi. The focus of that research was to collect data and to serve it over the Internet. Toshniwal et al. [5] have introduced a network-based sensor monitoring system that is portable for various applications. Voinescu et al. [6] describe in their research a USB device, which can be used as a wireless network gateway between ZigBee sensor nodes. These are some examples of studies related to this theme.

The structure of the paper is as follows: Section 2 represents the basis and the starting point of the system development work as well as briefly introduces the sensor technology utilized. Section 3 describes the architecture of our prototype system, data collection process, and the user interface for the graphical representation of collected data. Finally, section 4 summarizes the paper.

II. BACKGROUND – CASE: PORI HEALTH CARE CENTER

This case study was a part of the KiiauData (Smart analysis of property systems data, 2013-2015) project

funded by TEKES [7], where one of the main aims was to study potential new technologies for managing and controlling circumstances in buildings in a smart way. This case study is one example of the technology pilots carried out during the research project. This technology pilot - a portable sensor system - was developed as a collaboration project between Tampere University Technology (TUT) and the City of Pori. For this research case, the 5-storey hospital building of the City of Pori main health center was selected, which had undergone an extensive renovation, completed in the fall of 2014. The owners, i.e., the City of Pori, were particularly interested in identifying whether the property automation updated at the same time was functioning properly.

The research objective set was to implement a sensor system that would enable the flexible collection of the most reliable measurement data possible on changes in the conditions of the premises (rooms) and present the results in the clearest and most illustrative form possible. The research began by mapping out all the potential technologies for identifying changes in room conditions. In this case, the research was restricted to monitoring temperature and humidity data. An additional starting point for the system was that it should be *transferable* from one point to another, and *scalable* (addition of sensors) at a later date. Also, there was plenty of discussion on an appropriate measurement *results presentation* tool and methods from the aspect of analyzing and interpreting the collected data.

At the beginning of the case study, the sensor technology to be utilized in the system was selected. As a result of the specification and design stage, it was decided to build the system using RFID/NFC sensors. Considering the technical aspect of living conditions monitoring sensor networks, the application can be divided into three parts: 1) measuring the data, 2) transferring the data from the sensors to the service and 3) processing the data in the service. For part 2), one emerging approach for transferring data from sensors is NFC [8] or RFID (Radio Frequency Identification) technology [9], which was also utilized in this study.

Using RFID or NFC to gather data from sensors enables users to utilize a simple existing method for wireless data transfer, and thus develop scalable and cost efficient applications and solutions. NFC complements several widely used consumer level wireless technologies, by utilizing the key elements in existing standards for contactless card technology (ISO/IEC 14443 A&B and JIS-X 6319-4) and also enabling devices to share information at a distance of less than 4 centimeters with a maximum communication speed of 424 kbps. [10]. Choosing NFC as a communication method also reduces the energy requirement on the sensor side. All the energy needed for data communication will be received from the NFC reader, so the sensor battery capacity can be better utilized for data collection. This makes more frequent sampling possible while still maintaining long operation periods and lower maintenance requirements.

III. SENSOR SYSTEM IMPLEMENTATION

This chapter describes the implementation of the sensor system. First, the architecture of the system is described and then the selected sensor type and the data collection process. Finally, we introduce the user interface for the graphical representation of the collected data.

A. System Overview

Fig. 1 shows an overview of the system. The *Service* is available over the Internet where both the *Web User Interface* and *Data Collection* application can be connected. The service uses JavaScript Object Notation (JSON) to transmit data objects and has two REpresentational State Transfer (REST) interfaces, one for receiving the gauge data and the other for posting the data. The web user interface is a JavaScript-based web page accessible with a web browser. Here, the NFC tag collects the data independently, and maintenance staff is able to periodically transfer the collected data from the tag.



Figure 1. System overview.

Using the web interface, the collected data can be viewed embedded on a location map, plotted on a chart, or downloaded as raw data (as a comma-separated value file, CSV) for use with, for example, a spreadsheet application such as Microsoft Excel. A separate web page exists for system administrators for inserting new tag locations and uploading new data collected from the tags. In a practical use case, the tag locations are very rarely changed, and the web interface is generally used only for observing the tag data. Uploading new data is a relatively simple operation, in which the administrator uses the web browser to upload the log file transferred from the tag.

B. Sensors and Data Collection

The study began with a survey of the sensor types suitable for the purpose. As a result, 14 sensor types were found that could be adapted for the planned application. The key selection criterion was the customization capability of the sensors (e.g. code), as this feature ruled out most of the sensors found. The NFC-capable sensor KT-255F [11], which can store up to 16000 data points in its non-volatile memory, was chosen for this project. The device is shown on the left of Fig. 2 (below). It would also have been possible to buy the sensors separately and use a microcontroller or a single-board computer (for example, BeagleBone Black [12] or Raspberry Pi [13]) to implement an open source sensor device. In our case, temperature and humidity measurements were required and thus, a simple commercial device was a simpler and

easier solution. A proprietary device brings certain disadvantages when compared to a more open platform, such as dependency on the manufacturer's software and support and possible difficulties on moving from one manufacturer to another in the future, or making modifications to the existing hardware. The KT-255F sensor uses the FeliCa standard [14] in combination with a proprietary encrypted data format, which makes it difficult to read the tag using other than manufacturer certified reader software. Also, the encryption is not signed with a unique key, so the data can be read as long as you have the manufacturer's software and a FeliCa capable reader device, making encryption somewhat irrelevant. The data cannot be modified using an external reader, but the sensor can be reset to the default values. In practice, this means that any user with a suitable reader can read the data or reset the tag, but the user cannot insert malicious data. In our use case, this is not a major concern, as the tags are located in a public place (a hospital). The collected data, even though it is not published, is public, and in theory, you could gather the same data by simply walking around the building with a temperature meter in your hand. In addition, the data is used only for illustrating an overview of the conditions inside public areas, and accessing critical areas of the building requires privileged access, making tampering with the tags very difficult.

In our use case, however, the advantages outweigh any possible disadvantages. The tag device is in many areas situated in a publicly visible location, making a finalized product a more appropriate solution. The chosen commercial device is of sturdier build and is not as easily damaged if accidentally hit or dropped as a self-made sensor system would be. As an added benefit, the tag contains an LCD screen, which shows the current temperature, which according to the patients and the hospital staff was a nice feature, although not the most crucial part of the system.

Reading the tags is performed using a laptop with an external NFC tag reader and reader software. The software is capable of generating a PDF document with histograms and data values. The PDF format is unnecessarily difficult to convert to a format transferable to a database, and all the same graphs and charts can be generated by any other software, such as the web interface described in this paper. For that reason, it is much easier to collect the raw data only, which contains a comma-separated list of the values and their timestamps, and an ID unique to each tag, which can be used to detect automatically which tag was read. The data is automatically parsed by the service, and timestamps are compared to ensure the coherence of the data, and more specifically, to remove any duplicates in case the log files contain previously uploaded results. After the upload, the values are also checked for validity against a preset range of acceptable values (e.g., the temperature should be between 21 and 24 degrees Celsius). If a value is outside the range, an alert notification is automatically generated and shown in the web interface.

To summarize, for the sake of simplicity, the software is used to read the data log file, which is then uploaded by the system administrator or service personnel to the service using the administration web page. After the upload, the data will be immediately visible on the web interface. No other action is required, and the operation takes one to five minutes per tag, depending on the amount of data. The battery life of a tag is about two years, but the memory of the tag can contain only a limited amount of measurements. In our case, the tag sensors measure the temperature and humidity every 15 minutes, which fills the available memory in about five months. In practice, the measurements would be read more often than every five months, preferably once per month or even every two weeks for relatively up-to-date data. It should be noted that in this case, the data is used only for statistical and analytical purposes, i.e., verification of the property automation functionality, and there is no requirement for real-time data collection.

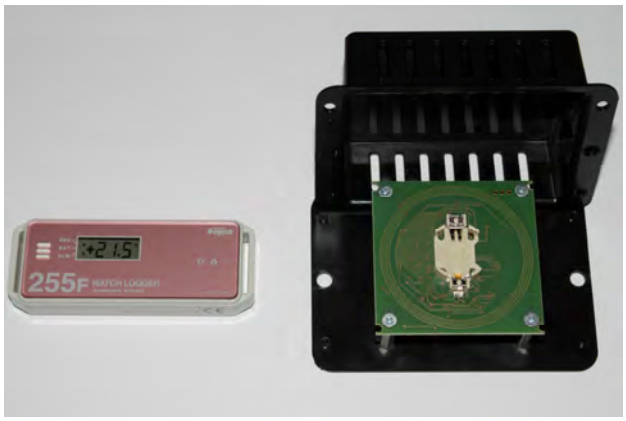


Figure 2. KT-255F Sensor and GoSense TH-Stat ID sensor.

In addition to the chosen sensor model KT-255F, other types were also sourced. Fig. 2 (right) presents another suitable option. However, this sensor, type name: GoSense TH-Stat ID NFC Temp & Humidity Sensor [15], varies clearly from the selected model, since it is not a ready-made commercial model nor is there any software available for configuring, controlling, and data handling. In some situations this can also be a benefit, since all data interfaces are open and specifications are freely available, which makes it possible to use this sensor in a wider area of applications. This sensor uses the standard ISO15693 air-interface for communication, which gives more flexibility in choosing the reader device compared to FeliCa, and can store up to 1 Mb measurement data. When considering this pilot study and its proof-of-concept type goals, using this sensor would have required considerably more effort in software development, which favored the use of the chosen KT-255F sensor.

When considering this application (living environment monitoring), it can naturally be carried out in different ways. One typical solution would be installing the necessary monitoring sensors permanently to the space under monitoring, and transfer measurement data through wires to the monitoring server. However, this approach requires more installation and cabling work than wireless options. The traditional way to utilize wireless data collection is to use radio links to transfer data, but then the battery life of the sensors may become an issue. The selected approach, using battery-powered data loggers and standardized NFC technology to deliver data from the sensors, makes it possible for sensors to gather a large

amount of data with high sample frequency and still keep the battery lifetime around two years, which is more than enough for this application. Along with this chosen architecture, the entire measuring system can be considered portable, as it can easily be taken to other locations for subsequent measuring periods.

C. *Web User Interface and Visualization*

The user interface is intended for the graphical representation of collected data. In addition to this, a meter location view was created to enable a quick overview of the collected data. This view is shown in Fig. 3. The web application supports internationalization (i18n) features but the data itself is not translated; it includes the names of the locations and the meters. The meter location view consists of the following sections:

- Top panel, which lists all the locations available for the authenticated user.
- Main view displaying locations and meters.
- Meter statistics, which can be seen when the cursor is moved on top of a meter marker. Fig. 4 illustrates the meter statistics window.
- Bottom toolbar with buttons *Show meters* and *Clear selections*.

The meter location view can show any kind of image and draw meters on top of it. The testing phase of the system consisted of installing sensors in the newly renovated facilities of the main health center of the City of Pori, mapping the sensor locations to floor plans, and finally defining acceptable boundaries for room temperature and relative humidity. In the event that the recorded data is outside the limits, an exclamation point will appear inside the corresponding meter marker. This tool can help the pinpointing of issues that occur in a certain room or floor.

Fig. 3 and Fig. 4 show floor plans of the main health center. The health center underwent extensive renovation work, and it was decided that collection of room temperature and humidity data was needed in order to verify a comfortable indoor environment. There are a total of 32 active sensors in this building, eight (8) for each floor, with a sampling period of 15 minutes.

Fig. 4 shows a close-up screenshot of a floor plan with an alert window opened. This alert appears in room number 24 on the third floor of the building (*3. Kerros Huone 24*). The window reports that both the humidity (8 % relative humidity) and the temperature have been too low (20.2 degrees Celsius), compared to the set limits. Other information displayed in the window is the average and standard deviation of all measured data types.

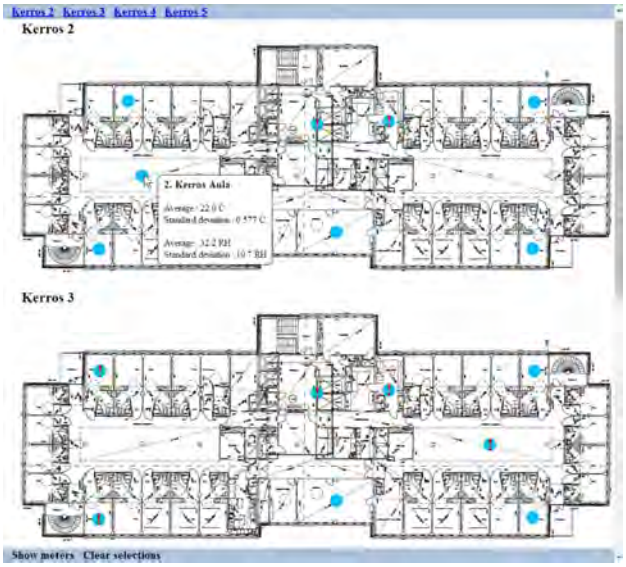


Figure 3. Meter locations overview.

The collected data can also be displayed in a chart view. The user may choose to show data from only one meter by double clicking the desired meter marker, or multiple meters can be shown by clicking any amount of meter markers, and then clicking *Show meters* (shown in Fig. 3). Fig. 5 (below) is a screenshot of a chart data view from the second floor lobby (*2. Kerros Aula*).

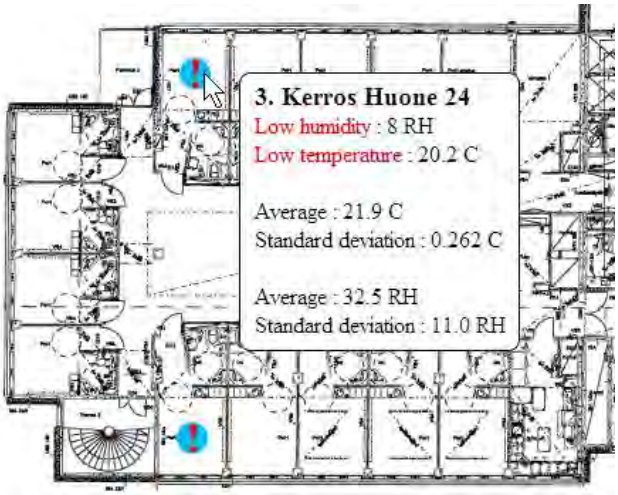


Figure 4. Close-up of floor plan with an alert window.

The upper chart zooms into the latest two weeks by default. The chart has options for zooming and panning for navigating through the data. The smaller chart on the bottom shows the overview of all collected data points. The chart can also be used for navigating through the data by selecting the desired range of dates.

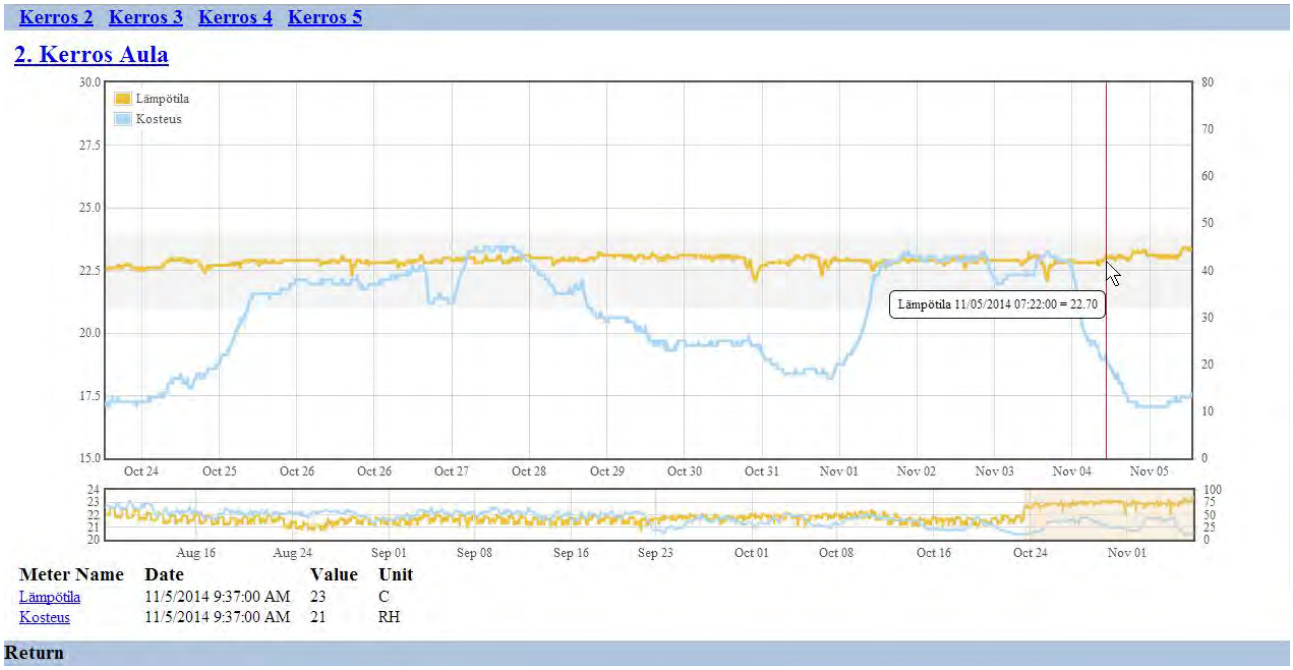


Figure 5. Meter value chart.

The example chart (Fig. 5) shows temperature (*Lämpötila*, higher line) in yellow and relative humidity (*Kosteus*, lower line) in blue. The chart can also draw the preferred zone for the temperature; here it has been defined as between 21 and 24 degrees Celsius. The cursor can be used to track the exact sensor value at a specific time and date shown in the tooltip. The *Return* button on the bottom toolbar can be used to go back to the list of locations view.

This prototype system has been in operation at the City of Pori main health care center since the fall of 2014. The primary goal of the system is to monitor the indoor conditions of the property and any changes that may occur over a long period of time. The health care center is under a large renovation project with an adjacent hospital building currently under construction. When the new construction project is complete, the sensor system will also be installed in the new location. The system can assist in detecting changes in the living and working conditions in both buildings during the renovation period, and after the projects are complete.

IV. Summary and Future Research

The paper presented a flexible and reliable prototype application for verifying changes in conditions inside a property and presents the measurement data collected to the system user in a clear manner. The portable condition measurement system makes it possible to reliably monitor and check the functionality of the property automation system at the renovated pilot target (main health center). The system introduced assists maintenance staff and also supports managers who are responsible for ensuring the correct operation of the devices. This system is one step toward more reliable measurement data, and it also improves the visual presentation of collected condition data for analysis. The portable, RFID-based, measurement sensor business is only just taking off globally. Conventional measurement sensors are still dominant, but in the near future both small- and low-current sensors will have a clear place on the market. In the future, the focus will be on studying how this technology can be applied to the packet transfer business. This technology does not need any wiring; sensors are easy to move and are easy to locate in terms of size and weight in different kinds of transfer packets. This is our next topic related to the sensor network research theme.

References

[1] C.Y. Leong, K.C. Ong, K.K. Tan, O.P. Gan, "Near Field Communication and Bluetooth Bridge System for Mobile Commerce", IEEE International Conference on Mobile Commerce, Industrial Informatics, pp. 50-55. August 16-18, 2006.

[2] H. Ihasalo, "Transforming building automation data into building performance metrics - design, implementation and evaluation of use of a performance monitoring and management system", Aalto University publication series Doctoral Dissertations, 26/2012, ISBN 978-952-60-4540-5 (electronic).

[3] S. Banerjee, D. Sethia, T. Mittal, U. Arora, and A. Chauhan, "Secure sensor node with Raspberry Pi," Impact-2013, pp. 26–30, November 2013.

[4] M. Saari, P. Sillberg, P. Rantanen, J. Soini, and H. Fukai, "Data collector service - practical approach with embedded linux". In 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1037–1041, 2015.

[5] K. Toshniwal, and J. M. Conrad, "A web-based sensor monitoring system on a Linux-based single board computer platform," Proceedings of the IEEE SoutheastCon 2010, pp. 371–374, March 2010.

[6] A. Voinescu, D. Tudose, and D. Dragomir, "A lightweight, versatile gateway platform for wireless sensor networks," In Networking in Education and Research, RoEduNet International Conference 12th Edition, pp. 1–4, September 2013.

[7] Finnish Funding Agency for Technology and Innovation, Tekes, http://www.tekes.fi/en. (retrieved November 6, 2015)

[8] International Organization for Standardization. ISO/IEC 18092:2013, Information technology -- Telecommunications and information exchange between systems -- Near Field Communication -- Interface and Protocol (NFCIP-1).

http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=56692 (retrieved November 6, 2015)

[9] Finkenzeller Klaus. RFID Handbook: Radio-frequency identification fundamentals and applications, Wiley, 1999.

[10] NFC Forum, What is NFC?, http://nfc-forum.org/what-is-nfc/ (retrieved November 6, 2015)

[11] Fujita Electric Works Ltd, KT-255F Specifications (Japanese), http://f-log.jp/?p=117 (retrieved November 6, 2015)

[12] G. Coley, BeagleBone Black System Reference Manual. 2014.

[13] V. Vujović, and M. Maksimović, "Raspberry Pi as a wireless sensor node: performances and constraints," 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1247–1252, 2014.

[14] Sony Corporation, FeliCa Card User's Manual, http://www.sony.net/Products/felica/business/tech-support/data/card_usersmanual_2.0.pdf (retrieved November 6, 2015)

[15] GoSense Wireless Ltd, 20TH-StatID Datasheet, http://www.gosense-wireless.com/GoSense%20TH-StatID%20DS%20Rev1.pdf (retrieved November 6, 2015)

Architecture of an Interoperable IoT Platform Based on Microservices

Tomislav Vresk* and Igor Čavrak**
*Končar elektronika i informatika, Zagreb, Croatia
**University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
tvresk@koncar-inem.hr, igor.cavrak@fer.hr

Abstract - The vision of the Internet of Things enabled the development of a wide spectrum of services, applications and ecosystems, deemed infeasible not long ago. However, lack of standardization poses a number of questions still requiring proper addressing. Due to the need of supporting large number of users and significant data processing throughput, Internet of Things requires a specific approach towards the problem of providing sufficient scalability and performance, clearly pointing towards the distribution of effort among a large number of small and specialized services. Reflecting on the importance of coexistence of heterogeneous systems supporting the Internet of Things, this paper presents an overview of the specific problems inherent to the IoT and the proposal of an architecture of a microservice based middleware aimed at connecting heterogeneous IoT devices. The middleware functionality is achieved irrespective of the size and complexity of a given device network, both from the data model aspect and from the aspect of connecting existing and newly created middleware components.

I. INTRODUCTION

The emerging communication technology breakthrough, based on the concept of each device having the ability to communicate with its surrounding environment while being able to take leverage of data gathered by other nearby connected devices, has grown into a phenomenon called the Internet of Things (IoT). Studies conducted in the business sector indicate that currently 20% of American households have already integrated some kind of IoT platform, while 68% of them intend to do it within the next 10 years. Independent market research and advisory firms forecast 6.4 billion connected things will be in use worldwide in 2016, up 30% from 2015, and will reach 20.8 billion by 2020 [1].

Although the concept of a global connectivity represents the foundation of the IoT that jump started the fourth industrial revolution, due to complexity inherited trough ability to integrate infinite number of technologies and multi-purpose communication protocols it also reveals the problems which lie in the nature of the IoT itself.

The paradigm raises questions on not only the security and privacy aspects but calls into question the laws governing the administration of such a network [2]. For the IoT to reach its full potential, a single internationally accepted communications standard is needed [3]. Even though standardization will have a significant impact on the paradigm evolution, the possibility of enforcing *de jure* and/or *de facto* standards by the predominant companies should not be ignored; monopoly on the interoperability technology may have rights associated with it, owned by one or several market players. In addition, such rights may be exploited to generate monopoly rents, counteracting the competition in standard-based interoperable products and services.

Properly implemented standard guarantees interoperability, which represents a crucial starting point for the successful integration and information sharing between distributed systems. Energy efficiency and information security are just some of the industry segments where standardization represents a mandatory obligation. In order to overcome this critical barrier, it is important to enable global ubiquitous connectivity through usage of publicly available communication standards, so that the data generated by devices can be shared and cross pollinated to create new and useful cross-functionality. At the same time, there are other segments and areas that do not require any way of standardization. Therefore, it is perfectly legitimate to ask whether is it really possible to standardize IoT, and more importantly, is standardization the only prerequisite standing in the way of enabling the global interoperability.

To address the scalability and interoperability problems, we define an IoT concept and platform based on orchestration of different IoT system components (devices, data sources, data processors, storage etc.) implemented in the form of microservices. The microservice based approach aims to hide the complexity stemming from variation of end-device properties by employing a uniform approach towards modelling of both physical and logical IoT devices and services, including uniform device addressing schemas. The approach also fosters interoperability and extensibility by encapsulating diverse communication protocols into proxy microservice components thus allowing for easy replacement of existing and adding new system components, laying the groundwork for open and dynamic IoT ecosystems.

Section II of this paper addresses the inherent problem of scalability in IoT systems and the microservice-based approach to tackling it. Section III describes our proposal of IoT-specific microservice architecture and organization. Section IV describes the example implementation of an

IoT system based on the microservice approach, while Section V concludes the paper.

II. MICROSERVICES

The rapid development of technology enabled ubiquitous connectivity resulting in a continuous growth of a number of connected devices, but it also revealed many challenges related to development and architecture of current information systems. While the network traffic generated by smart phones depends solely on the number of active users, IoT solicits a different approach where network data traffic results from communication among a vast number of both permanently and intermittently connected devices. Due to cohesion of wide variety of solutions and interacting devices, it is almost impossible to accurately and effectively predict data traffic patterns. As a result, such erratic behaviour can lead to exponential growth of network traffic and unpredictable system load.

In order to handle the growing amount of work, such as demand spikes, it is essential to implement scalability mechanisms and strategies that will maintain the system stability. There are two basic strategies for scaling: vertical and horizontal [4]. In case of vertical scaling, additional resources are added to a single processing node. As a result, the node can handle more work and provide additional capacities; however, vertical scaling is obviously limited by factors such as cost effectiveness, physical constraints and availability of specialized hardware. In contrast, horizontal scaling adds more nodes to the overall system. Opposite to vertical scaling, horizontal scaling represents a set of design practices and planned activities leading to an inherent distribution within the system. If the system supports horizontal scaling, it can be theoretically enlarged to thousands of machines. This unique ability defines the horizontal scaling approach as the key building stone of a large-scale architecture design.

The microservices approach, a relatively new term in the software architecture patterns [5], could present one of the solutions to the scalability problem in the IoT domain. Microservice pattern advocates organizing distributed applications as a suite of services, where each of the services is running in its own independent process context designed to segregate one service from another. Internal cross-service communication is implemented trough proprietary interfaces, their implementations predominantly relying on lightweight TCP/UDP-based application protocols.

At the core of the microservices approach lays the possibility to independently deploy and scale each service; a service may be deployed as several instances and different services may be hosted on the same server. When building and deploying distributed systems, the focus on automated provisioning is crucial. Such tools and services, essential for establishing the service provisioning and management chain, can be built using any programming language or development framework.

Each microservice is responsible for one or more closely related functions. There are a couple of different ways of decomposing the application into services. One approach is to use verb-based decomposition and define services that implement a single use case. The other option is to decompose the application by noun and create services responsible for all operations related to a particular entity such as customer management. An application might use a combination of verb-based and noun-based decomposition.

Opposite to the microservice architecture, the modern monolithic architecture provides a natural way of development, while combining all logic into a single process divided and organized into classes, methods and packages [6]. Such way allows the use of a deployment pipeline to ensure changes are properly tested and deployed into the processing environment. At the same time, a small change in the application code requires the application/service to be completely rebuilt and redeployed.

Taking into consideration that IoT relies on the integration of a large number of technologies, it is reasonable to conclude that a monolithic based approach does not represent a suitable solution. Advances in technology have driven a large amount of new architectures and design patterns. As such, the feature-based framework microservice approach allows the reuse of components leading towards the highly scalable service-oriented ecosystem.

III. IOT MICROSERVICE COMPONENTS

Currently, the main focus of the ongoing IoT-related research is on resolving the connectivity challenge; however, there is something that connects them all: telemetry. Even though telemetry represents one of the crucial functionalities of each connected device, there still remain more communication patterns, such as commands, queries and notifications, and accompanying asynchronous and synchronous data processing, that are often completely excluded from taking into consideration.

Synchronous or transactional operations in an inherently asynchronous IoT environment tend to introduce additional layers of complexity by employing distributed transaction management mechanisms to tie various actions together. As such, implementation of all communication towards other distributed components trough a single interoperable interface represents one of the biggest challenges in the considered architectural design.

From a high level perspective, the microservice component model can be interpreted as a black box decomposed into logical groupings - layers (Fig. 1). Layers help to differentiate between different kinds of tasks performed by components, making it easier to create a design that supports reusability of components. The basic function set includes cross-service API responsible for service data context exchange, coordination and remote service management.

End user authentication is implemented using token-based authentication. It allows consumers such as end users or services to provide credentials in order to obtain a token which allows fetching a specific resource. Once the

token has been obtained, the owner is granted access to a specific resource for a specific time period.

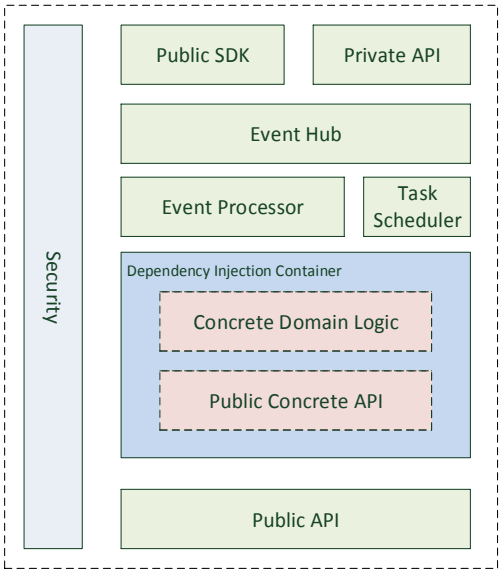


Figure 1. Microservice high level overview

Interactions among different microservices are restricted to data exchange using only microservice-exposed interfaces. By utilizing distributed data flows, services can make use of computing resources across the network. Note that such robust data flow management is required to handle the velocity, variety, and volume of exchanged data. The implementation of process logic is facilitated by inheriting functionality from pre-existing microservice core components, with the addition of communication mechanism implementations by using framework-provided API-s. By taking such an approach, a concrete microservice instance can represent a data management REST interface [7], CoAP end point [8] or simply a permanent storage data historian.

Rapidly changing technology represents one of the biggest challenges in achieving ubiquitous connectivity; every day, new technology is developed while old technologies being obsoleted or upgraded. Currently, there is a broad set of protocols being promoted as the silver bullet of IoT communication for the higher-level M2M protocol in the protocol stack [9], bringing the standard IP-centric protocols into the realm of smart devices and smart objects. One of the biggest challenges that IoT ecosystems face are the presence of low-powered devices; such devices need to function for prolonged periods of time in unfriendly environments, restrained by limited power resources and supporting frequent data exchanges over lossy communication links. These unique challenges make the as-is use of the existing TCP/IP-based communication protocols less than ideal.

Several factors pose significant impact on device battery life; the capacity being the main one. However, in a battery life estimation equation, capacity is a constant, while consumption appears to be only other relevant factor. Similar to battery life estimation, there are many factors that determine power consumption; frequency, payload size and communication protocol are just the essential ones. One of the key energy/power considerations in wireless communications for low-powered devices is the duty-cycle of the wireless radio transceiver. Communication protocols that allow the transceiver to remain inactive for longer periods of time are far more energy efficient than those requiring the radio to be almost constantly powered on. By taking all variables into consideration, it is reasonable to conclude that such unique scenario requires support for lightweight communication mechanism. As such, RESTful HTTP over TCP is particularly attractive for connecting consumer premise devices, given the near universal availability of HTTP stacks for various platforms.

A. Data Model

Data are at the heart of every information system. Without connected devices being able to capture, transfer, analyse, report and act on data, the benefits of the IoT would not be achievable. As as result, we are challenged with questions: who owns this data and who should have access to it? Issues related to interoperability, automation, and data analytics naturally lead to semantic-oriented perspective [10]. Applying semantic technology promotes interoperability among resources, information models, data providers and consumers, and facilitates data access, integration and knowledge extraction.

Prior to describing our proposed data model, key concepts such as *virtual device*, *thing* and *resource* will be introduced (Fig. 2). A *thing* can be any object, even abstract data context in the real world. *Things* are represented as *logical devices* having a unique ID, name, type, attributes and a list of *logical nodes*. A *logical device* contains one or more *logical nodes*, which represent various components of the physical device. A *resource* is a computational element providing access to a particular *logical device element*. In order to represent a physical device as a logical device, decoupling into corresponding virtual components in most cases will not provide enough relevant data. Attributes are part of the information model used for resource description, and can include additional context data like resource location, sensor accuracy etc.

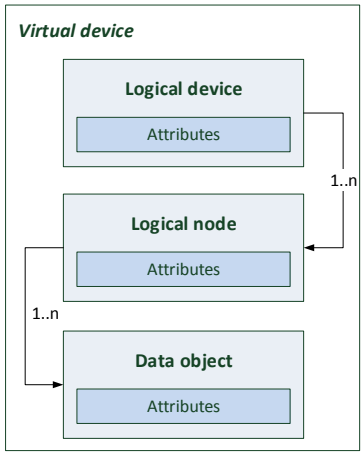


Figure 2. Logical device data model hierarchy

As an example, the modelled IoT device could contain a thermal energy smart meter logical device extended with surrounding environment temperature and a day ahead thermal energy usage estimation. The logical device would consist of two logical nodes: *measurement* and *estimation*. *Measurement* logical node would contain surrounding environment properties such as environment temperature data object and thermal energy smart meter data set. *Estimation* logical node would contain a day ahead thermal energy usage estimation data object representing a data output of analysis performed by concrete microservice via data inputs gathered from other microservice data outputs.

The biggest challenge arising from digitalization of physical assets is preserving the sufficient level of detail. Data can be defined as a representation of facts, concepts or instructions in a formalized manner suitable for communication, interpretation, or processing by human or electronic machine. When data are processed, interpreted, organized, structured or presented so as to make them meaningful or useful, they are called information. Such information provides context for data and therefore can be represented with a set of abstract primitive data types that are used to define data object in a virtual conceptual model.

B. Address Model

The primary objective of the address space is to define a method for data providers to represent resources to potential users. The object data model has been designed to meet this objective, with resources being addressable using the uniform addressing scheme. Address space can be directly mapped to REST interface exposing resources in a RESTful service as unique URL-s. However, logical device data elements are subject to structural changes, thus it is much more difficult to implement human readable addressing scheme compared to other more conventional approaches. Depending on the usage context, we are proposing two alternative addressing schemes for accessing the desired resource; *unified human readable scheme* and *unified global scheme*.

The *unified human readable scheme* aims to organize resources using the meaningful human readable URLs. As previously defined, each logical device represents a virtual model of some physical object and as such can be differentiated to less complex components. Because of the tree-like properties of the model, this path resembles a fully-qualified file-name notation. The path lists all the objects on the route from the root of the model to the object in question. Furthermore, fully-qualified file-name notation usually has a fixed delimiter between object names, and unified human readable notation uses slash to separate the logical device from the rest of the path. As an example of the human readable scheme, the setup contains a weather station model composed of multiple logical nodes containing the latest temperature information:

uhm/ws-1-location/ measurement -latest-temperature

uhm: address model

ws-1-location: logical device

measurement: logical node

latest: logical node containing leaf data objects

temperature: data object

To satisfy the REST constraints, each URL should map to a single resource and all access to this resource should be provided using that specific URL. However, opposite to other addressing concepts, this notation uses paths that resemble a fully-qualified file-name notation, meaning that they can easily be mapped to a URL.

In the *unified global scheme*, each resource is automatically assigned a globally unique resource identifier. With reference to a given set of objects, the unique identifier is any identifier which is guaranteed to be unique among all identifiers used for those objects and for a specific purpose within the distributed system. Although the unified global scheme is much easier to implement and, unlike human readable scheme, is in full compliance with RESTful nature, in the end it represents a trade off between good usage practices and human readable semantics.

As an example of the unified global addressing scheme, the setup contains a weather station model composed of multiple logical nodes containing the latest temperature information:

ugm/data-object/5428d753-b033-4c4e-98f9-d9c1b4a1f8ff

uum: address model

data-object: resource type

IV. IMPLEMENTATION SCENARIO

The typical cloud computing architecture, presented on Fig. 3, introduces architecture that meets all the relevant IoT criteria. The purpose of this example scenario is to create cohesion between devices using standard communication protocol and third party web service which involves establishment of microservice data flow chain whose outputs act as inputs for predictive analytics processing, resulting with the future thermal energy consumption estimation.

The scenario includes a number of smart meters installed on a chosen location. Since the energy efficiency segment is subject to standard, the status of individual meters is read via an open communication protocol. Most of the energy meters are installed in a constant power supply capable environment, but due to specific environmental properties (i.e. prohibitive costs) they tend to have a limited IP-capable connectivity. Even though this problem has been partially minimized with the

advancement of technology trough GPRS/3G and upcoming LTE-M/NB-IOT [11], Wireless M-Bus based solutions [12] are still having primary role in meter reading system architectures. Such Wireless M-bus based energy meters represent non IP capable devices and require additional IP-capability integration. In order to provide the IP-capability, Wireless M-Bus to Modbus TCP converter is integrated into existing network.

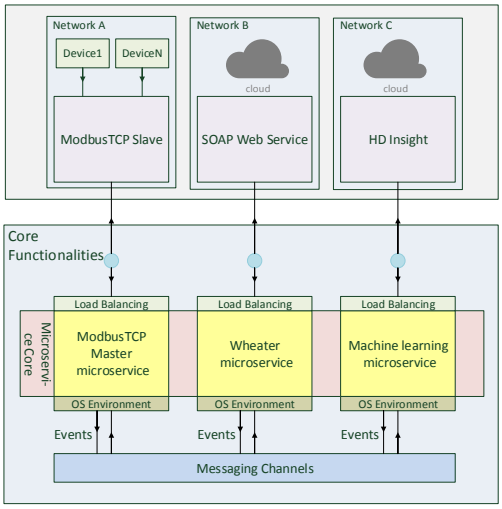


Figure 3. High level platform overview

The architecture consists of several layers. The first layer contains the load balancing nodes. The following layer contains microservice applications responsible for transfer, processing and transformation of data trough a dataflow pipeline. The pipeline encapsulates an entire series of computations that accepts input data from both internal and external sources, transforms that data to provide useful intelligence, and produces output data. Opposite to the traditional generic pipeline approach, microservice based dataflow pipeline supports bi-directional communication through usage of integrated mechanisms. For the purposes of current architecture, a dataflow pipeline is formed by integrating three microservices: Modbus TCP master, meteorological (weather) service and machine learning predictive analytics.

Even though current architecture is based on broad research and predefined functional requirements, some real world problems were discovered only during the implementation phase. Such problems could be classified into two main groups: problems originated due to lack of chosen development framework knowledge and problems caused due to communication protocol limitations.

A. Modbus TCP master microservice

Modbus is considered an application layer messaging protocol, providing master/slave communication between devices connected through buses or networks [13]. Since Modbus protocol is just a messaging structure, it is independent of the underlying physical layer, and is therefore suitable for TCP encapsulation [14]. As such, Modbus TCP uses the client/server architectural style.

Opposite to modern event-driven publish/subscribe communication protocols, Modbus TCP is based on passive dataflow mechanisms requiring specific time interval polling. Designed and built around base microservice core, the developed Modbus TCP microservice provides essential Modbus TCP protocol stack support for data acquisition of Wireless M-Bus over Modbus TCP gateway connected devices. Originally published in the 80s, simple and robust, during the past years it has become de facto standard communication protocol. However, the address field in the Modbus-defined frame is only 8 bits wide. As such, Modbus is restricted to addressing 247 devices on one data link, limiting the number of field devices that may be connected to a master station (Fig. 4). To get beyond this limit, a protocol modification could be made to use two-byte address format, requiring both master and slaves to support it. However, such an approach would not be in compliance with the standard and would not represent a viable solution.

Relying on the key concepts of this paper, a solution to this problem could be found by adding multiple instances of the same microservice with unique IP or port parameters, distributing the global data context over multiple nodes without the need for complex protocol stack modifications. Defined by the standard, each Modbus capable device bases its data model on a set of tables with distinguishing characteristics.

Besides the implementation of Modbus protocol stack, Modbus TCP microservice includes additional key functionalities including task based slave context polling, in-service data historization and cross-service data exchange.

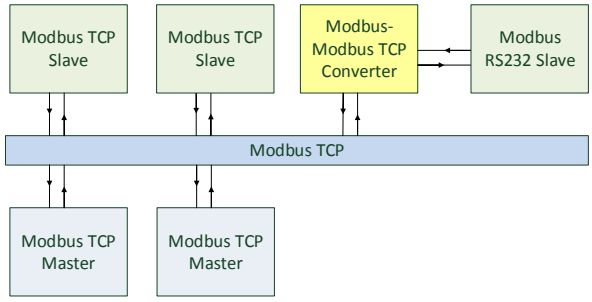


Figure 4. Modbus TCP network architecture

B. Meteorological microservice

In residential IoT systems (home automation scenarios) environmental temperature represents one of the key points in the energy consumption estimation equation. By examining the present energy consumption estimation model, it is clear that thermal energy consumption in relation with time does not represent a complete set of parameters required for performing the future estimation analysis successfully.

The main role of meteorological microservice is microclimate data acquisition, in-service data historization and cross-service exchange. Opposite to Modbus TCP microservice, which sets focus on implementation of Modbus communication protocol stack, meteorological microservice encapsulates non-standard web service in order to implement microclimate data acquisition for a preconfigured geographical location.

C. Machine learning microservice

Machine learning is quietly taking over in the mainstream and in a short period of time it has become a vital segment of every intelligent application [15]. The main role of machine learning predictive analytics microservice is data acquisition, model processing and model output stream warehousing.

Currently there is no standard methodology by which to carry out data mining projects. Due to lack of standardization each manufacturer uses its own in-house designed predictive model development methodology. However, our main focus is set on integration of Microsoft HD Insight (Hadoop) [16] framework based on CRISP-DM methodology [17]. The CRISP-DM process model aims to make large data mining projects less costly, more reliable, repeatable, manageable and faster. The process consists of the following four steps.

Defining goals - defining the right problems and goals represents the first step in predictive analytics. This step starts by asking what do we want to achieve and what is the final result.

Obtaining the data - after the analytical environment has been set up and the goals defined, the next step includes importing the right dataset upon which data model will be created. Even though data can be imported from many sources in this particular scenario processing dataset includes data gathered from both Modbus TCP master and meteorological microservices. The resulting data set contains current consumption, ambient temperature, corresponding data point time stamps and assigned geographical locations.

Data pre-processing - a dataset usually requires a certain amount of pre-processing before it can be analysed. Pre-processing available data involves adjusting the available data using the standard API which allows input row data manipulation and transformation using available methods such as merging datasets, grouping and summarizing data, checking for missing values and replacing them with appropriate values and converting values to another type.

Selecting and applying the learning algorithm - predictive analytics is the process of building models from historical or current data in order to forecast future outcomes. Predictive modelling is a process used in predictive analytics to create a statistical model of future behaviour.

A predictive model is made up of a number of predictors, variable factors that are likely to influence future behaviour or results. Data is collected for the relevant predictors, a statistical model is formulated, predictions are made and the model is validated (or revised) as additional data becomes available.

In order to estimate the future energy consumption simplest linear regression model has been chosen for our example system [18]. Regression analysis is a data mining process that identifies a model that depicts the correlation between an outcome and one or more predictors. The result is a trained regression model that can be used to score new samples to make predictions.

V. CONCLUSION

The vision of the 4th industrial revolution, triggered by the innovative technological concepts and solutions, emphasizes the need for connecting heterogeneous devices within a single global environment, but simultaneously raises a number of questions still remaining unanswered. In order to overcome a number of challenges, such as fragmentation barrier of vertically oriented closed systems, it is crucial that the attention is pointed towards the open platform architecture. Even though there is no universal answer to standardization problem, it should be noted that scalability of applications makes prerequisite for the efficient use of Internet of Things resources.

In this paper we presented an overview of specific IoT related problems and introduced the proposal of an architecture of a microservice based middleware whose aim is to ensure cohesion between different types of devices, services and communication protocols while maintaining the system scalability. In addition to provided concepts, we described a regression model estimation processing dataflow pipeline based on different type data source integration using the proposed architecture.

REFERENCES

[1] Gartner, “Forecast: The Internet of Things”, Gartner, 2013.

[2] N. Dhanjani, “Abusing the Internet of Things”, O'Reilly Media, 2015.

[3] G. Santucci, “The Internet of Things: Between the Revolution of the Internet and the Metamorphosis of Objects”, Vision and Challenges for Realising the Internet of Things, 2010.

[4] M. Abbott and M. Fisher, “The art of scalability: Scalable web architecture, processes, and organizations for the modern enterprise”, Addison-Wesley Professional, 2009.

[5] Microservices http://microservices.io/patterns/microservices.html, (20.4.2015.)

[6] M. Richards, “Software Architecture Patterns”, O'Reilly Media, 2015.

[7] A. Rodriguez, “RESTful Web services: The basics.” https://www.ibm.com/developerworks/webservices/library/ws-restful, (20.4.2015.)

[8] IETF The Constrained Application Protocol (CoAP), http://tools.ietf.org/html/rfc7252

[9] A. McEwen and H. Cassimally, “Designing the Internet of Things”, John Wiley & Sons, 2014.

[10] European Research Cluster, “IoT Semantic Interoperability: Research Challenges, Best Practices, Recommendations and Next Steps”, 2015.

[11] Nokia Network, LTE-M – “Optimizing LTE for the Internet of Things”, 2015.

[12] Wireless M-Bus, “Communication systems for meters and remote reading of meters - Part 4: Wireless meter readout (Radio meter reading for operation in SRD bands)”, 2013.

[13] Modicon, “Modicon Modbus Protocol Reference Guide”, Modicon Inc. Industrial Automation Systems, 1996.

[14] Modbus.org, “MODBUS Messaging on TCP/IP Implementation Guide V1.0b”, 2006.

[15] A. Smola and S.Vishwanathan, “Introduction to Machine Learning”, University of Cambridge, 2010.

[16] T. White, “Hadoop: The Definitive Guide”, O'Reilly Media, 2009.

[17] P. Chapman, T. Reinartz, C.Sheare and R Wirth, “CRISP-DM 1.0: Step-by-step data mining guide”, CRISP-DM consortium, 2000.

[18] L. Schmidt-Thieme, “Machine Learning: Linear Regression", University of Hildesheim, 2007.

Performance Estimation in Heterogeneous MPSoC Based on Elementary Operation Cost

Nikolina Frid, Danko Ivošević and Vlado Sruk
University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
nikolina.frid@fer.hr, danko.ivosevic@fer.hr, vlado.sruk@fer.hr

Abstract – Effective use of resources available on heterogeneous MPSoC platforms can only be achieved through careful resource allocation and scheduling. The diversity of processing and memory elements will manifest itself in the total time and resources required to perform a task or execute an application. Choosing the right platform element is the key and the first step is performance estimation.

This paper tackles the issue of finding the most suitable processing element for each part of the software application through a novel approach – elementary operation cost. The cost of each elementary operation is experimentally determined through a set of carefully devised benchmarks and is used for estimating duration of complex functions found in common applications such as JPEG, AES etc. By raising the abstraction level on which the execution time is calculated from instruction to operation level, common problems in performance estimation such as pipelining and branch prediction can be avoided and estimation accuracy is improved. Demonstrated results show that the average error rate of estimated execution times for various benchmarks is around six percent compared to the actual execution times.

Keywords: Embedded Computer Systems, Heterogeneous Computing, FPGA, Design Space Exploration, Performance Estimation

I. INTRODUCTION

Heterogeneous multiprocessor systems have continuously been used in embedded systems for almost a decade due to their ability to provide better performance at lower cost, area and power demand. Much work has been done to achieve, in each new generation of systems, even better performance at lower cost and design time while maintaining software flexibility. Exploring large design space poses the greatest challenge in MPSoC design. Choosing the right platform architecture which will meet software application demands in early stages of design process is the key to reducing development cost and time-to-market. The widely accepted approach [1] to overcome this challenge is to start the design space exploration at high levels of abstraction allowing for fast evaluation of available options. In early design stages the hardware platform is still not finished and performance must be evaluated either through simulation or analytical methods. This article proposes a novel analytical method for software performance estimation on heterogeneous MPSoC platforms based on elementary operation cost concept. This method ensures high accuracy of results while maintaining high level of abstraction and thus enabling fast design space exploration.

II. RELATED WORK

In early days of heterogeneous platforms research and design, Instruction Set and Cycle-Accurate Simulators (ISS/CAS) have been invaluable tools for early performance estimation [2]. However, their high accuracy came at the cost of long simulation time and large design effort. More recent work has adopted the TLM [3] approach which allows much faster functional and performance (i.e. timing) evaluation with possibility to even simulate the behaviour of RTOS. While TLM gives very good results in functional simulation, the issues arise in achieving high accuracy of timing estimation. Main challenge is to accurately describe software application performance on different processing elements and provide it as an input to TLM. Several different approaches exist, mostly relying on some form of application source code annotation and transformation into SystemC, which is suitable for later use in TLM.

Authors in [5] rely on source level annotation. They all break down the application structure into Basic Blocks (mostly using GCC compiler) and use ISS or CAS to estimate performance for each block. In later stages the application is simulated (speed comparable to real-time execution of application) with different inputs, mostly as a part of a transaction-level model. Rate error remains under ten percent, in most work around six percent, within acceptable boundaries.

A slightly different approach is presented in [4]. The authors introduce annotation at intermediate representation (IR) level, a lower level of abstraction, in order to capture compiler optimizations more accurately.

However, all these approaches have trouble with the inter-block dependency (pipeline effects). The problem is partially solved by introducing pair-wise execution of neighbour blocks or simulation of as many possible basic block combinations at the cost of total simulation time.

A rather different approach is presented in [9], where authors use analytical estimation instead of simulation.

This method employs neural networks which receive the number of each type of instruction of software application and give the estimation of number of cycles (i.e. total duration) as output. The results are little less accurate with around 17 percent average error rate, and the network training time is a bit longer, but the system is more flexible and faster in later stages.

III. Elementary Operation Concept

Fast and efficient exploration of the large design space, present in MPSoC design, requires early performance estimation, as mentioned earlier. Consequently, the performance estimation must be conducted at very high level of abstraction using application and platform models. It is also very important that this models not only enable fast execution but also retain a very high degree of accuracy in performance estimation. Thus it is extremely important to accurately describe the heterogeneous platform.

In earlier work regarding the impact of heterogeneity on MPSoC performance [10] it has been observed that the performance of a particular type of processing element in execution of simple tasks (e.g. calculation of square root of an integer number) highly correlates with the performance of that same element when executing a more complex task (e.g. JPEG compression). In this paper that concept is further elaborated.

A. Concept Overview

Great majority of authors agree that the key to source-level performance estimation is to identify a unit of code small enough to enable modular and reusable approach and in the same time large enough to diminish the pipeline effects. Assembly level instructions satisfy only the first condition. Basic blocks, at C source level, are much more resilient to pipeline effects but their reusability is often questionable – two different functions (tasks) almost never have two identical basic blocks. Thus ISS must be used each time a new function requires timing estimation. Moreover, pipeline effects (i.e. branch prediction) between two basic blocks must be separately handled.

On the other hand, it can be observed that there exist types of operations (not necessarily a single line of code) which are present in many different functions written in C. Although it is hard to list all possible types of operations which can be used in writing C functions it can be safely concluded that it is a finite set. And while the research discussed in this paper has certainly not identified all of them, through careful examination of different types of functions it is possible to do so.

At this point it can be concluded that there exist several distinct subsets of these elementary operations that accurately reflect the implicit features of a processing element architecture: integer and floating point arithmetic operations, logic and memory operations. All of these operation types are also supported by processor through dedicated parts of datapath. Further examination and experiments reveal that distinction must also be made between operations on local and global variables, and

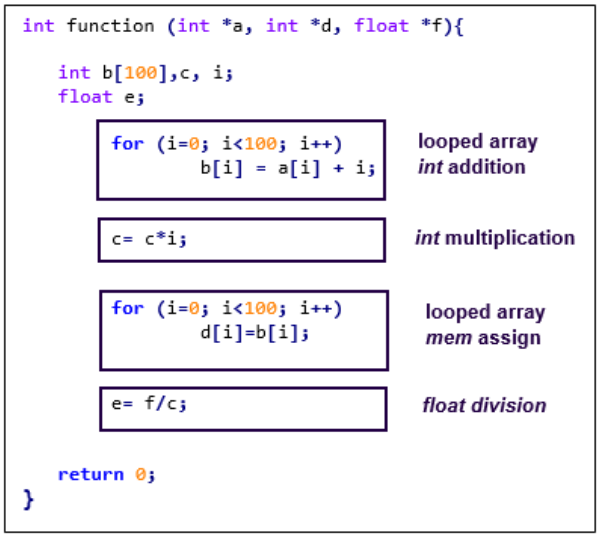


Figure 1. Elementary operation examples

operations with arrays. Looped execution of a certain type of operation should also be taken into consideration. An example of several types of elementary operations is illustrated in Figure1.

Once a large enough set of elementary operations is defined, the cost is determined based on execution time of these operations for each type of processing element by using a simulator or the actual target. The data collected is reusable, without any additional measurements, for estimating performance of any software application task at hand. The granularity of elementary operation is large enough to encompass pipeline effects. Also, an elementary operation often crosses basic block boundaries thus cancelling inter-block effects.

Test benches for measuring elementary operation execution times have been carefully devised to include a large set of possible operation types. CHStone [10] had been used as a starting point, mostly to identify the distinction between integer and floating point arithmetic operations, and between simple (e.g. addition, shift) and more complex operations (e.g. multiplication, division). However, major modifications and improvements have been done to better suit the earlier described features of a heterogeneous architecture.

B. Elementary operation cost measurement

Experiments and measurements have been conducted on two different RISC processor architectures implemented on Xilinx ZedBoard Zynq®-7000 All Programmable SoC.

An ARM Cortex A9 processor was used in the following configuration: operating frequency @ 667MHZ, 32 KB L1 cache and 512 KB L2 cache with both instructions and data stored in local DDR3 SDRAM memory operating at 533 MHz.

MicroBlaze, a softcore processor was used in the following architecture configuration: 5-stage pipeline

TABLE I. ELEMENTARY OPERATION EXECUTION TIMES

Target processor		ARM Cortex A9 [ms]	MicroBlaze [ms]
INT ADD	loop	1,2E-5	1,15E-4
	unrolled loop	9,23E-6	2,93E-5
INT MUL	loop	1,352E-5	2,9E-4
	unrolled loop	6,9E-6	1,02E-4
FLOAT ADD	loop	1,35E-5	1,635E-3
	unrolled loop	1,18E-5	1,48E-3
FLOAT MUL	loop	1,5E-5	1,74E-3
	unrolled loop	1,414E-5	9,83E-4

with hardware multiplier, FPU and barrel shifter @ 200 MHz (200 MHz clock frequency is the maximum achievable frequency on PL side of Zynq system) with both the instructions and data stored in local BRAM memory.

In Table 1 an example of elementary operation execution times for different types of addition and multiplication operation is presented. The distinction is made between integer and floating point operations and whether the operation is performed in a loop or not.

Besides using the obtained data for estimating duration of a complex task execution on a certain type of processor, this data can also be used for quick identification of implicit features of processor architecture and datapath. For example, Figure 2 illustrates the impact of loop unrolling for operations specified in Table 1. It is clearly visible that due to the internal architecture, it is much more beneficial to do loop unrolling on a MicroBlaze processor than on ARM Cortex A9 – speedup of integer addition goes up to four times, and floating point multiplication almost doubles in performance.

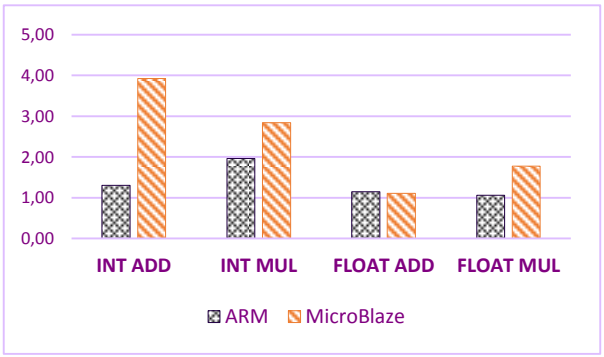


Figure 2. Loop unrolling speedup comparison

IV. EXPERIMENTAL RESULTS

Several different benchmark sets have been used to test the ability to accurately estimate task execution time based on elementary operation cost concept: MiBench [12], JPEG [13,14] and AES [15].

The total of seven benchmarks were used: two benchmarks from MiBench set (*Cubic* and *Sqrt*), three from JPEG set (*Shift, DCT* and *Zig-Zag*) and two from AES set (*SubBytes* and *ShiftRows*). These particular benchmarks have been chosen to test all elementary operation subsets. *Cubic and Sqrt* benchmarks test integer and floating point arithmetic operations. *Sqrt, Shift, DCT* and *Zig-Zag* integer, memory and logic operations, while AES benchmarks test memory operations.

The number of each type of elementary operation included in these benchmarks needed for timing estimation, has been analysed with the help of a previously developed tool [16]. This tool has been further improved to enable automation of the process of identifying total number of occurrence of each type of elementary operation in a given function.

After the analysis of elementary operations contained within each benchmark, timing estimation was calculated based on previously obtained operation execution times. The results are given in Table 2 under *Est.* These results were compared with the results of real-time execution of

TABLE II. COMPARISON OF ESTIMATED AND ACTUAL EXECUTION TIMES

Benchmark			*Cubic*	*Sqrt*	*JPEG Shift*	*JPEG DCT*	*JPEG Zig-Zag*	*AES SubBytes*	*AES ShiftRows*
Target	ARM Cortex A9	*Est. [ms]*	4,99E-03	1,77E-03	2,24E-03	1,34E-02	4,03E-03	1,33E-03	5,39E-04
		Act. [ms]	4,76E-03	1,69E-03	2,27E-03	1,54E-02	3,89E-03	1,44E-03	5,16E-04
		Error	*-4,83%*	*-4,73%*	*1,32%*	*12,99%*	*-3,60%*	*7,64%*	*-4,46%*
	MicroBlaze	*Est. [ms]*	2,77E-01	9,28E-03	1,35E-02	9,06E-02	1,42E-02	5,44E-02	5,64E-04
		Act. [ms]	2,56E-01	8,74E-03	1,47E-02	9,79E-02	1,38E-02	5,30E-02	5,07E-04
		Error	*-8,20%*	*-6,18%*	*8,16%*	*7,46%*	*-2,90%*	*-2,64%*	*-11,24%*

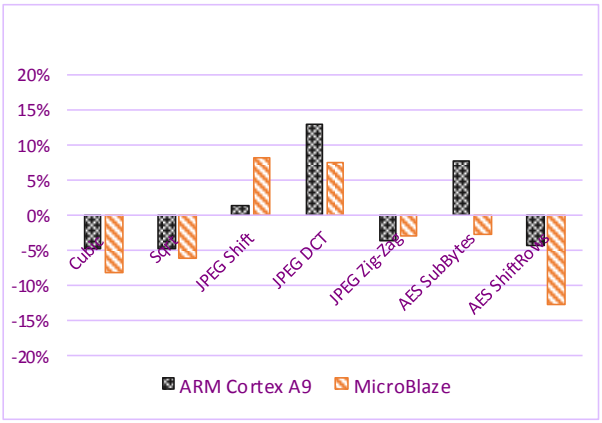


Figure 3. Estimation error rates

the same set of benchmarks on the Zynq platform for both target processors. Benchmark execution times on the Zynq platform are indicated in the Table 2 under *Act.*

Comparison of the results, illustrated in Figure 3 shows that the average error rate is around six percent with the peak at twelve percent. Considering that timing estimation is done at very high level of abstraction, these results are within acceptable range and also comparable to the results of methods presented in related work mentioned previously.

It is important to note that while these experiments were conducted on a real physical platform, they could have also been conducted, without any modification on an ISS or cycle-accurate simulator. In this case the simulator for these processors was not available.

V. Conclusion

In this article a novel approach to source-level performance estimation has been proposed. The elementary operation cost concept enables very fast design space exploration with high accuracy of results. The average error rate at six percent is comparable to the results of methods presented in related work, and thus within acceptable limit, especially the performance estimation is done in a very early design stage where the main goal to detect the most suitable type of processing element for a certain type of software application task and determine the optimal total number of processing elements.

First steps in future evolution of this approach will be full automation of the entire process and providing support for evaluation of effects of different memory configurations. Further improvements will need to face the challenge of handling compiler-introduced optimizations. The possibility to use this approach for simulation-based software performance evaluation, possibly in a transaction-level environment, will also be considered.

References

[1] K. Keutzer, A. R. Newton, J. M. Rabaey, and A. Sangiovanni-Vincentelli, "System-level design: orthogonalization of concerns and platform-based design," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 19, no. 12, pp. 1523–1543, Dec. 2000.

[2] L. Cai and D. Gajski, "Transaction level modeling," in *Proceedings of the 1st IEEE/ACM/IFIP international conference on Hardware/software codesign & system synthesis - CODES+ISSS '03*, 2003, p. 19.

[3] S. Abdi, G. Schirner, Y. Hwang, D. D. Gajski, and L. Yu, "Automatic TLM generation for early validation of multicore sSystems," *IEEE Des. Test Comput.*, vol. 28, no. 3, pp. 10–19, May 2011.

[4] S. Chakravarty, Z. Zhao, and A. Gerstlauer, "Automated, retargetable back-annotation for host compiled performance and power modeling," in *2013 International Conference on Hardware/Software Codesign and System Synthesis, CODES+ISSS 2013*, 2013, pp. 1–10.

[5] K. L. Lin, C. K. Lo, and R. S. Tsay, "Source-level timing annotation for fast and accurate TLM computation model generation," in *Proceedings of the Asia and South Pacific Design Automation Conference, ASP-DAC*, 2010, pp. 235–240.

[6] Z. Wang and J. Henkel, "Accurate source-level simulation of embedded software with respect to compiler optimizations," *Des. Autom. & Test Eur. Conf. & Exhib.*, vol. 0, pp. 382–387, 2012.

[7] E. Cheung, H. Hsieh, and F. Balarin, "Fast and accurate performance simulation of embedded software for MPSoC," in *Proceedings of the Asia and South Pacific Design Automation Conference, ASP-DAC*, 2009, pp. 552–557.

[8] P. Gerin, M. M. Hamayun, and F. Pétrot, "Native MPSoC co-simulation environment for software performance estimation," in *Proceedings of the 7th IEEE/ACM international conference on Hardware/software codesign and system synthesis - CODES+ISSS '09*, 2009, p. 403.

[9] M. Oyamada, F. R. Wagner, M. Bonaciu, W. Cesario, and A. Jerraya, "Software Performance Estimation in MPSoC Design," in *2007 Asia and South Pacific Design Automation Conference*, 2007, pp. 38–43.

[10] N. Frid, D. Ivosevic, and V. Sruk, "Heterogeneity impact on MPSoC platforms performance," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2015, pp. 1071–1076.

[11] Yuko Hara, Hiroyuki Tomiyama, Shinya Honda and Hiroaki Takada, "Proposal and Quantitative Analysis of the CHStone Benchmark Program Suite for Practical C-based High-level Synthesis", *Journal of Information Processing*, vol. 17, pp.242-254, 2009.

[12] M. R. Guthaus, J.S. Ringenberg, D. Ernst, et al., "MiBench: A free, commercially representative embedded benchmark suite", in *Proceedings of IEEE 4th Annual Workshop on Workload Characterization"*, Austin, TX, 2001., pp. 3-14.

[13] G. K. Wallace, "The JPEG Still Picture Compression Standard" *Communication of the ACM*, vol 34., Issue 4, ACM Press, New York, NY, USA, pp.30-44 , 1991.

[14] Arai Y., Agui T., Nakajima M., "A Fast DCT-sQ Scheme for Images", Trans. IEICE, Vol. E71, No.11 pp. 1095-1097, 1988.

[15] NIST, "FIPS 197 Advance Encryption Standard (AES)", http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf.

[16] D. Ivošević and V. Sruk, "Unified flow of custom processor design and FPGA implementation," in *IEEE EuroCon 2013*, 2013, pp. 1721–1727.

Sustav za lociranje atmosferskih pražnjenja u identifikaciji kvarova TK mreže uzrokovanih atmosferskim prenaponima

V. Milardić*, B. Franc*, M. Budimirović **
* Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska
** Hrvatski telekom d.d., Zagreb, Hrvatska
bojan.franc@fer.hr

Sažetak – Na korisničkoj terminalnoj opremi koja služi za pružanje tzv. 3play usluge (internet + telefonije + televizija) su česti kvarovi uslijed prenapona uzrokovani atmosferskim pražnjenjima, što ima za posljedicu nezadovoljstvo korisnika te dodatne troškove održavanja. Kvarovi se najčešće događaju u ruralnim sredinama gdje se nalaze nadzemne niskonaponske i telekomunikacijske mreže te u područjima s većim specifičnim otporom tla i većom gustoćom udara munja.

U radu je prikazan proračuna induciranih prenapona u telekomunikacijskoj mreži uslijed udara munje. Nadalje, prikazana je korelacija podataka o atmosferskim pražnjenjima s kvarovima/ispadima u TK mreži kroz povezivanje sa sustavom za upravljanje i osiguranje kvalitete širokopojasnih usluga. Prikazanom korelacijom moguće je utjecati na automatiziranje i skraćivanje vremena otklanjanja prijavljenih smetnji i kvarova, povećanje broja slučajeva koji se uspješno rješavaju prilikom prvog poziva, što u konačnici omogućava poboljšanje kvalitete pružanih usluga te povećanje korisničkog zadovoljstva i lojalnosti pružatelju.

I. UVOD

Na korisničkoj terminalnoj opremi koja služi za pružanje tzv. 3play usluge (internet + telefonije + televizija) su česti kvarovi uslijed prenapona (atmosferskih pražnjenja), što ima za posljedicu nezadovoljstvo korisnika te dodatne troškove održavanja. Kvarovi se najčešće događaju u ruralnim sredinama (nadzemna niskonaponska i telefonska mreža) te u područjima s većom gustoćom udara munje.

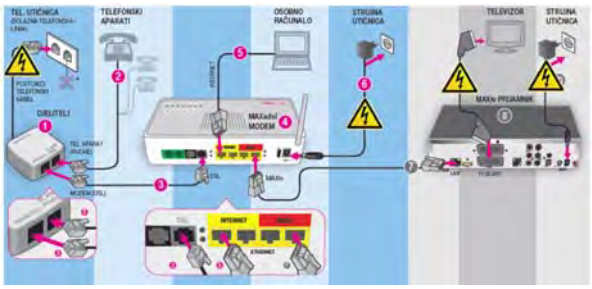
Gromobranska i prenaponska zaštita objekata izvodi se shodno zonama zaštite, a efikasno projektirana prenaponska zaštita ima više stupnjeva (II ili III) [1] .

Za efikasnu prenaponsku zaštitu korisničke terminalne opreme, a također i ostale opreme u objektu, potrebno je slijedeće:

- Objekt posjeduje temeljeni uzemljivač;
- U objektu je provedeno sveobuhvatno izjednačavanje potencijala;
- Svi vodovi koji ulaze u objekt (telefonski, niskonaponski, kabelska i satelitska televizija itd.) su adekvatno prenaponski štićeni na mjestu ulaska u objekt (I stupanj); Vodovi ulaze u objekt na istom mjestu (s iste strane objekta).
- Svi osjetljivi uređaji i uređaji koji su spojeni (priključeni) na više mreža (TV, PC, modem, MAXtv prijemnik, fax itd.) su dodatno prenaponski štićeni (II stupanj).

Zanemarivo mali broj objekata u RH je izvedena na gore opisani način, odnosno velika većina objekata nema temeljni uzemljivač, nije izvedeno izjednačavanje potencijala te ne postoji nikakva prenaponska zaštita. To naravno za posljedicu ima česte kvarove spomenutih uređaja.

Na slici 1. je prikazano uobičajeno 3play okruženje. Na slici su naznačeni najčešći putevi dolaska atmosferskih prenapona. Atmosferski prenapon može stići preko telefonske mreže, preko niskonaponske (NN) mreže te preko antenskog sustava.



Slika 1. Uobičajeno 3play okruženja

Ovdje je potrebno napomenuti da se moderni telefoni napajaju preko NN utičnice što je također mogući put dolaska prenapona. Osobno računalo ili prijenosnik se također napaja iz NN utičnice što je također mogući put dolaska prenapona. Moderni TV uređaji su spojeni na NN utičnicu, TK utičnicu i antenski sustav što su sve mogući putevi dolaska prenapona.

Iz navedenog je jasno zašto nastaju česta oštećenja modema i MAXtv prijemnika.

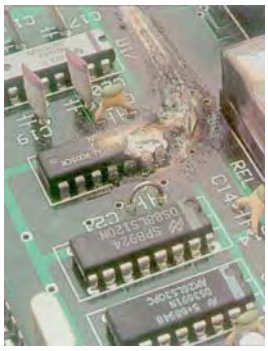
Prenaponi, prouzročeni udarom munje osim oštećenja mogu izazvati i prekid normalnog rada te degradaciju komponenti sustava, što je naravno manje opasan slučaj.

Prekid normalnog rada podrazumijeva slijedeće. I ako nije prouzročena fizička šteta na komponentama ili sustavu, logički ili analogni nivo sustava je promijenjen što uzrokuje: gubitak podataka, oštećenje podataka ili programske podrške, neobjašnjivi ispad kompjutera, modema itd. Sustav

treba resetirati (obično isključiti i ponovno uključiti) i nakon toga će ponovo normalno funkcionirati.

Degradacija komponenti je nešto opasnija (ozbiljnija). Dugotrajno izlaganje prijelaznim prenaponima niže razine, bez obzira na uzrok, degradira elektroničke komponente i krugove, reducirajući životni vijek opreme i povećavajući vjerojatnost kvara.

Prijelazni prenaponi više razine mogu prouzročiti uništenje komponenti, elektroničkih krugova i ulazno/izlaznih kartica, kao na slici 2. Vidljivo ne oštećene elektroničke komponente na slici 2. su također uništene.



Slika 2. Prijelaznim prenaponima uništeni svi mikročipovi na pločici (premda samo jedan ima jasne znakove uništenja)

Prijelaznim prenaponima uništeni ISDN uređaja na slici 3. nema tako očigledne tragove uništenja na elektroničkoj pločici. Tragovi uništenja su vidljiviji na plastičnom kućištu bijele boje.



Slika 3. ISDN uređaj uništen prijelaznim prenaponima

Sva tri ranije spomenuta efekta prouzrokuju ispad opreme ili cjelokupnog sustava što nadalje znači nezadovoljstvo korisnika te dodatne troškove održavanja i popravka.

II. INDUCIRANI PRENAPONI

Udar munje u objekt nije tako čest događaj. Isto tako, udar munje u nadzemnu telekomunikacijsku (TK) ili niskonaponsku (NN) mrežu nije tako čest. Mnogo češće, munja pogodi u blizinu objekta ili TK (NN) mreže a inducirani prenaponi preko TK i/ili NN mreže stižu u objekt i ugrožavaju električnu i elektroničku opremu unutar objekta. Inducirani prenaponi ne prouzrokuju velika energetska opterećenja uređaja prenaponske zaštite (UPZ) kao direktni udari munje ali mogu oštetiti električnu i elektroničku opremu unutar objekta, ako nije instalirana adekvatna prenaponska zaštita.

Računanje induciranih prenapona na nadzemnom vodu TK (NN) mreže, uslijed udara munje, se provodi na slijedeći način:
- prvo se usvoji model povratnog udara munje, koji specificira prostornu i vremensku raspodjelu struje munje uzduž kanala munje za vrijeme faze povratnog udara;
- zatim se računa promjena elektromagnetskog polja proizvedena strujom munje u točkama duž voda. Poboljšanje točnosti se postiže modeliranjem efekta širenja EM polja od točke udara munje.
- na kraju se izračunavaju prenaponi koristeći model sprege između EM polja i vodiča voda.

Izraz za računanje induciranog prenapona, koji se zbog svoje jednostavnosti najčešće koristi u praksi dat je u (1).

$$V_{max} = Z_0 \frac{I \cdot h}{d}\left(1+\frac{1}{\sqrt{2}}v\frac{1}{\sqrt{1-0{,}5v^2}}\right) \qquad (1)$$

gdje je: $Z_0 = \frac{1}{4\pi}\sqrt{\frac{\mu_0}{\varepsilon_0}} = 30$

I - amplituda struje munje za koju je usvojen valni oblik step funkcije;
h - visina voda iznad tla;
d - najkraća udaljenost mjesta udara i voda;
v - odnos između brzine povratnog udara i brzine svjetlosti.

Izraz (1) izveo je Rusck iz kompliciranog matematičkog modela [2]. Izraz daje maksimalnu vrijednost induciranog prenapona V_{max}, za beskonačno dugi vod, iznad savršeno vodljive zemlje (tla), u točci koja je najbliža mjestu udara munje.

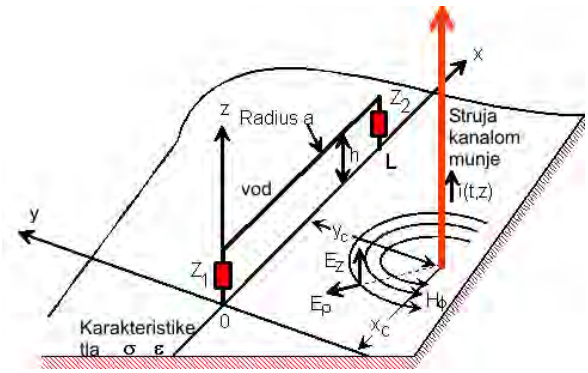
Točnije računanje induciranih prenapona je relativno kompleksno jer je i sam udar munje veoma složen fenomen i ni dan danas nije potpuno razjašnjen. U tom smislu postoje različiti modeli povratnog udara munje, različiti modeli računanja EM polja te različiti modeli sprege između kanala munje i voda. Efekte konačne vodljivosti zemlje (tla) i korone, također treba uzeti u obzir.

Ipak, određeni aspekti međudjelovanja munje i voda, mogu se modelirati jednostavnije. Premda model ne mora biti opsežan ili precizan, može se koristiti za određivanje najvažnijih parametara. Slučaj bliskog udara munje od oblaka prema zemlji prikazan je na slici 4.

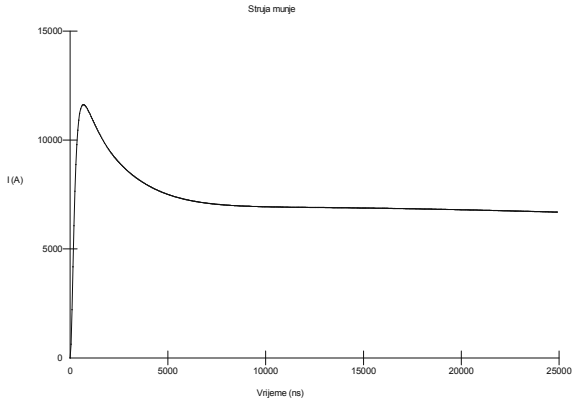
A. Proračun induciranih napona u TK mreži

Proračuni induciranog prenapona koji slijede, provedeni su kompjuterskim programom prema [3]. Za proračun induciranih prenapona u TK mreži, uslijed indirektnog udara munje, usvojena je konfiguracija kao na slici 4. Duljina TK mreže je 1,8 km, a prosječna visina vodiča TK kabela iznad tla je 6,0 m. Mjesto udara munje (kanal) je udaljen yc=100 m, (200 m, 300 m …1000 m) od TK mreže a 900 m od

početka odnosno kraja mreže. Radijus vodiča TK mreže je 0,03 cm (0,3 mm). Vodič TK mreže je na početku i kraju mreže zaključen valnom impedancijom da bi se eliminirala refleksija induciranog prenaponskog vala. Za zemlju je usvojena relativna dielektričnost 1 i specifični otpor ρ=800 Ωm. Konstanta prigušenja duž kanala munje (α) je 0,588 km^{-1}. Relativna brzina vala v u kanalu je 0,43 (=1,29*10^8 m/s). Parametri strujnog vala povratnog udara su usvojeni kao na slici 5.

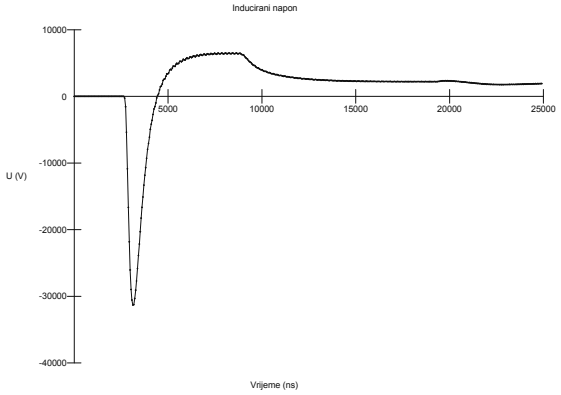


Slika 4. Vod izložen bliskom udaru munje (pražnjenju) između oblaka i zemlje [3]

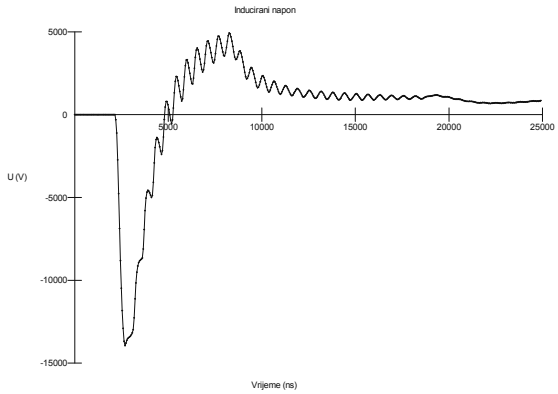


Slika 5. Struja munje za proračun induciranih prenapona

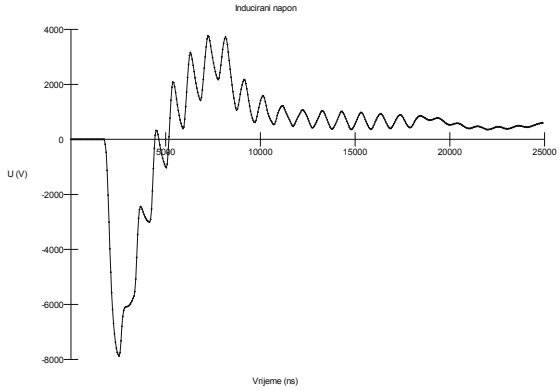
Rezultati proračuna su na slijedećim slikama.



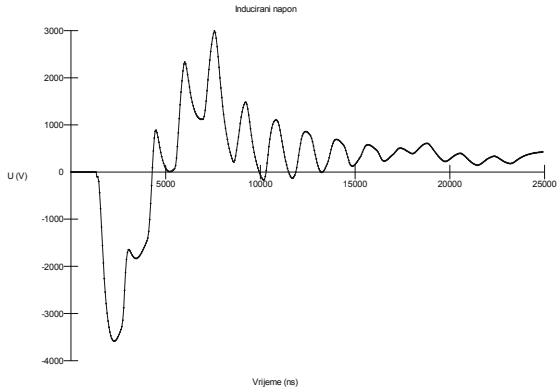
Slika 6. Inducirani prenaponi na početku (i kraju) TK mreže (otpor tereta prilagođen valnoj impedanciji, y_c= 100 m)



Slika 7. Inducirani prenaponi na početku (i kraju) TK mreže (otpor tereta prilagođen valnoj impedanciji, yc= 300 m)



Slika 8. Inducirani prenaponi na početku (i kraju) TK mreže (otpor tereta prilagođen valnoj impedanciji, yc= 500 m)



Slika 9. Inducirani prenaponi na početku (i kraju) TK mreže (otpor tereta prilagođen valnoj impedanciji, yc= 800 m)

Ako se udaljenost mjesta udara munje mijenja i od TK mreže (y_c=100 m, 200 m, 300 m …1000 m) i od početka TK mreže (x_c=100 m, 300 m, 500 m, 700 m i 900 m) dobije se tablica I. Udaljenosti od početka TK mreže su upisane vodoravno a udaljenosti od TK mreže uspravno. Za svaku udaljenost od početka TK mreže (npr. x_c=100 m) i udaljenost od TK mreže (npr. y_c=200 m) u tablici I su dvije vrijednosti izračunatog induciranog napona u kV. Prva se odnosi na

inducirani napon na početku TK mreže a druga na inducirani napon na kraju TK mreže. Iznimno, za udara munje točno na sredinu TK mreže (x_c=900 m) inducirani naponi su isti i na početku i na kraju TK mreže te je, shodno tome, u tablici I. navedena samo jedna vrijednost. Predznak „-„ u tablici I označava da je apsolutni maksimum induciranog napona negativan.

Na isti način se mogu izračunati i inducirani naponi za udare munje na udaljenosti x_c= 1100 m, 1300 m, 1500 m i 1700 m od početka TK mreže. Rezultati su slični kao u tablici I.
U tablici I su crveno označeni izračunati inducirani prenaponi apsolutno veći od 6 kV. Proračunom je potvrđena poznata činjenica da udari munje bliže TK mreži induciraju prenapone viših apsolutnih vrijednosti.

TABLICA I. INDUCIRANI NAPONI NA POČETKU I KRAJU TK MREŽE

x_c (m) / y_c (m)	100		300		500		700		900
100	21,8	-27,8	-20,6	-30,3	-26,7	-31,4	-29,8	-31,8	-31,4
200	14,6	-20,1	10,5	-21,3	-15,5	-21,8	-18,7	-21,5	-20,7
300	11,1	-14,9	8,84	-15,1	-10,2	-15,0	-12,6	-14,6	-14,0
400	8,75	-11,6	7,47	-11,7	-7,02	-11,7	-9,15	-11,3	-10,5
500	7,15	-9,67	6,21	-9,45	5,23	-9,18	-6,64	-8,70	- 7,87
600	5,94	-8,01	5,30	-7,61	4,65	-7,25	-4,89	-6,73	-5,96
700	5,00	-6,59	4,45	-6,18	4,10	-5,79	-3,71	-5,27	-4,57
800	4,23	-5,50	3,85	-5,10	3,62	-4,69	3,10	-4,18	-3,59
900	3,74	-4,65	3,43	-4,26	3,02	-3,84	2,93	-3,39	-2,84
1000	3,33	-4,00	3,00	-3,58	2,80	-3,19	2,34	-2,76	2,32

Rezultati na slikama 6. do 9. i tablici I su dobiveni računanjem induciranog prenapona za jedan vodič.

U [4] je pokazano da u sustavu s više vodiča dolazi do redukcije induciranog prenapona, u odnosu na izračunati inducirani prenapon pojedinačno za svaki vodič. Ta redukcija za sustav od više vodiča, kao što je TK kabel, iznosi ≈25 %. Također, postojanje uzemljenog metalnog plašta TK kabela ima za posljedicu daljnju redukciju induciranih prenapona. Ta redukcija ovisi o broju četvorki u kabelu i za udar munje, koji sadrži široki raspon frekvencija, može se usvojiti 0,8. Redukcija induciranih prenapona se može fizikalno objasniti djelovanjem svih vodiča protiv promjene incidentnog EM polja struje munje, koje uzrokuje indukciju. U TK mreži su instalirani i plinski odvodnici prenapona, obvezno na početku TK mreže u TK centrali te na mjestima prijelaza podzemne u nadzemnu mrežu. Plinski odvodnici prenapona ograničavaju razinu prenapona u TK mreži. Osim navedenog, tipična TK mreža duljine 1800 m je izvedena 1500 m podzemno a 300 m nadzemno. Za dio TK mreže izvedene podzemno inducirani naponi su niži. Usporedbom dobivenih rezultata u tablici I s rezultatima prikazanim u [5]-[6], za kabel ukopan u tlo na dubini 0,9 m, u tlu specifičnog otpora ρ=555-625 Ωm, može se zaključiti da su rezultati za kabel ukopan u tlo oko 34 % izračunatih.

Shodno svemu navedenom može se procijeniti da bi udari munje koji su bliže od 300 m od TK mreže mogli prouzročiti kvar uređaja koji su spojeni na TK mrežu (npr. modema ili MAXtv prijemnik).

III. SUSTAVI ZA LOCIRANJE ATMOSFERSKIH PRAŽNJENJA U OTKRIVANJU KVAROVA U TK MREŽAMA

Moderno sredstvo za praćenje atmosferskih pražnjenja danas predstavljaju sustavi za lociranje atmosferskih pražnjenja (SLAP). U razvijenim zemljama diljem svijeta se sustavi za lociranje atmosferskih pražnjenja primjenjuju u mrežama i sustavima raspoređenim na velikim prostranstvima kao što su elektroenergetski sustavi, TK mreže, mreže RTV odašiljača, mreže naftovoda i plinovoda, sustavi osiguranja, vojne instalacije, meteorološki servisi, prometna infrastruktura, sustavi zaštita od šumskih požara i drugo. Sustavi za lociranje atmosferskih pražnjenja pružaju informaciju o vremenu, lokaciji i intenzitetu udara munje u približno stvarnom vremenu, a ti se podaci mogu korelirati s podacima o ispadima i kvarovima u mreži, što može doprinijeti kvaliteti praćenja pogona mreže, otkrivanju uzroka i lokacije nastalih kvarova te minimiziranju potrebnog vremena i troškova u otklanjanju kvarova te ponovnoj uspostavi rada sustava.

Za efikasno korištenje podataka prikupljenih sustavom za lociranje atmosferskih pražnjenja u vođenju tehnoloških sustava potrebna je prikladna programska podrška. Kako se podaci SLAP-a primjenjuju u raznim tehnološkom mrežama i sustavima, programska podrška prilagođava se specifičnim potrebama pojedinog sustava. Opće gledano, funkcionalnosti SLAP sustava u vođenju i analizi tehničkih mreža su:

1. osnovne funkcionalnosti:
   - vizualizacija atmosferskih pražnjenja u približno stvarnom vremenu i alarmiranje o grmljavinskoj aktivnosti,
   - analiziranje, izrada izvještaja i vizualizacija povijesnih podataka o atmosferskim pražnjenjima,
   - statistički proračuni nad podacima o atmosferskim pražnjenjima,
   - izrada karata gustoće atmosferskih pražnjenja šireg područja;
2. napredne funkcionalnosti:
   - prostorna korelacija između podataka o atmosferskim pražnjenjima i alarmnim zonama generiranim oko promatranih objekata mreže,
   - izrada visoko-rezolucijskih karata gustoće atmosferskih pražnjenja oko promatranih objekata mreže,
   - korelacije između atmosferskih pražnjenja i smetnji i kvarova registriranih u promatranoj mreži u približno stvarnom vremenu (slika 10).



Slika 10. Primjer korelacije udara munje i kvara u mreži s određivanjem lokacije kvara na vodu

A. *Sustav za upravljanje i osiguranje kvalitete širokopojasnih usluga*

Sustav za upravljanje i osiguranje kvalitete širokopojasnih usluga (*eng. Broadband Service Assurance, BBSA*) je sustav čiji je glavni cilj potpora poslovnim procesima, otklanjanju smetnji i kvarova na T-HT širokopojasnim uslugama. Jednostavno, intuitivno i jedinstveno agentsko sučelje koje koriste agenti korisničke podrške omogućava im automatizirano vođenje procesa otklanjanja smetnji i kvarova u srhu postizanja slijedećih ciljeva:

- skraćivanje vremena otklanjanja prijavljene smetnje,
- povećanje broja slučajeva/prijava koji se uspješno rješavaju prilikom prvog poziva,
- smanjenje broja slučajeva/prijava koji se eskaliraju prema višoj razini podrške,
- paralelno izvršavanje akcija,
- kolektiranje, analiza, vizualizacija i popravak (akcija),
- OneScreen/"OneView" - status/problem prikazan pri otvaranju sesije (nakon unosa identifikatora) ,
- korisnička orijentiranost,
- jednostavnija inicijacija agenata korisničke podrške.

Sustav unutar sebe ima definirane algoritme za analizu i otklanjanje grešaka na širokopojasnim uslugama, a sve u svrhu učinkovitijeg, unificiranog, preciznog i bržeg otklanjanja smetnji na širokopojasnim uslugama. Unutar BBSA platforme se interpretiraju podaci koji se sakupljaju po elementima mreže koji sudjeluju u isporuci pojedine usluge. Zaključci i akcije pri otklonu smetnje se baziraju na realnom mjerenju s BSS/OSS (*eng. Business Support System / Operational Support System*) sustava.

B. *Vremensko-prostorna korelacija kvarova/ispada korisničke opreme i detektiranih atmosferskih pražnjenja*

U svrhu koreliranja kvarova i ispada u TK mreži s atmosferskim pražnjenjima implementirana je web usluga SLAP sustava koju BBSA poziva na zahtjev. Prilikom otvaranja slučaja/prijave kvara, BBSA prosljeđuje SLAP-u koordinate korisničke lokacije (točnije, lokaciju izvoda na koji je korisnik priključen) te vremenski interval unutar kojeg se smatra da je došlo do nastanka kvara/ispada. Na temelju ta dva ulazna podatka, SLAP prostorno-vremenski korelira događaj s atmosferskim pražnjenjima te pronalazi potencijalne kandidate za uzrok kvara/ispada. Nakon korelacije SLAP vraća BBSA broj koreliranih atmosferskih pražnjenja, podatke o atmosferskim pražnjenjima (vrijeme, lokaciju, amplitudu struje udara, grešku lociranja) te podatke o udaljenosti pojedinog udara od zadane lokacije.

Korelacijski proces između BBSA i SLAP-a izvršava se nešto više od godinu dana (početak je bio u veljači 2015. godine) te je do sada obrađeno oko 212.000 upita, pri čemu je analizirano 58.811 različitih lokacija na području Hrvatske. Unutar BBSA identificirano je pet izvoda (lokacija) u TK mreži s najvećim brojem kvarova/ispada tokom 2015. godine.

Od promatranih pet izvoda, najčešći kvarovi/ispadi tokom 2015. godine su zabilježeni na izvodu u Mošćeničkoj Dragi, gdje je 5 puta tokom godine grmljavina izazvala kvarove/ispade. U tih 5 štetnih događaja registrirano je ukupno 18 ispada/kvarova. Najviše kvarova/ispada na jednom izvodu tokom 2015. godine registrirano je na izvodu u Delnicama, gdje je uslijed dva udara munje registrirano ukupno 48 kvarova/ispada (tablica 2).

TABLICA II. KVAROVI UZROKOVANI UDARIMA MUNJA NA PROMATRANIM IZVODIMA U PROTEKLOJ GODINI

Lokacija	Štetni događaji*	Kvarovi uslijed munje
Rijeka (Delnice)	2	48
Rijeka (Delnice)	4	21
Rijeka (Mošćenička Draga)	5	18
Rijeka (Sveti Jakov)	1	17
Istra (Nova Vas)	2	16
* Štetnim događajem se smatra grmljavinska aktivnost u jednoj minuti na jednom izvodu koji je izazvao jedan ili više kvarova/ispada (npr. u istoj minuti na jednom izvodu ispadne više korisnika zbog udara munje)		

C. *Primjer toka događaja uslijed nastanka kvara na izvodu usred udara munje*

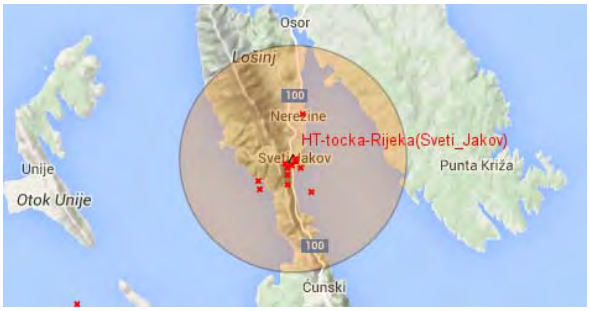
Prilikom pojave grmljavinske aktivnosti nad nekim područjem, SLAP u približno stvarnom vremenu registrira atmosferska pražnjenja (slika 11).



Slika 11. Grmljavinska aktivnost 3.10.2015. na širem području Kvarnera

U pro-aktivnom pristupu nadzora rada mreže, SLAP će postaviti upozorenje nad objektima koji se nalaze u neposrednoj blizini grmljavinskog nevremena (po završetku grmljavinske aktivnosti upozorenje će se ukinuti). U slučaju pojave kvara/ispada u komunikaciji prema krajnjem korisniku, BBSA određuje lokaciju izvoda korisnika te vrijeme prekida komunikacijskog kanala na temelju čega generira upit prema SLAP-u. SLAP provodi postupak prostorno-vremenske korelacije te vraća listu koreliranih udara munja u BBSA. U primjeru je prikazan događaj na dan 3.10.2015. kada je zabilježena grmljavinska aktivnost te kvarovi korelirani s udarima munja. Na slici 12 prikazan je jedan izvod na kojemu je BBSA zabilježio više istovremenih kvarova/ispada unutar jedne minute (jedan štetni događaj sa 17 gubitaka konekcije na istom izvodu).

Navedene kvarove (unutar jedne minute na istom izvodu) SLAP je korelirao s 12 atmosferska pražnjenja.



Slika 12. Atmosferska pražnjenja korelirana s kvarovima na izvodu

Korelirani podaci o atmosferskim pražnjenjima koje SLAP proslijeđuje BBSA su prikazani u tablici 3.

TABLICA III. KORELIRANI PODACI ZA KVAR NA IZVODU 3.10.2015. U SV. JAKOV

Broj atmosferskih pražnjenja koreliranih s kvarovima na izvodu	**12**

Vrijeme [UTC]	Lon	Lat	Struja [kA]	Greška [m]	Udaljenost munje od izvoda [m]
6:53:35,246	14,3744	44,6279	-33,0	169	1987
6:53:35,285	14,4030	44,6268	-14,1	134	1666
6:53:35,305	14,3736	44,6312	-12,5	198	1814
6:54:05,976	14,3897	44,6336	-63,2	186	750
6:54:06,006	14,3971	44,6364	-11,3	315	516
6:54:06,026	14,3884	44,6381	-13,8	158	412
6:54:06,056	14,3898	44,6297	-16,2	121	1165
6:54:06,129	14,3982	44,6577	-3,7	286	2017
6:54:06,417	14,3896	44,6365	-41,8	186	464
6:54:06,120	14,3944	44,6400	7,4	86	119
6:54:06,121	14,3945	44,6390	12,9	153	164
6:54:06,140	14,3922	44,6373	7,5	181	299

Konačno, agentu korisničke podrške je već neposredno nakon prijave smetnje/kvara dostupna informacija da li je kvar mogao biti uzrokovan udarom munje ili ne. Ta informacija može pomoći u odabiru i smanjenju vremena provođenja postupka otklanjanja kvara.

D. Beneficij korištenja podataka o atmosferskih pražnjenja u korelaciji sa smetnjama i kvarovima u telekomunikacijskim uslugama

Grmljavinska aktivnost i atmosferska pražnjenja pružateljima telekomunikacijskih usluga mogu prouzrokovati povećan broj korisničkih smetnji, vršna opterećenja pri sanaciji kvarova odnosno kraće razdoblje između prijava smetnji (MTBF). Povezivanjem smetnji, najčešće kvarova oštećenjem korisničke opreme, i događaja atmosferskih pražnjenja (vanjskih utjecaja) dobiva se informacija o najvjerojatnijem uzroku korisničke smetnje i mogućnost proaktivne reakcije, specifičnog pristupa rješavanju i u konačnici bržeg otklona smetnje. Ranim prepoznavanjem ovakvih smetnji moguće je bolje organizirati interne resurse, primjerice prevenirati korisničke pozive, optimizirati rute za intervencije zamjena oštećene korisničke opreme ili u rješavanje uključiti vanjske resurse, primjerice isporuka zamjenske opreme putem službe za distribuciju čime se može postići značajno operativno rasterećenje. Takvim pristupom moguće je troškovno optimizirati poslovanje u smislu spomenutog smanjenja operativnog opterećenja u rješavanju smetnje s kraja na kraj (call centar – tehničar), minimizirati trajanje prekida u korištenju usluge i najvažnije time povećati korisničko zadovoljstvo i lojalnost pružatelju usluge.

IV. ZAKLJUČAK

Udar munje u objekt nije tako čest događaj. Isto tako, udar munje u TK ili NN mrežu nije tako čest. Mnogo češće, munja pogodi u blizinu objekta ili TK (NN) mreže a inducirani prenaponi preko TK i/ili NN mreže stižu u objekt i ugrožavaju električnu i elektroničku opremu unutar objekta. Inducirani prenaponi ne prouzrokuju velika energetska opterećenja uređaja prenaponske zaštite kao direktni udari ali mogu oštetiti električnu i elektroničku opremu unutar objekta, ako nije instalirana adekvatna prenaponska zaštita, što je u RH skoro redovit slučaj.

Proračuni induciranih prenapona uslijed udara munje pokazuju da udari munje koji su bliže od 300 m od TK mreže mogu prouzročiti kvar uređaja koji su spojeni na TK mrežu (npr. modema ili MAXtv prijemnik).

Grmljavinska aktivnost i atmosferska pražnjenja pružateljima telekomunikacijskih usluga mogu uzrokovati povećan broj korisničkih smetnji, vršna opterećenja pri sanaciji kvarova i posljedično negativno korisničko iskustvo odnosno kraće razdoblje između prijavljenih smetnji (MTBF). Sustavi za lociranje atmosferskih pražnjenja nude dodatne informacije korisne u vođenju telekomunikacijskih sustava. Kako se podaci o atmosferskim pražnjenjima mogu primijeniti u različitim segmentima vođenja TK sustava, za realizaciju je potrebno definirati zahtjeve, modelirati funkcije te razviti adekvatnu programsku podršku. U radu je prikazana korelacija podataka o atmosferskim pražnjenjima s kvarovima/ispadima u TK mreži kroz povezivanje sa sustavom za upravljanje i osiguranje kvalitete širokopojasnih usluga (BBSA). Predstavljenom korelacijom moguće je utjecati na automatiziranje i skraćivanje vremena otklanjanja prijavljenih smetnji i kvarova, povećanje broja slučajeva koji se uspješno rješavaju prilikom prvog poziva, što u konačnici omogućava poboljšanje kvalitete pružanih usluga te povećanje korisničkog zadovoljstva i lojalnosti pružatelju usluge.

ZAHVALE

Ovaj rad su djelomično financirali Hrvatska zaklada za znanost u sklopu projekta “Development of advanced high

voltage systems by application of new information and communication technologies" (DAHVAT) i Regionalna Energetska Agencija Kvarner d.o.o.

LITERATURA

[1] I. Uglesic, V. Milardic: Surge Protection of Low-Voltage Systems Connected to Overhead Line, 28th International Conference on Lightning Protection, Kanazawa, 18-22 September 2006.

[2] C. A. Nucci: Lightning Induced Voltages, Invited Lecture 4, 26th International Conference on Lightning Protection, Cracow, 2-6 September 2002.

[3] F. M. Tesche, M. V. Ianoz, T. Karlsson: EMC Analysis Methods and Computational Models, John Wiley & Sons, NY, 1997.

[4] F. Rachidi, C. A. Nucci, M. Ianoz, C. Mazzetti: Response of Multiconductor Power Lines to Nearby Lightning Return Stroke Electromagnetic Fields, IEEE Transactions on Power Delivery, Volume 12, No. 3, July 1997.

[5] E. Petrache, F. Rachidi, M. Paolone, C. A. Nucci, V. A. Rakov, M. A. Uman: Lightning Induced Disturbances in Buried Cables - Part I: Theory, IEEE Transactions on Electromagnetic Compatibility, vol. 47, No. 3, August 2005.

[6] M. Paolone, E. Petrache, F. Rachidi, C. A. Nucci, V. A. Rakov, M. A. Uman, D. Jordan, K. Rambo, J. Jerauld, M. Nyffeler, J. Schoene: Lightning Induced Disturbances in Buried Cables - Part II: Experiment and Model Validation, IEEE Transactions on Electromagnetic Compatibility, vol. 47, No. 3, August 2005.

SNUPI – Sustav za nadzor i upravljanje procesima infrastrukture podatkovnog centra

Momir Zmijanac,*,
* Combis d.o.o. / Sektor mrežne infrastrukture i povezivanja, Zagreb, Hrvatska
Momir.Zmijanac@combis.hr

Sažetak – U uvjetima sve veće konkurencije na tržištu potrebno je osigurati maksimalnu raspoloživost sustava, povećati sigurnost i pouzdanost, te istovremeno smanjiti troškove poslovanja, odnosno troškove energije, održavanja i broj zaposlenih u operativnom poslovanju.

Navedene zahtjeve moguće je postići automatizacijom i optimizacijom performansi nadziranih infrastrukturnih elemenata pomoću sustava nadzora. Takav sustav ujedno nudi i trenutni vizualni uvid u ispravnost i dostupnost infrastrukture, obavijesti o trenutnom operativnom stanju, obavijesti o očekivanom problemu (trend), te alarm ukoliko je potrebna trenutna fizička reakcija održavatelja. Osim obavljanja nadzornih zadataka, sustav kontinuirano prikuplja podatke o ugrađenom hardveru i instaliranom softveru, te daje potpuni uvid u imovinu i znatno olakšava optimalno planiranje i korištenje resursa.

Upravljanje energijom uz maksimalnu dostupnost infrastrukture i razvoj modularnih standardiziranih modela samo su neki od razloga koji sustav za nadzor i upravljanje procesima infrastrukture (eng. data center infrastructure management, DCIM), odnosno SNUPI čine neizostavnim alatom za poslovanje sve važnijih podatkovnih centara današnjice.

Ovaj rad daje jedinstven uvid u osnove i izazove jednog takvog sustava kao integrirane nadzorne i upravljačke platforme, s primjerom i osvrtom na trenutno stanje u Hrvatskoj.

I. UVOD

Zamislimo automobil bez instrument ploče, bez brzinomjera, putomjera ili bez pokazivača razine goriva, kao i bez svih onih dodatnih pokazatelja koji olakšavaju održavanje i povećavaju sigurnost u prometu. Takav automobil bi mogli upotrijebiti samo u nuždi izvan prometnica, a u prometu danas ga nitko ne bi niti htio koristiti.

Razvojem automobila razvijali su se i sustavi koji vozaču pomažu pri donošenju odluka o sigurnosti, održavanju, o tome kada treba natočiti gorivo ili provjeriti tlak u gumama, a pojavom navigacije i o tome kojim putem se optimalno kretati. Takvi sustavi su se razvijali kao dio jedne veće, međusobno povezane nadzorno upravljačke cjeline koja ujedno upravlja i energetskom učinkovitosti automobila.

Kada govorimo o sustavima za nadzor i upravljanje unutar podatkovnih centara, kao i kod automobila, primijetit ćemo da je podatkovni centar isto tako ispunjen različitim sustavima kojima je potreban istovremeni nadzor ili upravljanje, od kojih svaki odvojeni sustav može imati svoje mjerne instrumente i signalizaciju kojima je potrebno rješenje koje ih povezuje, odnosno objedinjuje, kao i automobilu njegova instrument ploča.

Zadatak nadzorno upravljačkog sustava jednog podatkovnog centra također je održavanje ispravnog i pouzdanog rada svih povezanih sustava, uzimajući u obzir planirana održavanja, izmjene i maksimalnu raspoloživost sustava.

II. PROCESI INFRASTRUKTURE PODATKOVNOG CENTRA

Pod pojmom podatkovni centar (engl. data center) smatramo posebno izgrađen objekt ili posebno uređenu prostoriju unutar objekta koja je namijenjena za smještaj ključnih računalnih sustava; uređaja za pohranu podataka (eng. storage), uređaja za slanje i primanje podataka od mnogostrukih klijenata, odnosno poslužitelja (eng. server), te ključne komunikacijske opreme poput preklopnika (eng. switch), usmjerivača (eng. router), te sličnih uređaja.

Oprema unutar podatkovnog centra se smješta u za to posebno dizajnirane ormare (eng. rack) i to prema prethodno definiranom rasporedu kako bi se osigurao dovoljan energetsko termički kapacitet, odnosno opskrbu električnom energijom te odvođenje toplinske energije - hlađenja. Sredstvo za isporuku električne energije uređajima unutar tih ormara je „pametna“ energetska letva (eng. power distribution unit, PDU), koja nam ujedno omogućava i mjerenje potrošnje energije.

Da bi računalna oprema zadržala optimalnu radnu točku potrebno je iskorištenu energiju, koja se pretvorila u toplinu, vratiti u okolinu, odnosno potrebno je pothlađivati opremu. Preduvjet za održavanje konstantnih infrastrukturnih uvjeta su osjetila - senzori koji aktiviraju povratnu spregu, te tako vrše regulaciju. Najstarija automatika u podatkovnim centrima se temeljila na senzorima za praćenje temperature i protoka zraka. Takvi senzori su najčešće spojeni na lokalnu mrežu u podatkovnom centru i sve interakcije sa istima se obično

obavljaju pomoću standardnih LonWorks, Modbus, BACNet, SNMP ili sličnih mrežnih komunikacijskih protokola.

Osim mrežne povezanosti, samim senzorima je potreban i pouzdan izvor energije. Postavljanje većeg broja povezanih senzora zbog cijene, a i zbog samog kabliranja predstavlja osjetan logističko financijski izazov koji je ispočetka teško opravdati.

Pouzdan rad opreme zahtjeva hardversku i softversku usklađenost opreme, te optimalni prostorni smještaj uzimajući u obzir postojeći raspored energetsko-toplinskog i težinskog opterećenja. Iako se za ugradnju nove opreme prvo prazno mjesto u prvom praznom ormaru čini kao najprivlačnija lokacija, takav nesustavni pristup ne daje kvalitetno i optimalno rješenje.

Upravljanje infrastrukturom podatkovnog centra često se i dalje obavlja pomoću *ad hoc* alata i tehnologija, ali rješavanje posebnih problema sa posebnim alatima postaje sporo i nepraktično, te oslanjanje na velik broj različitih rješenja postaje neučinkovito i povećava troškove.

III. POTREBA ZA SUSTAVOM NADZORA I UPRAVLJANJA

Poput automobila bez instrument ploče, zamislimo sada i podatkovni centar u kojem nemamo uvid u ispravnost ili dostupnost infrastrukture, u kojem nemamo informacije o stvarnoj popunjenosti i iskorištenosti resursa. Kako koordinirati održavanje ili zamjenu opreme, a osigurati maksimalnu raspoloživost? Kako optimalno planirati i koristiti resurse, a istovremeno smanjiti troškove, biti energetski učinkoviti, te povećati pouzdanost i sigurnost?

Podatkovni centri, kao i ljudi koji ih vode, sve češće se susreću sa sve većim izazovima. Raspoloživost, optimizacija, učinkovitost, performanse, pouzdanost, usklađenost, računalni „oblaci“, virtualizacija i sigurnost postaju svakodnevna terminologija.

Znamo li točno koja oprema je ugrađena u podatkovnom centru? Znamo li trenutni kapacitet, što prostorni, što energetsko-toplinski? Možemo li u kratkom roku napraviti izvještaj koji pokazuje gdje se točno nalazi koja oprema, te od kojih proizvođača? Znamo li kolika je trenutna potrošnja energije? Možemo li predvidjeti troškove?

Dodajmo tome stalan pritisak smanjivanja troškova, očuvanja okoliša i vrlo ograničena sredstva, dolazimo do ogromnog i naizgled nemogućeg zadatka. Ne zadovoljimo li sva navedena očekivanja i ako nemamo odgovore na sva navedena pitanja, posljedice mogu biti vrlo štetne ili katastrofalne. Očito je da je podatkovnom centru potrebno jedinstveno rješenje koje prikuplja podatke sa svih mjernih uređaja, te na temelju podataka pruža sveobuhvatan pregled svih informacija potrebnih za učinkovito upravljanje sve složenijom infrastrukturom, potreban mu je sustav za nadzor i upravljanje procesima infrastrukture, skraćeno SNUPI.

U podacima su informacije, u informacijama je znanje, a u znanju je moć. Kada znamo naš podatkovni centar do najmanjih detalja, uz trenutni pristup traženim informacijama, prethodno navedena očekivanja i pitanja više i nisu tako zastrašujuća.

IV. ENERGETSKA UČINKOVITOST I SMANJENJE TROŠKOVA

Podatkovni centri igraju značajnu ulogu u ukupnoj potrošnji energije u sektoru informacijskih tehnologija te predstavljaju veliki potencijal za ostvarenje ušteda energije i povećanje energetske učinkovitosti (eng. power usage effectiveness, PUE).

$$PUE = \frac{Ukupna\ dostavljena\ energija}{Energija\ utrošena\ na\ računalnu\ opremu} \qquad (1)$$

Učinkovita primjena energije je jedan od glavnih ciljeva podatkovnog centra, pa tako i nadzorno upravljačkog sustava. Najpopularnija metoda (1) obračuna učinkovite primjene energije je omjer ukupno dostavljene energije i energije utrošene na računalnu opremu podatkovnog centra. Što je podatkovni centar učinkovitiji, to PUE teži prema 1. Prosječan PUE podatkovnih centara današnjice iznosi oko 1,8 [1].

Kako bi uopće upravljali potrošnjom energije, potrebno je istu mjeriti. Tu nam u pomoć uskaču već spomenute „pametne“ energetske letve preko kojih možemo lokalno i udaljeno mjeriti potrošnju opreme po ormaru, a moguća je i izmjera potrošnje po utičnici na samoj letvi. Kada takve letve povežemo sa nadzornim sustavom, dobijemo potpunu sliku potrošnje energije od strane računalne opreme unutar podatkovnog centra.

Računalnu opremu je potrebno i hladiti. Rashladna oprema troši gotovo trećinu ukupne potrošnje energije podatkovnog centra, te samim time predstavlja priliku za uštedu [2]. Rashladna oprema je također povezana s nadzorno upravljačkim sustavom.

Navedena sinergija sustava omogućava i optimalno smještanje nove opreme. Pomoću informacija koje dobijemo od sustava napajanja i rashladnog sustava možemo optimalno rasporediti opremu unutar sistem sale čime izbjegavamo „tople točke“, te rashladni sustavi mogu raditi u optimalnom režimu rada. Ukoliko nije moguće fizički premjestiti opremu, tehnologija virtualizacije u suradnji s nadzornim sustavom nam omogućava premještanje i ujednačavanje opterećenja, sve u svrhu optimalnog rada i uštede energije.

Bez nadzornog sustava odabir lokacije za novu i postojeću opremu bi se svodio na tehničke specifikacije proizvođača opreme uz informaciju o postojećem rasporedu opterećenja podatkovnog centra. Praksa je pokazala da su takve specifikacije najčešće pretjerane i većina opreme ne dostiže niti 50% predviđenih opterećenja. Posljedica ovoga je da mnogi podatkovni centri zbog takvog zastarjelog pristupa imaju planirani višak energetskih i rashladnih resursa, koji su istovremeno nedovoljni na kritičnim točkama, zbog neoptimalnog rasporeda same opreme.

Softversko rješenje nam omogućava vizualizaciju problema i rješenja, odnosno izračun optimalne lokacije za ugradnju nove opreme. Na taj način se uravnotežuje energetsko opterećenje podatkovnog centra, optimalno iskorištava prostor i ostali resursi.

Dodatna korist je i učinkovitije iskorišteno radno vrijeme djelatnika. Potpunim uvidom u infrastrukturu, koji nam pruža nadzorni sustav, uz potpuni i trenutni pristup svim potrebnim informacijama lakše postižemo učinkoviti rad.

Nadalje, vizualizacija čini sve zadatke lakšima. Većina ljudi je vizualno orijentirana. Vizualni prikaz složenih informacija je mnogo lakše razumjeti od beskonačnih proračunskih tablica, iz čega slijedi da je razumljivo i intuitivno grafičko sučelje imperativ dobrog softverskog rješenja.

V. SNUPI (SUSTAV ZA NADZOR I UPRAVLJANJE PROCESIMA INFRASTRUKTURE) KAO JEDINSTVENI I NEIZOSTAVNI ALAT

Uz cjeloviti alat na raspolaganju znatno smanjujemo složenost sustava, dok istovremeno povećavamo pouzdanost i učinkovitost. Također, smanjujemo i troškove. Takav alat bi u osnovi trebao:

- nuditi potpuni, trenutni i točan uvid u imovinu, ispravnost i dostupnost infrastrukture
- omogućiti integraciju sa postojećom (starijom) opremom i elementima infrastrukture
- pružiti točne podatke o trenutnoj potrošnji energije te omogućiti predviđanje potrošnje u određenom vremenskom intervalu
- omogućiti optimalno korištenje i planiranje resursa, odnosno dati vizualni uvid u iskorištenost prostornih i energetskih kapaciteta
- omogućiti planiranje i izradu nacrta smještaja opreme
- obavještavati o iznenadnom ili očekivanom problemu kroz model trendova, te alarmirati održavatelja ukoliko je potrebna trenutna fizička reakcija
- olakšati softversku i hardversku usklađenost opreme
- generirati izvještaje i potrebnu dokumentaciju
- nuditi jednostavno, modularno i po potrebi udaljeno upravljanje i nadzor
- imati razumljivo i intuitivno grafičko sučelje, uz naglasak na responzivno i dinamičko web rješenje
- biti otvorenog tipa i prihvaćati opremu različitih proizvođača zbog lakše zamjene opreme, otvorenog tržišta i optimizacije troškova

Ako je podatkovni centar proizvodno postrojenje koje proizvodi informacije, onda je sustav za nadzor i upravljanje rukovoditelj jednog takvog postrojenja, a informacija mu je najvrjedniji i strateški najvažniji proizvod.

Ključni atributi za poslovanje podatkovnog centra, kao i svakog drugog proizvodnog postrojenja su uspostava standardiziranih procesa i ostvarivanje predvidljivih rezultata; kako bi se uspješno i učinkovito poslovalo moraju biti uspostavljeni standardi koji moraju biti održivi. Kada proizvodni procesi postaju rutina, produktivnost raste, a količina grešaka pada. Također, ti procesi moraju dostavljati predvidljive rezultate na vrijeme, svaki put. SNUPI je kao stvoren za to jer iskorištava tu ponovljivost procesa kako bi omogućio predviđanje performansi podatkovnog centra uz prisutnost varijabilnih opterećenja i ostalih promjena.

Nadalje, kako bi taj proizvodni proces bio održiv na duge staze, podatkovni centar mora imati način prenošenja znanja na novo osoblje. Za nešto kompleksno kao podatkovni centar, precizna i vrlo detaljna dokumentacija je kritična. Ne samo da je važno dokumentirati trenutno stanje postrojenja i svu opremu u njemu, potrebno je dokumentirati i sve promjene, stoga je nužno da SNUPI podržava normiranje, uvoz i izvoz podataka u standardnim formatima, kao što su npr. CSV ili XLS.

VI. OD TRENDOVA DO UPOZORENJA I ALARMA

Podaci koje sustav prikuplja u stvarnom vremenu čine samo jedan od izvora informacija o stanju imovine i zapravo predstavljaju vrlo malen udio. Takvi podaci se najčešće odnose na trenutnu potrošnju energije IT opreme i stanje njihovih virtualnih komponenata. Ovi podaci se prikupljaju tradicionalnim mrežnim protokolima kao što su SNMP ili putem modernih Web API-ja.

Druga vrsta podataka koja se prikuplja na nadzornom sustavu dolazi od energetske, rashladne i ostale tehničke infrastrukture. To su uređaji koji komuniciraju nizom starijih protokola, uključujući i SNMP, Modbus, LonWorks, BACnet, a u nekim slučajevima čak i drevni ASCII RS-232, poznatiji kao „serial port“. U većini slučajeva podaci prikupljeni sa ovih izvora se prikupljaju tek nekoliko puta po satu, kao što su temperatura i vlažnost zraka, ulazni i izlazni napon UPS-ova, napon, snaga i ukupna potrošnja cijele energetske mreže podatkovnog centra, itd. Takvi podaci zatim se zapisuju i analiziraju tijekom dužeg vremenskog razdoblja od analitičke strane SNUPI softvera u potrazi za bilo kakvim trendovima koji bi se mogli prepoznati. Općenito, što je više podataka prikupljeno, to bolje. Temeljem trendova unutar SNUPI-ja se generiraju upozorenja o degradaciji kvalitete i degeneraciji sustava, te o potrebnom prediktivnom održavanju koje je znatno pouzdanije i jeftinije od korektivnog i interventnog održavanja.

Za usporedbu u zdravstvu – sistematski pregled daje mogućnost za preventivnu zdravstvenu akciju, odnosno produžuje život i smanjuje potrebu za hitnom službom.

Ukoliko se dogodi iznenadni kvar izvan okvira očekivanih trendova, tek tada dolazi do alarma i tada je

potrebna trenutna fizička reakcija održavatelja (alarm = cash out), odnosno interventno održavanje sustava.

Vratimo li se nakratko usporedbi s automobilom, periodična provjera razine ulja je vjerojatno dobra navika, ali i u potpunosti suvišna kod „inteligentnih“ vozila koja vam govore kada trebate planirati održavanje; ista paralela se može vidjeti u podatkovnom centru sa nadzornim sustavom kada je u pitanju upravljanje imovinom ili održavanje postrojenja i opreme.

VII. PRIMJERI IZ HRVATSKE – DC UTRINE I DC SELSKA

2002. godine pristupilo projektnom zadatku izrade sustava centralnog nadzora rada elektroenergetskih i klima sustava, te praćenja klimatskih uvjeta u telekomunikacijskim objektima Hrvatskog Telekoma. Hrvatski Telekom ima ugovorno vezane tvrtke (kolokatore) koje unajmljuju HT-ove prostorije i iz njih nude svoje telekomunikacijske usluge. HT je dužan kolokatorima osigurati ugovorom definiranu energiju i klimatske uvjete. Svrha projekta je bilo postizanje slijedećih ciljeva:

- uvesti nadzor i točnu evidenciju uzroka i vremena trajanja „nenormalnih“ uvjeta kako kolokatori eventualne probleme sa energijom ne bi iskorištavali za opravdavanje bilo kakvih problema sa svojim sustavima, te u takvim situacijama tužili HT zbog gubitka prometa
- izgraditi sustav ranog upozorenja na incidentne uvjete kako bi se izbjeglo ili smanjilo trajanje ispada telekomunikacijske opreme
- uspostaviti nadzor nad potrošnjom električne energije telekomunikacijske opreme; potrošnja električne energije je direktno vezana za instaliranu telekomunikacijsku tehnološku opremu, odnosno ugovoreni kapacitet prometa, te se praćenjem potrošnje električne energije vrši ujedno i nadzor nad ugovorenim kapacitetom prometa kolokatora
- smanjenje troškova održavanja

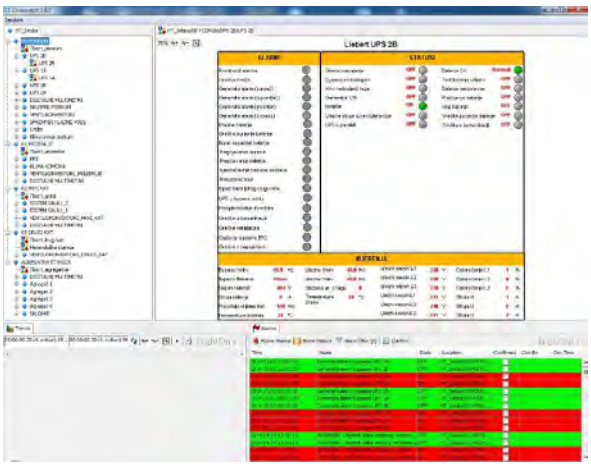
U projektu je bilo potrebno koristiti standardna rješenja nadzora, mjerenja, upravljanja i regulacije koji su trenutno bili instalirani u pogonima koje je obuhvaćao ovaj investicijski zahvat. Neke od smjernica za projektiranje su bile:

- neovisnost od ponuđača – projektirano rješenje treba biti standardni proizvod najnovije generacije nekog od vodećih svjetskih isporučitelja sličnih rješenja
- modularnost i skalabilnost – projektirano rješenje treba imati mogućnost jednostavnog dodavanja novih objekata i komponenti bez izmjena i narušavanja performansi rada postojećih sustava,
- standardiziranost – kod projektiranja koristiti standardizirane komponente (npr. senzori 4-20mA), standardne komunikacijske protokole (npr. TCP/IP, Modbus, LonWorks…), te standardne informatičke komponente (Microsoft OS, SQL relacijska baza, itd.)
- sigurnost – na osnovu analize rizika definirati kritične komponente sustava te mjere za sprječavanje ispada sustava i kritičnih komponenti (redundancija, povećana sigurnost, itd.)
- maksimalno iskoristiti već instaliranu opremu na objektima

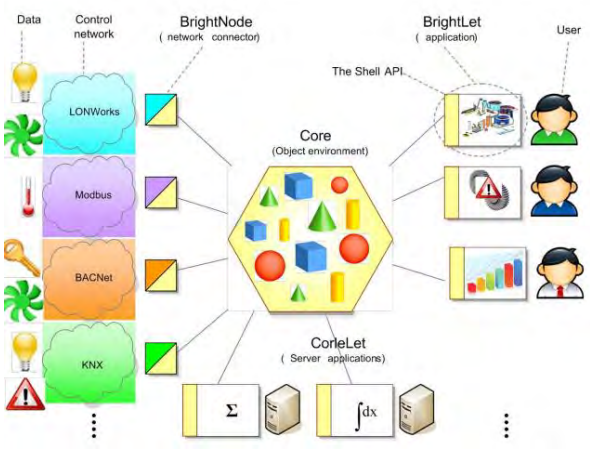
Kada se pristupilo projektnom zadatku, na lokaciji DC Utrine nalazila se oprema od otprilike 35 različitih proizvođača, sa otprilike 60 različitih uređaja na 5 različitih komunikacijskih protokola koji nisu bili međusobno povezani. Sa sveprisutnim ciljem smanjenja troškova, naglasak je bio na maksimalnom iskorištenju već instalirane opreme. Ovakav pristup problemu također je osigurao i izlazak iz tzv. silos scenarija ili „vendor lockinga“, odnosno neće se dopustiti niti jednom proizvođaču da bude nezamjenjiv opremom drugog proizvođača.

Početna ideja je, između ostalih, bila i implementacija običnog SCADA (eng. supervisory control and data acquisition) sustava nadzora, no zbog raznolikosti opreme i protokola ubrzo je postalo jasno da SCADA ne odgovara svim zahtjevima. Cijena integracije bi eksponencijalno rasla sa brojem proizvođača i aplikacija. Također, SCADA kao takva nije nudila nikakvu redundanciju i sva komunikacija prema niže spojenim uređajima se odvijala kroz jedan server, bez mogućnosti odvajanja serverske i klijentske strane, odnosno bez mogućnosti „pravog“ mrežnog rada – „pad“ servera bi značio kompletan „pad“ cijelog sustava.

Kako tada na tržištu nije postojalo gotovo rješenje koje bi zadovoljilo sve navedene zahtjeve, prvenstveno zbog raznolikosti uređaja i protokola koje je trebalo obuhvatiti, te kasnije i zbog rasprostranjenosti opreme, bilo je potrebno izraditi novo jedinstveno rješenje. Tako je nastao SNUPI, koji je integrirao sve uređaje u jednu zajedničku infrastrukturu.



Slika 1. SNUPI - Sustav nadzora i upravljanja procesima infrastrukture



Slika 2. Infrastruktura SNUPI sustava

SNUPI je danas u potpunosti distribuiran, mrežno orijentiran, „multi-protokol“ i „multi-vendor“ sustav, te „ispadanje“ servera ne utječe na procese. SNUPI je definiran kao interni „oblak“ sustav (eng. cloud), što znači da programu, evidencijama i dokumentaciji možemo pristupiti sa većeg broja uređaja u bilo koje vrijeme i sa više različitih lokacija, sve što nam je potrebno je veza prema Internetu, odnosno Intranetu.

Sastoji se od međusobno povezanih Fieldbus mreža, distribuiranih na više lokacija unutar RH, koje su ontološki povezane u jednu cjelinu preko infrastrukture objektnih usmjerivača, odnosno rutera (eng. router). Trenutni objektni ruteri procesiraju oko 60.000 objekata u realnom vremenu (eng. real-time). Objekt je u ovom smislu diskretan paket koji se sastoji od stanja i ponašanja, tj. u smislu SNUPI-ja objekti su raznorazne logičke komponente sustava, razni uređaji, njihovi senzori i drugo. Ukupan broj kontrolnih uređaja na sustavu trenutno broji preko 600 s oko 45.000 različitih varijabli i parametara.

Sustav sadrži i softversku sabirnicu (eng. bus) koja omogućava usmjeravanje podataka i upravljačkih signala između povezanih uređaja putem optimiziranog UCNP protokola. UCNP ujedno pruža i administracijsku cjelinu tako da se putem istog protokola rade sve promjene na mreži i to bez prekida rada mreže, odnosno izmjene se vrše samo na mrežama i uređajima na koje se odnose. Sustav uvijek čuva prethodnu konfiguraciju ukoliko je potreban povratak na stanje prije izmjena. Korisnik niti u jednom trenutku neće primijetiti intervenciju na mreži. Napravljen je i model mikro servisa (eng. corelet) koji omogućuje da se u potpunosti kontrolira svaka nadogradnja, odnosno da se ne mora nužno nadograđivati osnovna struktura sustava već se za potrebe integracije dodatne opreme ili aplikacija samo grade novi mikro servisi.

Sustav je potpuno distribuiran i prilagodljiv, te potpuno skalabilan i otvoren prema svim vrstama integracije, od razine adaptera do razine serverskih aplikacija. Cijela infrastruktura koristi TCP/IP stack izgrađen na ACE-u (eng. adaptive communication enviroment) tako da se sustav koji trenutno radi na Linux i Windows OS može jednostavno prilagoditi i prebaciti na bilo koji drugi operativni sustav po potrebi.

Povezanost sustava je izvedena tako da se iz nadzorne točke, ovisno o pravima krajnjeg korisnika, vidi svaki uređaj na svakoj mreži u cijelosti, uključujući i sve njegove ulazno-izlazne varijable, odnosno njegove interakcije s ostalim uređajima. Istovremeno se do nadzorne točke prenose i parametri kvalitete rada u obliku broja izgubljenih paketa pri komunikaciji između uređaja, odnosno prikupljaju se i analiziraju sve smetnje u komunikacijskim slojevima za svaku Fieldbus mrežu.

Korištene tehnologije Fieldbus mreža u SNUPI-ju su LonWorks, BACnet, Modbus i SNMP, a informacije se prosljeđuju putem TCP/IP i UDP/IP protokola. Pristupni objektni usmjerivači, odnosno ruteri pakiraju informaciju dobivenu kroz osnovni Fieldbus protokol u transportnu objektnu ovojnicu, te je nepromijenjenu prosljeđuju prema mjestima za obradu te informacije. Informacija se prenosi ili iz jedne mreže prema drugoj ili se prosljeđuje prema nekom uređaju koji je specijaliziran za praćenje i snimanje te informacije, alarmiranje ili prikaz na ekranu. Informacije se trenutno snimaju u SQL bazu podataka, a iste se mogu dalje obrađivati u bilo kojem ICT programu. Osim obrađivanja, moguće je izvršavati i naredbe koje izmjenjuju vrijednosti varijabli u samim uređajima.

Tipovi uređaja koji se nalaze na Fieldbus mrežama su programibilni, poput PLC uređaja (eng. programmable logic controllers) i raznih adaptera, te neprogramibilini poput npr. kontrolera ventilatora. Tu se još nalaze i specijalizirani uređaji PAC kategorije (eng. programmable automation controller), poput raznih senzora i pokretača, te njihovih kontrolera. Sustavom se trenutno prati okolina; ventilacija, grijanje, hlađenje i rasvjeta, te stanje UPS-ova, generatora, DC napajanja, statusi energetskih razvodnih ormara i potrošnja energije.

Ukupni broj uređaja i mreža može se povećavati u skladu s izgradnjom dodatne mrežno-transportno-procesne infrastrukture u smislu topologije i/ili aplikacija. Također, ukupan broj korisnika nije ograničen SNUPI tehnologijom i dodatni korisnici se mogu priključivati na internu cloud infrastrukturu. Nema ograničenja ni u broju fizičkih objekata koje sustav može nadzirati kao ni u broju mreža koje se mogu priključiti na sustav. Izrađen je i Builder alat koji omogućuje uvid u cijelu strukturu pojedine mreže. Na taj način izbjegao se model klasičnog „gateway-a“, te se kroz „point & click“ radnje može preusmjeriti vrijednost iz jednog uređaja na jednoj mreži na drugi uređaj na drugoj mreži, a pritom mreže mogu biti iste ili različite po tipu, kako na lokalnoj razini na pojedinom čvoru (eng. node) adaptera i lokalnih mikro servisa, tako i na višoj razini između node-ova. Ukoliko je potrebno postaviti dodatne uvjete za prijenos informacija, moguće je u sučelju Buildera napisati skripte u JavaScript-u, te je pored osnovnih skripti moguće stvarati i dodatne nove objekte.

SNUPI koristi Internet kao medij, s time da se sva komunikacija u stvarnom vremenu kriptira i dekriptira prema AES256 tipu enkripcije. Na taj način je sva interna komunikacija prilikom napuštanja Fieldbus mreže u potpunosti zaštićena od bilo kakvih nekontroliranih ili neovlaštenih aktivnosti.

Pored osnovne operativnosti sustav ima i ugrađen Cloud Fieldbus DNS koji omogućuje brzu izmjenu svakog objektnog rutera kao i aplikacijskog servera u obliku

rezervnog hardvera, koji se priključi na mjesto oštećenog uređaja i samim konfiguriranjem njegove lokacije se aktivira interni automatski servis koji taj uređaj završno konfigurira i integrira s ostatkom sustava.

Sustav ima otvoreni SDK (eng. software development kit) za C++ i Java programske jezike, a trenutno se radi na izradi RESTful Web aplikacije, odnosno sučelja baziranog na Node.js i Express.js „frameworku".

Obogaćen iskustvima iz DC Utrina, SNUPI je proširen na DC Selska gdje je dodatno unaprijeđen redundantnom izvedbom, a danas udaljeno nadzire i omogućava udaljeno upravljanje sa infrastrukturom u HT objektima diljem Hrvatske. Osim navedenih lokacija, u gradu Zagrebu se pod budnim okom SNUPI-ja još nalaze lokacije Draškovićeva, Prečko, Vukovarska i Kruge, a van grada Zagreba su obuhvaćeni gradovi Varaždin, Split, Osijek i Rijeka.

VIII. ZAKLJUČAK

Upravljanje procesima infrastrukture podatkovnog centra je donedavno bio fragmentiran i nedovoljno istražen tehnički izazov. Ovaj fizički sloj upravljanja je bio ili u potpunosti zanemaren ili rješavan empirijskim metodama predimenzioniranja i planiranja prevelike, tj. nepotrebne količine resursa. Opća smjernica do sada je bila jednostavno stvoriti obilje resursa, tako da se gornje granice nikada ne bi ni ispitale. Tek uz nedavni dramatični porast troškova energije i ubrzano kretanje IT industrije prema gustoj virtualizaciji, predimenzioniranje i gomilanje resursa se više ne smatra isplativim poslovnim ili strateškim planom.

Moderna nadzorno upravljačka rješenja, kao i SNUPI, dodiruju svaki aspekt rada podatkovnog centra, te vraćaju uloženo na nekoliko načina. Ljudi s različitim funkcijama i različitim ciljevima se oslanjaju na nadzorne sustave, na temelju različitih potreba koje imaju. Takav sustav je prije svega zanimljiv IT organizaciji, koja se sastoji od pojedinaca čija je karijera vođenje podatkovnog centra, odnosno briga o njegovom učinkovitom i neprekidnom radu. Odmah nakon njih je od ogromne važnosti i organizaciji koja se brine o napajanju, hlađenju, kao i o održavanju infrastrukture zgrade podatkovnog centra. Financijski odjel itekako je zainteresiran za odnos uloženog i dobivenog, pogotovo za troškove održavanja podatkovnog centra. Konačno; izvršni tim mora ozbiljno i pozitivno shvatiti utjecaj nadzornih sustava na zadatke upravljanja podatkovnim centrom.

Rukovoditelji podatkovnih centara su ušli u novu fazu učinkovitog upravljanja, koja obuhvaća spajanje i koordinaciju IT-a i procesne infrastrukture sa svrhom smanjenja troškova i stvaranja novih modela pružanja usluga.

LITERATURA

[1] Data Center Industry Survey 2014, Uptime Institute

[2] Energy Logic: Reducing Data Center Energy Consumption by Creating Savings that Cascade Across Systems, Emerson

CIS

International Conference on

INTELLIGENT SYSTEMS

Steering Committee

Chairs: Slobodan Ribarić, University of Zagreb, Croatia
Andrea Budin, Ericsson Nikola Tesla, Zagreb, Croatia

BiForD

Special Session on

BIOMETRICS & FORENSICS & DE-IDENTIFICATION AND PRIVACY PROTECTION

Chair: Slobodan Ribarić, University of Zagreb, Croatia

Program Committee:
Jean-François Bonastre, France
Bojan Cukic, USA
Patrizio Campisi, Italy
Joe Cannataci, Malta
Paulo Lobato Correia, Portugal
Jana Dittmann, Germany
Andrzej Drygajlo, Switzerland
Ivo Ipšić, Croatia
Carmen Garcia Mateo, Spain
Lily Meng, UK
Emilo Mordini, Italy
Ioannis Pitas, Greece
Bhiksha Raj, USA
Daniel Ramos, Spain
Tieniu Tan, China
Zheng-Hua Tan, Denmark
Isabel Trancoso, Portugal
Marcos Faundez Zanuy, Spain

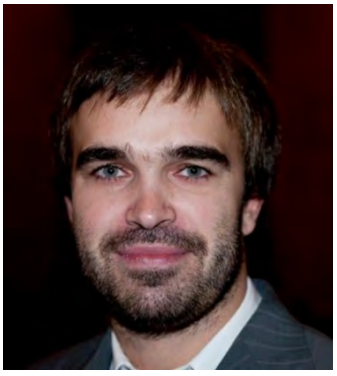
cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY
COST ACTION IC1206
DE-ID De-identification for privacy protection in multimedia content
COST Action IC1106

KEYNOTE SPEECH

Face Alignment: Addressing Pose Variability in Face Recognition Systems

Vitomir Struc
University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia
vitomir.struc@luks.fe.uni-lj.si

Pose variability represents one of the major problems not satisfactory solved yet by the existing face-recognition technology. Recent advancements in facial landmark localization and image alignment have made it possible to devise new methods and techniques to mitigate this problem and successfully match facial images captured under significantly different viewpoints (or poses). In this talk, we will discuss some of the methods that can be used to address pose variability in facial recognition systems and present our recent work on *(a)* facial landmark detection from 2D and 3D data, *(b)* face frontalization, and *(c)* reposing of facial images. We will elaborate on some of the recent trends in the area of facial alignment and highlight other application areas where the presented techniques can be used as well.



Vitomir Struc is an Assistant Professor at the Faculty of Electrical Engineering at the University of Ljubljana. He received the PhD degree in electrical engineering from the University of Ljubljana in 2010. He is the recipient of the "2015 young researcher award" of the University of Ljubljana. Vitomir participated in several national and EU funded R&D projects and authored (or co-authored) more than 50 research papers for leading international peer reviewed journals and conferences related to different areas of computer vision, image processing, pattern recognition, and biometrics. He served on the organizing committees of several workshops and international conferences and is currently a guest editor for the special issue on "The Best of Face and Gesture 2015" of the Image and Vision Computing Journal. Vitomir is a member of the IEEE, IAPR, the national contact point for the EAB and the former vice-president and current treasurer of the Slovenian Pattern Recognition Society.

PAPERS

Shape and Texture Combined Face Recognition for Detection of Forged ID Documents

Daniel Sáez-Trigueros, Heinz Hertlein, Li Meng*
School of Engineering and Technology
University of Hertfordshire
Hatfield AL10 9AB, UK
d.saez-trigueros@herts.ac.uk, L.1.MENG@herts.ac.uk

Margaret Hartnett
IDscan Biometrics Ltd
London E14 9QD, UK

***Abstract*— This paper proposes a face recognition system that can be used to effectively match a face image scanned from an identity (ID) document against the face image stored in the biometric chip of such a document. The purpose of this specific face recognition algorithm is to aid the automatic detection of forged ID documents where the photography printed on the document's surface has been altered or replaced. The proposed algorithm uses a novel combination of texture and shape features together with subspace representation techniques. In addition, the robustness of the proposed algorithm when dealing with more general face recognition tasks has been proven with the Good, the Bad & the Ugly (GBU) dataset, one of the most challenging datasets containing frontal faces. The proposed algorithm has been complemented with a novel method that adopts two operating points to enhance the reliability of the algorithm's final verification decision.**

Keywords—face recognition, shape and texture combined features, detection of forged IDs

I. INTRODUCTION

Even though the first successful face recognition algorithms dated from the late eighties, it is still a vibrant area of research with new and better techniques appearing every year. One of the main reasons for the popularity of face recognition is the wide range of potential applications, including access control, identification systems, surveillance, and identity verification, to name but a few. This paper investigates the applicability of face recognition to the authentication of biometric ID documents, i.e., the process of analysing an ID document in order to prove its legitimacy. ID documents are specifically designed with security features (watermarks, holograms, special materials, etc.) to avoid counterfeiting and forgery. A counterfeit document is a complete reproduction of a document from scratch to resemble an officially issued document, whereas a forged document is a genuine document that has been illegally altered in some way.

Some organisations have studied how to detect fraudulent identity documents by examining their security features [1], [2]. This can be done manually by human operators, or automatically by processing the document electronically. In practice, the most effective method is a combination of both: documents are electronically validated and referred to a human operator when the automatic electronic validation fails for any reason. Nowadays, millions of ID documents are manually examined by human operators every day in different businesses and organisations. Thus, this study is motivated by the growing need for automatic or semi-automatic methods for authenticating ID documents. This paper introduces an additional automatic validation check to detect forged documents by using face recognition for the comparison of the image printed on the ID document to the image stored in the ID document's biometric chip. This validation check aims to reduce one common method of forgery, namely photography substitution [1], [2].

The problem presented here is a specific face recognition case since it effectively compares whether the images themselves are identical (i.e. derived from the same camera shot). Hence, the main difficulties arise from the presence of watermarks, holograms, reflections and other imperfections on the scanned image as shown in Fig. 1. Other studies [3], [4] have investigated the problem of matching degraded face images scanned from passports to high-resolution digital face images. Those studies have proposed pre-processing methods to improve the quality of the scanned face images for their later use in a face recognition algorithm. While these pre-processing steps are likely to improve the overall performance, they are intrinsically domain specific. By not applying any domain specific pre-processing, the proposed algorithm is more generalised and can be used to compare any two face images. In order to test its accuracy for both the specific application considered here as well as more general face recognition applications, the proposed algorithm has been evaluated on two different datasets, namely (i) a proprietary dataset (hereinafter referred to as the BiometricID dataset) containing face images scanned from ID documents and digital images from the biometric chip of those documents; and (ii) the Good, the Bad & the Ugly (GBU) dataset, which contains pairs of frontal face images with three levels of difficulty [5].

The proposed face recognition method is based on the fusion of texture-based features and shape-based features. In particular, Scale-Invariant Feature Transform (SIFT) descriptors are used to extract the texture features, and a set of coefficients that represent relative distances between pairs of facial landmarks is used to describe the shape of the face. These two different types of features are further processed using Principal Component



Fig. 1. Example of image printed on an ID document (left) and image stored in the biometric chip of the same document (right).

Analysis (PCA) and Linear Discriminant Analysis (LDA) to project the features to a lower dimensional and more discriminating space. Furthermore, the use of two operating points is proposed to provide a greater degree of control over the verification decision.

The remainder of the paper is structured as follows. Section II provides a subject review with a focus on the approaches adopted in this study. Section III describes the proposed algorithm in detail and Section IV presents and analyses the experimental results. Finally, conclusions are drawn in Section V.

II. Subject Review

A. Face Alignment with Constrained Local Neural Field

The term face alignment is typically used in the literature to refer to the automatic detection of facial landmarks in a face image or video in order to support further processing. Those landmarks can be used to normalise faces to a canonical view using 2D or 3D transformations or to extract shape features out of the face image. The algorithm proposed in this paper makes use of the Constrained Local Neural Field (CLNF) algorithm [6], an extension of the Constrained Local Model (CLM) [7]. CLM and CLNF are based on the popular Active Appearance Model (AAM) method [8]. AAM is a statistical model of shape and texture created from a training set of manually annotated face images and subsequently used to fit new unseen images. The fitting process starts by placing landmark points on an image using the mean location of each landmark in the training set. Then, the texture residual between the current estimate and the model is calculated, and the shape parameters updated in order to minimise that residual. In both CLM and AAM methods, a set of rigid and non-rigid shape parameters $\mathbf{p} = [s, \mathbf{R}, \mathbf{t}, \mathbf{q}]$ models the positions of the predefined set of landmarks

$$\mathbf{x} = s\mathbf{R}(\bar{\mathbf{x}} + \mathbf{\Phi q}) + \mathbf{t}, \tag{1}$$

where $\mathbf{x}$ are the locations of the set of facial landmarks defining the face shape in the given image, $\bar{\mathbf{x}}$ the mean shape of the faces in the training set, $\mathbf{\Phi}$ the principal component matrix describing the modes of variation among the face shapes and $\mathbf{q}$ a vector of weights that control the non-rigid shape defined by $\mathbf{x}$. The rigid shape transformations are controlled by a scaling term s, a translation term $\mathbf{t}$, and a rotation matrix $\mathbf{R}$.

CLM methods offer better performance than AAM by using local descriptors to represent the texture surrounding each landmark location instead of utilising a global texture model of the whole face. The CLNF method, in particular, uses a local descriptor based on a neural network with one hidden layer and similarity and sparsity spatial constraints in the output to enhance the accuracy [6].

B. Scale-Invariant Feature Transform

SIFT descriptors have been extensively used in object and face recognition [9]. The original SIFT algorithm [10] finds key points on an image and calculates a descriptor for each one of these key points. This section summarises how to calculate the descriptor. The process of finding the key points is not discussed here since the algorithm proposed in Section III calculates descriptors on fixed locations within the image.

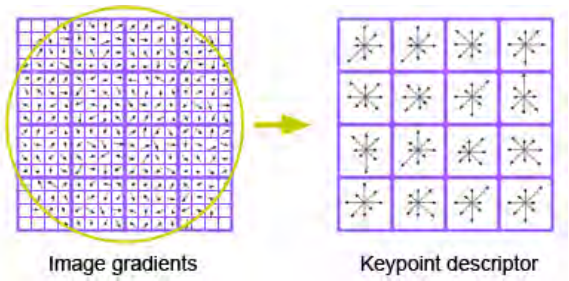


Fig. 2. Orientation histograms created from gradients sampled over a 16 x 16 region. The Gaussian weighting function is represented with a circular window on the left figure. Image taken from [11].

The first step is to compute gradients in a region of 16 x 16 pixels around the centre of the key point. Then, the magnitude of each gradient is weighted using a Gaussian weighting function to avoid abrupt changes in the descriptor and give less emphasis to gradients located far from the centre. Orientation histograms of 8 bins are then created using gradients within 4 x 4 sub regions. In an orientation histogram the bins can be represented by vectors pointing to different directions, where the length of each vector is the magnitude of the corresponding bin (see Fig. 2). The final descriptor is formed by concatenating the magnitudes of the bins of all the orientation histograms in the 16 x 16 region. Since there are 4 x 4 histograms, each containing 8 bins, the dimension of the descriptor is 128 [10].

C. PCA and LDA in Face Recognition

The *Eigenfaces* algorithm based on PCA was one of the first successful face recognition techniques [12], and although its performance has since been surpassed by more advanced algorithms, PCA is still a relevant technique used in many modern face recognition algorithms. One of the limitations of PCA is that it does not use class labels, meaning that faces from the same identity and faces from different identities are treated in the same way.

LDA is a similar approach to PCA but uses the class (identity) labels to find a projection that minimises the variation within the classes while maximising the variation between classes, i.e., the identity labels are used to reduce intra-person variations while increasing inter-person variations. PCA is typically used before LDA to reduce the dimensionality of the input feature vector, as LDA does not perform well when the within-class scatter matrix is not well estimated. This happens when there are not enough samples in the training set compared to the high dimensionality of the input feature vector [13], [14]. This combination of PCA before LDA has been adopted in this work.

III. Proposed Algorithm

A. Face Normalisation

The Viola-Jones object detector [15] is applied to detect the position of the face(s) in a given image. In order to perform face alignment, the CLNF landmark detector is used to locate a set of 68 landmarks in each face [6]. Once the landmarks have been located, the positions of the pupils (which are found at the intersecting point defined by the landmarks surrounding the eye sockets) are used to normalise the face images to a common scale (128 x 128 pixels) and crop area, with the eyes located at fixed locations. Lastly, the image is converted to greyscale.

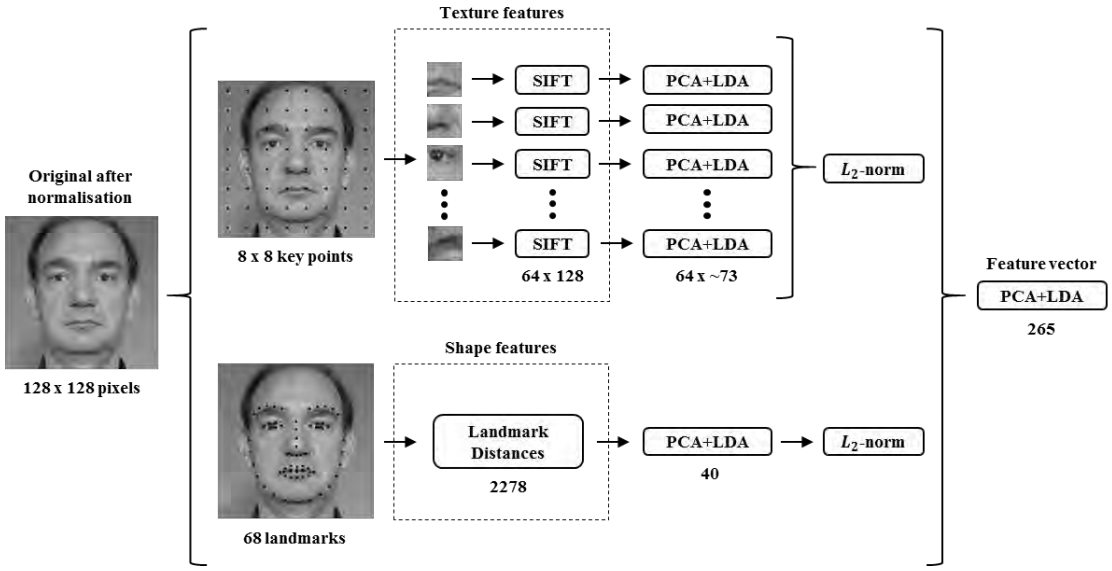


Fig. 3. Diagram illustrating the face representation step of the proposed algorithm. The figures below each block indicate the number of dimensions after each step.

B. Face Representation

The face images are represented in two steps: firstly, a number of features are calculated to extract relevant information that are more informative than the raw pixels; secondly, the extracted features are transformed using machine learning techniques in order to create feature vectors that are more discriminating and, ideally, unique to each person.

1) Feature Extraction

Two different types of features are calculated in the proposed algorithm: texture-based features and shape-based features.

Texture features are the most popular kind of features for face recognition. They can represent more information than shape features as they are directly computed from the raw pixels. Shape features are usually calculated from landmark locations within the face image, which means that their reliability depends heavily on the accuracy of the landmark localisation algorithm. On the other hand, shape features can improve robustness in situations when the appearance of a face changes but not its geometry, for example, comparing face images of the same person with and without facial hair, makeup, or glasses. In this work, shape features are fused with texture features to boost the recognition accuracy of the proposed algorithm.

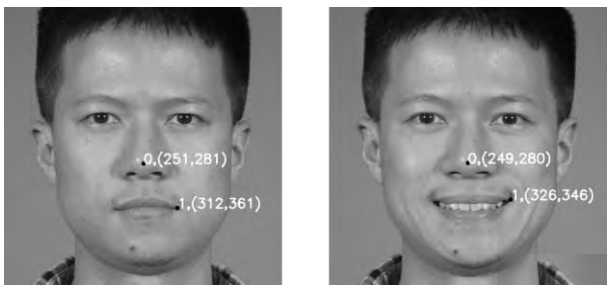


Fig. 4. The Euclidean distance between the highlighted landmarks when the expression is neutral (left image) is 100.60, and 101.41 when the expression is smiling (right image), i.e. the relative distance between the landmarks remains almost the same. On the other hand, the misalignment of the landmark in the right corner of the mouth when the person smiles is 20.52 with respect to the case in which the expression is neutral.

a) Texture features

Inspired by [16], [17], the texture features used in the proposed algorithm consist of SIFT descriptors densely extracted from local patches across the face image. In particular, 64 SIFT descriptors with a radius of 6 pixels are extracted from a 8 x 8 grid of fixed key points as shown in the top half of Fig. 2.

b) Shape features

The shape features used in the proposed algorithm are Euclidean distances between pairs of landmarks. The distances are calculated after normalising the face image so that the effect of scaling is removed. The total number of possible distances is given by $\binom{n}{2} = n(n-1)/2$, where n is the total number of landmarks. As shown in the bottom half of Fig. 3, 68 landmarks are used resulting in a total number of landmark distances of 2,278.

Other studies have considered the use of landmarks for face representation. For example, the use of landmark coordinates and distances between pairs of landmarks has been studied in [18]. Their ratio-based model is very similar to the approach adopted in this paper. However, they made use of fewer landmark points and removed half of the distances due to the similarity between both sides of the face. In contrast, the approach adopted in this work uses all the possible landmark points to represent as much shape information as possible. In [19] the use of landmark locations for affine-invariant shape representation is studied. However, shape variations encountered in real datasets are not only a result of affine transformations but also of non-rigid transformations caused by, for example, changes in facial expression. As shown in Fig. 4 the shape features based on distances between pairs of landmarks are more robust to facial expressions than the shape features based on landmark locations.

2) Subspace Representation

Multivariate statistical tools can be used to transform a high dimensional space spanned by a large number of features into a lower dimensional space that retains the most useful information for discriminating the original samples.

As suggested in Section II.C, the proposed algorithm uses a common method in face recognition wherein LDA is applied to the subspace obtained by first applying PCA to the input data. In this work, PCA+LDA projections are used to independently transform the space spanned by each SIFT descriptor and the space spanned by the shape features. Other studies [17], [20], [21] have used a random sampling technique [22] to reduce the dimensionality of very high dimensional feature spaces (e.g. the space that would result from concatenating all the shape and texture features). However, independently applying PCA+LDA to lower dimensional spaces as proposed here has the advantage of eliminating the need of such random sampling techniques, which might otherwise accidentally remove highly discriminating features.

Using a PCA that retains 98% of the original variance, the 128-dimensional SIFT descriptors become (on average) 73-dimensional when the FRGC dataset is used for training (see Section IV.B), and the 2,278-dimensional vector representing the landmark distances becomes 40-dimensional. The observed significant reduction in dimensionality of both the texture and shape vectors, notwithstanding the retention of 98% of the variance in the data, substantiates the notion that many distances between pairs of landmarks must be highly correlated as they are calculated from landmarks being located next to each other.

The next step is to combine the shape and texture features. First, all the PCA+LDA projected texture features are concatenated resulting in a 4,670-dimensional vector ($64 \times \sim 73$) and normalised to unit L_2-norm. The PCA+LDA projected shape features are normalised in the same way prior to being concatenated with the texture features. The feature vector resulting from concatenating the shape and texture features is 4,710-dimensional (4,670 texture features and 40 shape features) and might contain redundant information. For this reason, another PCA+LDA transformation is performed to project the information contained in the concatenated texture and shape features into a more discriminating feature space with even lower dimensionality. In this case a PCA retaining 90% of the variance is performed to reduce the dimensionality to 265 before applying LDA. A diagram describing the entire face representation step is depicted in Fig. 2.

Another way of fusing multiple modalities such as face texture and shape is to have separate recognition algorithms for each modality and then combine their scores [19]. However, feature-level fusion has the advantage of having to train a single algorithm, and eliminates the need to optimise the fusion weighting. Liu and Wechsler [23] used a feature-level fusion method similar to the one proposed here. However, the input features in [23] are the raw image pixels and the landmark coordinates instead of the more informative features considered in this study.

C. Face Matching

In this work, the matching score between two feature vectors $\mathbf{a}$ and $\mathbf{b}$ is calculated using the cosine similarity:

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}. \quad (2)$$

Typically, the operating point of the algorithm is determined by a threshold t used to decide whether two face images match or not. This threshold defines the true and the false acceptance rates (TAR and FAR) and the true and the false rejection rates (TRR and FRR). When analysing such a conventional biometric system that operates with one threshold only, the verification accuracy is completely specified by the pair of error rates FAR and FRR, as $\mathrm{TAR} = 1 - \mathrm{FRR}$ and $\mathrm{TRR} = 1 - \mathrm{FAR}$. As shown in Fig. 5, with one threshold and one operating point, FRR can be reduced arbitrarily at the expense of increasing FAR and vice versa.

In order to alleviate the conflict between FRR and FAR, an approach that uses two thresholds, i.e., two operating points, is adopted here. In this approach, the higher threshold t_h controls the accepted comparisons (TAR and FAR), whereas the lower threshold t_l controls the rejected comparisons (TRR and FRR). A matching score in between the two operating points would yield an undetermined result. As shown in Fig. 6, this approach allows both FRR and FAR to be reduced at the same time, as they are determined separately by two different thresholds. In real applications, it is often desirable to achieve very low error rates whilst it is acceptable to have a certain number of undetermined cases, which might be handled further in a specific way (e.g. by manual inspection). Therefore, the introduction of a second threshold represents a significant improvement in terms of the applicability of face verification in practice.

IV. Experimental Results

This section presents details about the protocol and the datasets adopted for the evaluation of the proposed algorithm (Section IV.A), the subsets used for training (Section IV.B), and the evaluation results (Section 0).

A. Evaluation Protocol and Datasets

The protocol adopted for the evaluation of the algorithm is based on the protocol as defined by NIST for its face recognition challenges [24]. In this protocol, the algorithms are tested by comparing all the images in a target set to all the images in a query set. The resulting scores are used to generate a Receiver Operating Characteristic (ROC) curve that plots TAR against FAR.

The proposed algorithm is evaluated on two different datasets. The BiometricID set is a proprietary database where for each subject there are face images scanned from his/her ID document (i.e. scanned images with artifacts such as watermarks, holograms, etc.), and the original digital face image obtained from the biometric chip of the same ID document (i.e. the RFID image). For some subjects there are more than one scanned images (as the artifacts can vary from one scan to the next), and some others do not have any scanned or RFID image. In such cases, all the images are used to form additional matching or non-matching pairs. For example, if a subject only has one scanned image and no RFID image, the scanned image is compared against all the RFID images in the evaluation set to produce extra non-matching pairs. In total, the target set contains 4,802 RFID images and the query set 8,801 scanned images forming 6,029 matching pairs and 42,256,373 non-matching pairs.

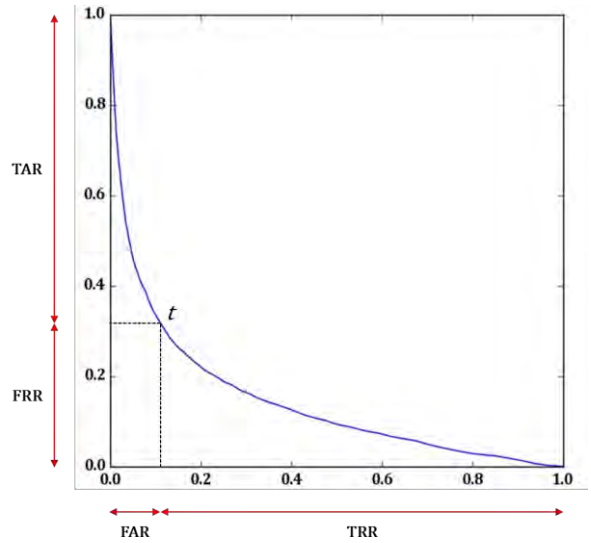


Fig. 5. ROC curve plotting FAR against FRR and a possible threshold. It can be seen the effect that increasing or decreasing the threshold t has on TAR, FRR, FAR and TRR.

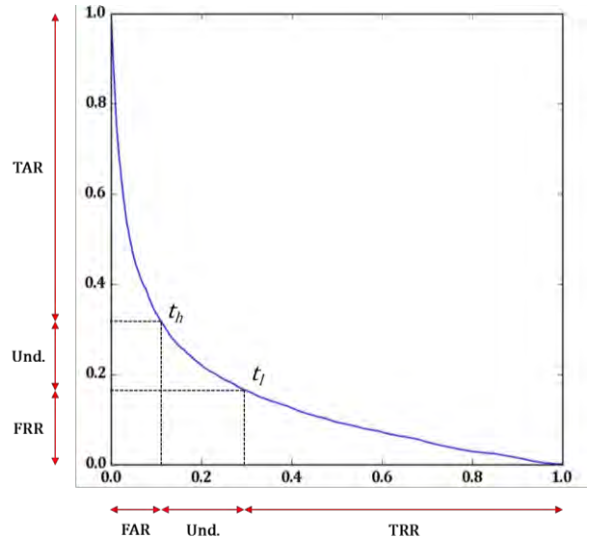


Fig. 6. ROC curve plotting FAR against FRR and two possible thresholds. In comparison with Fig. 4, the introduction of the lower threshold t_l allows to reduce FRR at the expense of reducing TRR and having comparisons where the result is undetermined. In real applications this is often desirable, as the number of comparisons between non-matching faces is usually much lower than the number of comparisons between matching faces.

The second database is the Good, the Bad and the Ugly public dataset [5], one of the most challenging sets available to evaluate the performance of face recognition algorithms on frontal images. The GBU dataset is divided into three partitions. The Good partition consists of face pairs easy to match, the Bad partition consists of face pairs with an average matching difficulty and the Ugly partition consists of face pairs difficult to match. The query and target sets of each partition contain 1,085 images from 437 subjects. The distribution of images on each query and target set, i.e., the number of images per subject, is the same across the three partitions. In total, there are 3,297 matching pairs and 1,179,928 non-matching pairs on each partition.

B. *Training*

Two different sets of images have been used to train the algorithm, one for each dataset evaluated in this study. One is a random subset of the BiometricID dataset for evaluation on the BiometricID dataset, and the other is a random subset of the FRGC dataset [25] for evaluation on the GBU dataset. Since the GBU dataset does not provide training images, the FRGC dataset is used as both of them contain images that were collected by the University of Notre Dame under similar conditions.

The BiometricID training set contains 1,000 images from 500 subjects, with 2 images per subject, one RFID image and one scanned image. Only 2 images per subject are used since any additional sample available for a subject is simply a different scan of the face image printed on the ID document. The training set used for evaluating the algorithms on the GBU dataset contains 5,320 images of a total of 266 subjects from the FRGC dataset with 20 images per subject. Recognition of face images on the BiometricID dataset is considerably easier than recognition on the GBU dataset in the sense that the algorithm requires less training data to perform well on the BiometricID dataset. This is because for the face images in the BiometricID dataset the PCA+LDA only needs to learn feature vectors to match two different versions of the same face image (RFID and scanned) and differentiate them from feature vectors generated for face images of other subjects; whereas the more general scenario presented by the GBU dataset requires similar (matching) feature vectors generated for various, and sometimes dramatically different, face images of each and every subject in the dataset.

C. *Results*

The open source implementation of the 4SF algorithm described in [16] has been chosen as the baseline for the performance evaluation in this work. The 4SF algorithm is a good example that makes use of local descriptors and subspace representation. To focus the comparison on the face recognition accuracy rather than on the face detection accuracy, the same face alignment technique has been adopted in both algorithms. For this reason, the 4SF algorithm was modified to use the CLNF landmark detector to locate the position of the eye pupils. Moreover, the 4SF algorithm is trained and evaluated using exactly the same images as the proposed algorithm.

As seen in the ROC curves obtained with the BiometricID dataset (Fig. 7) and the GBU dataset (Fig. 8), the proposed algorithm outperforms the baseline algorithm in all cases. This implies that (i) the proposed combination of texture and shape features possesses more discriminating information than texture features alone (at least when evaluating datasets that contain mainly frontal images), and (ii) the proposed PCA+LDA applied to multiple low dimensional spaces is a more effective dimensionality reduction strategy than the random sampling technique used in 4SF.

V. CONCLUSIONS

This paper presents a novel face recognition algorithm tailored for a specific application scenario of face recognition that involves face images with added security features such as watermarks and holograms for the detection of forged ID documents.

Considering the similar face geometry across faces of the same subject, the proposed algorithm fuses face shape features with the commonly used face texture features. The high dimensionality produced by the large number of shape and texture

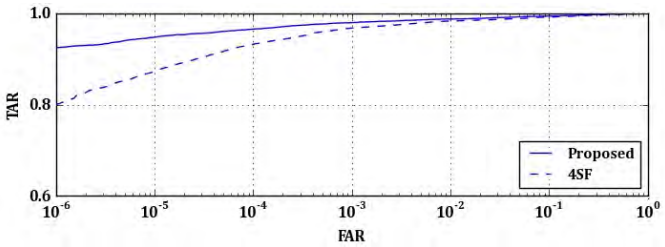


Fig. 7. ROC curves on the BiometricID dataset using the 4SF algorithm and the proposed algorithm.

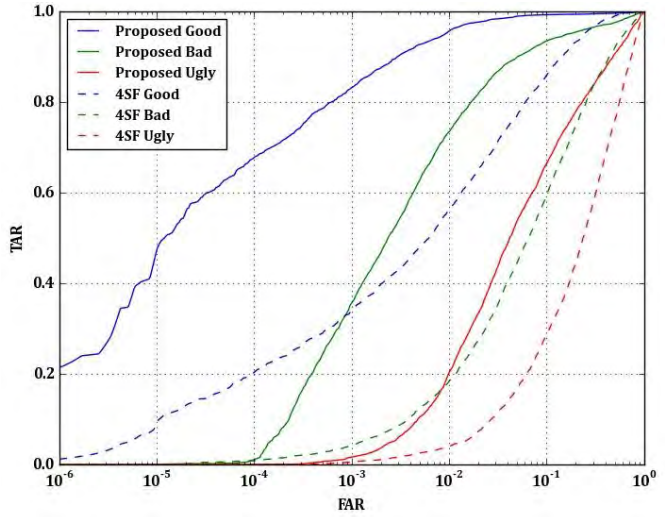


Fig. 8. ROC curves on the Good, Bad and Ugly partition of GBU dataset using the 4SF algorithm and the proposed algorithm.

features has been avoided by using multiple PCA+LDA transforms. The proposed algorithm has achieved high accuracy for the specific application scenario considered here. In addition, the robustness of the proposed algorithm for more generic face recognition tasks has been confirmed by its performance on the GBU dataset, offering a good balance between training and accuracy in the sense that there is no need to train using many thousands of images as is the case with the latest face recognition algorithms based on deep learning [26].

Finally, it has been shown how the applicability of any face recognition algorithm in real applications can be benefited with the use of two thresholds to have a better control on the face recognition rates at the expense of having comparisons where the result is undetermined.

REFERENCES

[1] 'Guidance on Examining Identity Documents'. UK Home Office, 18-Nov-2014.

[2] 'A Good Practice Guide On Pre-Employment Screening - Document Verification'. Centre for the Protection of National Infrastructure (CPNI), Jun-2007.

[3] T. Bourlai, A. Ross, and A. Jain, 'On matching digital face images against scanned passport photos,' in *Proc. First IEEE Int. Conf. Biometrics, Identity and Security (BIDS)*, 2009.

[4] T. Bourlai, A. Ross, and A. K. Jain, 'Restoring Degraded Face Images: A Case Study in Matching Faxed, Printed, and Scanned Photos,' *IEEE Trans. Inf. Forensics Secur.*, vol. 6, no. 2, pp. 371–384, Jun. 2011.

[5] P. J. Phillips *et al.*, 'An introduction to the good, the bad, & the ugly face recognition challenge problem', in *IEEE Int. Conf. Automatic Face & Gesture Recognition and Workshops,* 2011, pp. 346–353.

[6] T. Baltrusaitis, P. Robinson, and L.-P. Morency, 'Constrained local neural fields for robust facial landmark detection in the wild,' in *IEEE Int. Conf. Computer Vision Workshops (ICCVW)*, 2013, pp. 354–361.

[7] D. Cristinacce and T. F. Cootes, 'Feature Detection and Tracking with Constrained Local Models,' 2006, pp. 95.1–95.10.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, 'Active appearance models,' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[9] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli, 'On the Use of SIFT Features for Face Authentication,' in *Conf. Computer Vision and Pattern Recognition Workshop*, 2006, pp. 35–35.

[10] D. G. Lowe, 'Distinctive image features from scale-invariant keypoints,' *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[11] 'Computer Vision Project'. [Online]. Available: http://www.cc.gatech.edu/~hays/compvision/results/proj2/rchawla8/index.html. [Accessed: 05-Jan-2016].

[12] M. A. Turk and A. P. Pentland, 'Face recognition using eigenfaces,' in *IEEE Conf. Computer Vision and Pattern Recognition,* 1991, pp. 586–591.

[13] W. Zhao *et al.*, 'Discriminant Analysis of Principal Components for Face Recognition,' in *Face Recognition*, Springer Berlin Heidelberg, 1998, pp. 73–85.

[14] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, 'Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, 1997.

[15] P. Viola and M. J. Jones, 'Robust real-time face detection,' *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[16] J. C. Klontz *et al.*, 'Open source biometric recognition,' in *IEEE Conf. Biometrics: Theory, Applications and Systems (BTAS)*, 2013, pp. 1–8.

[17] Z. Li, U. Park, and A. K. Jain, 'A Discriminative Model for Age Invariant Face Recognition,' *IEEE Trans. Inf. Forensics Secur.*, vol. 6, no. 3, pp. 1028–1037, Sep. 2011.

[18] J. Shi, A. Samal, and D. Marx, 'How effective are landmarks and their geometry for face recognition?' *Comput. Vis. Image Underst.*, vol. 102, no. 2, pp. 117–133, May 2006.

[19] T. Wu, P. Turaga, and R. Chellappa, 'Age Estimation and Face Verification Across Aging Using Landmarks,' *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 6, pp. 1780–1788, Dec. 2012.

[20] B. Klare and A. K. Jain, 'Face recognition across time lapse: On learning feature subspaces,' in *Int. Joint Conf. Biometrics*, 2011, pp. 1–8.

[21] B. F. Klare *et al.*, 'Face Recognition Performance: Role of Demographic Information,' *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 6, pp. 1789–1801, Dec. 2012.

[22] X. Wang and X. Tang, 'Random Sampling for Subspace Face Recognition,' *Int. J. Comput. Vis.*, vol. 70, no. 1, pp. 91–104, Oct. 2006.

[23] C. Liu and H. Wechsler, 'A shape-and texture-based enhanced Fisher classifier for face recognition,' *IEEE Trans. Image Process.*, vol. 10, no. 4, pp. 598–608, 2001.

[24] S. Rizvi, J. Phillips, and H. Moon, 'The FERET verification testing protocol for face recognition algorithms,' in *Proc. Third IEEE Int. Conf. Automatic Face and Gesture Recognition,* 1998, pp. 48–53.

[25] N. US Department of Commerce, 'Face Recognition Grand Challenge (FRGC)'. [Online]. Available: http://www.nist.gov/itl/iad/ig/frgc.cfm. [Accessed: 04-Aug-2015].

[26] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, 'Deepface: Closing the gap to human-level performance in face verification,' in *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1701–1708.

This study is related to a collaboration project between the University of Hertfordshire and the company IDscan Biometrics Ltd as part of a Knowledge Transfer Partnership (KTP) programme.

Simple Method Based on Complexity for Authorship Detection of Text

Lukáš Meluch*, Ivana Tokárová*, Peter Farkaš*,**
*Institute of Telecommunications
Faculty of Electrical Eng. and Information Technology,
Slovak University of Technology,
Bratislava, Slovakia
xmeluchl@stuba.sk, xtokarovai@stuba.sk,
p.farkas@ieee.org

Frank Schindler**
**Institute of Applied informatics
Faculty of Informatics, Pan European University
Bratislava, Slovakia
frank.schindler@paneurouni.com

Abstract— **Authorship-identification is an actual research area with different applications. Particularly it is a question if only highly specialised and well equipped entities could identify an author or it could be made by broader community via exploiting text. In order to shed light on this question a very simple and computationally not demanding method for author identification was proposed based on lossless compression algorithms. In this paper another such method based on complexity is proposed and its verification on real data is presented. Results of experiments proved, that also such a simple method could be used successful identification of authors.**

Keywords— Authorship –identification; text; simple; method; complexity

I. INTRODUCTION

In connection with Big Data and Cloud it would be useful to know if entities with comprehensive hardware and sophisticated algorithms could successfully identify an author of a text or if it could be done by broader community by using simple, computationally not demanding methods. In order to contribute to answering this question a very simple method with low computational complexity for author identification was proposed based on lossless compression algorithms in [1].

In this contribution another simple method based on complexity is proposed and its verification on real data is presented. The paper is organized as follows. In Section II. new simple method for authorship identification from text is proposed. In Section III. experiments with real text information, namely sonnets and The Federalist Papers are described and their results are presented in tables. In Section IV. Some concluding remarks are given.

II. NEW SIMPLE METHOD FOR AUTHORSHIP IDENTIFICATION

In this manuscript we will under written text understand also text which is not written by hand but produced by digital terminals such as smartphones, laptops etc. In a digital form the graphical characteristics from a hand written text are lost and the authorship detection has to be based on other characteristics. It was observed, that such a text by a particular person is influenced by specific style of its author. It can influence the length of used words their position, repetition, frequency and other characteristics of the text. In digital domain, these artefacts could appear as bursts of identical subsequences (substrings). The consequence of this is that if we use lossless compression algorithm in which one text is used as a vocabulary for compression of the other one, the achieved reduction depends on the similarity of the two text data [1]. In this manuscript further simplification of this method is based on hypothesis that a complexity measurement of two concatenated textual strings from different authors is higher than the complexity of a string composed by pair of sequences from the same author. Therefore in this chapter we will first sketch some known measures of complexity and then present the simple method for authorship estimation based on their exploitation.

A. Complexity

Complexity in this paper will be used in its meaning as it is common in Information Theory. Namely complexity is used for information contents quantification of an object, usually a sequence via its shortest description. In other words if a sequence could be described by a short description it has also a low complexity. Longer sequence than a shorter one could have a lower complexity for example if some strings are repeated in the longer one. There are more known methods how to measure complexity. The basic concept was originally proposed by Kolmogorov [2] and later it inspired Ziv and Lempel to propose their measure for complexity in [3].

Let $\mathbf{S} = s_1 \quad s_2 \quad \cdots \quad s_n$ and $\mathbf{S}(i,j) = s_i \quad s_{i+1} \quad \cdots \quad s_j$ be a sequence with n symbols and its subsequence or string with length $j - i + 1$ respectively. The sequence $\mathbf{S}$ could be expressed as:

$$H(\mathbf{S}) = \mathbf{S}(1, s_1), \quad \mathbf{S}(s_1 + 1, s_2), \quad \cdots \quad \mathbf{S}(s_{m-1} + 1, s_m) \tag{1}$$

In (1) $H(\mathbf{S})$ is history of $\mathbf{S}$. One sequence could have many histories. Therefore the authors defined also complete history $E(\mathbf{S})$. In it no string could be repeated or in other words:

$$\mathbf{S}(i,j) \not\subset \mathbf{S}(1, j-1) \tag{2}$$

For example if $\mathbf{S} = A \quad A \quad B \quad C \quad D \quad A \quad A \quad B \quad D$, than its history could be represented with different expressions, for example:

$H(\mathbf{S}) = A, \quad A, \quad B, \quad C, \quad D, \quad A, \quad A, \quad B, \quad D$

$H(\mathbf{S}) = A, \quad A, \quad BCD, \quad A, \quad ABD$

$H(\mathbf{S}) = A, \quad A, \quad B, \quad CDA, \quad ABD$

In this case $E(\mathbf{S}) = A, \quad AB, \quad C, \quad D, \quad AABD$.

Ziv and Lempel defined complexity c as the number of symbols in $E(S)$, which fulfils (2).

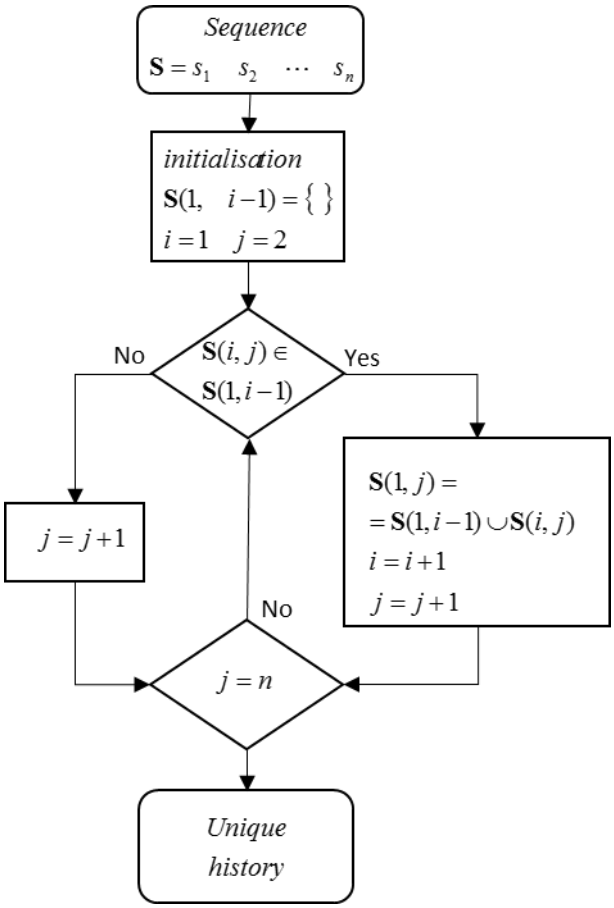


Fig. 1. Flow chart of the algorithm LZ 76

Let's observe the following examples.

$\mathbf{A} = A \quad A \quad C \quad G \quad T \quad A \quad C \quad C \quad A \quad T \quad T \quad G$

$\mathbf{B} = C \quad T \quad A \quad G \quad G \quad G \quad A \quad C \quad T \quad T \quad A \quad T$

$\mathbf{C} = A \quad C \quad G \quad G \quad T \quad C \quad A \quad C \quad C \quad A \quad A$

Unique histories of these sequences are:

$E(\mathbf{A}) = A, \quad AC, \quad G, \quad T, \quad ACC, \quad AT, \quad TG$

$E(\mathbf{B}) = C, \quad T, \quad A, \quad G, \quad GGA, \quad CTT, \quad AT$

$E(\mathbf{C}) = A, \quad C, \quad G, \quad GT, \quad CA, \quad CC, \quad AA$

We can see, that: $c(\mathbf{A}) = c(\mathbf{B}) = c(\mathbf{C}) = 7$

The complexity could be applied also on concatenation of two sequences in accordance with the algorithm in Fig. 2.

For example unique histories of pairs **AC** and **BC** are:

$E(\mathbf{AC}) = A, AC, G, T, ACC, AT, TG, ACGG, TC, AA$

$E(\mathbf{BC}) = C, T, A, G, GGA, CTT, AT, ACG, GT, CA, CC, AA$

$c(\mathbf{AC}) = 10 \quad c(\mathbf{BC}) = 12$

From this example one can conclude, that the sequence **C** is more similar to **A** than **B** is to **A** because:

$c(\mathbf{AC}) < c(\mathbf{BC})$

However in this case the length of the sequences **A** and **B** is approximately equal. Via experiments it could be shown that the validity of it is also dependent on the length of the inputted sequences.

Therefore in our method we will use following normalized quantities in order to estimate the similarity of two sequences:

$$d_1(\mathbf{S}_1\mathbf{S}_2) = \max\{c(\mathbf{S}_1\mathbf{S}_2) - c(\mathbf{S}_1), c(\mathbf{S}_1\mathbf{S}_2) - c(\mathbf{S}_2)\} \tag{3}$$

$$d_2(\mathbf{S}_1\mathbf{S}_2) = \frac{\max\{c(\mathbf{S}_1\mathbf{S}_2) - c(\mathbf{S}_1), c(\mathbf{S}_1\mathbf{S}_2) - c(\mathbf{S}_2)\}}{\max\{c(\mathbf{S}_1), c(\mathbf{S}_2)\}} \tag{4}$$

$$d_3(\mathbf{S}_1\mathbf{S}_2) = c(\mathbf{S}_1\mathbf{S}_2) - c(\mathbf{S}_1) + c(\mathbf{S}_1\mathbf{S}_2) - c(\mathbf{S}_2) \tag{5}$$

$$d_4(\mathbf{S}_1\mathbf{S}_2) = \frac{c(\mathbf{S}_1\mathbf{S}_2) - c(\mathbf{S}_1) + c(\mathbf{S}_1\mathbf{S}_2) - c(\mathbf{S}_2)}{c(\mathbf{S}_1\mathbf{S}_2)} \tag{6}$$

B. Method for authorship identification

In our method two text strings are used as input. First from a set of possible known author and the second from an unknown, which we would like to identify. The pairs of strings are inputted into a software realized in accordance with the flow chart depicted in Fig. 2. The output are normalized complexity measures given by (3) – (6). The first authorship detection attempt is based on normalized measures $d_2(\mathbf{S}_1\mathbf{S}_2)$ and $d_4(\mathbf{S}_1\mathbf{S}_2)$. If these parameters are minimal for the same author in the analyzed pairs, than the text is attributed to this author. In case of ambiguity also the other normalized measures $d_1(\mathbf{S}_1\mathbf{S}_2)$, $d_3(\mathbf{S}_1\mathbf{S}_2)$ are taken into account. In case that the ambiguity is not resolved than their complexity without normalization is also used as a decision factor. In order to do so, unique histories will be computed for all reference text

sequences and also for the tested ones using the algorithm in Fig. 1.

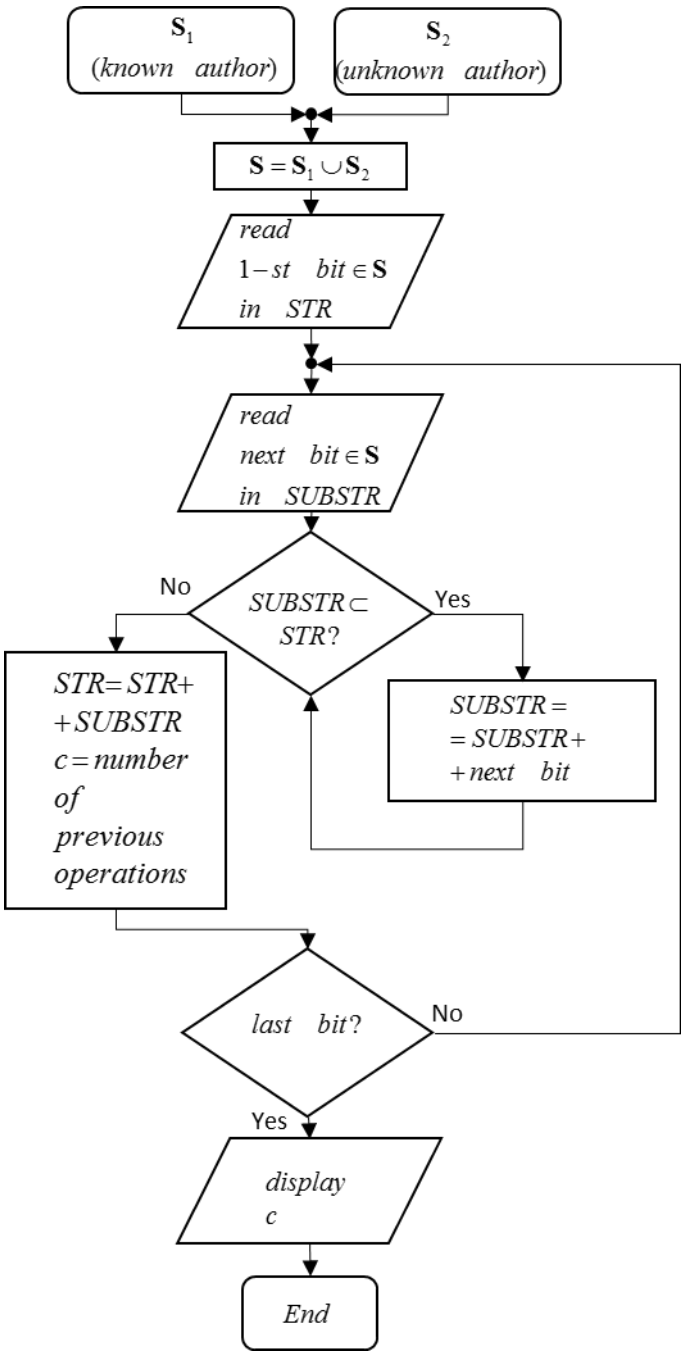


Fig. 2. Flow chart of the algorithm for a pair of sequences

III. METHOD VERIFICATION BY EXPERIMENTS WITH REAL TEXT

In this chapter experiments results are presented obtained using the method on real text strings, namely sonnets and The Federalist Papers.

A. Sonets

Sonnets are ideal for verification of a new authorship detection methods because their text is strongly influenced by the style of the author. In case the new methods are not successful with sonnets, than the probability is low, that they will be appropriate for other text strings.

In the first set of experiments we used one selected sonnet from Shakespeare as a reference string and two sets of sonnets, first one containing the same sonnet and 9 other sonnets from Shakespeare and the second one containing 10 sonnets from different randomly selected authors. The proposed method was implemented in MATLAB code. The results of this experiments are in Tab. 1

TABLE I. RESULTS OF EXPERIMENTS WITH SONNETS FROM SHAKESPEARE VERSUS ONE SONNET FROM SHAKESPEARE

Sonnet No.	Shakespeare		
	Complexity	$d_2(S_1, S_1)$	$d_4(S_1, S_1)$
1	***203***	***0.005***	***0.0099***
2	***360***	***0.756***	***0.8694***
3	***354***	***0.7574***	***0.8616***
4	***359***	***0.7633***	***0.8611***
5	***363***	***0.7559***	***0.8678***
6	***348***	***0.7723***	***0.8678***
7	***363***	***0.7703***	***0.8623***
8	***369***	***0.7840***	***0.8618***
9	***351***	***0.7673***	***0.8661***
10	***356***	***0.7624***	***0.8455***

TABLE II. RESULTS OF EXPERIMENTS WITH SONNETS FROM DIFFERENT AUTHORS VERSUS ONE SONNET FROM SHAKESPEARE

Sonnet No.	Different authors		
	Complexity	$d_2(S_1, S_1)$	$d_4(S_1, S_1)$
1	***374***	***0.8***	***0.885***
2	***370***	***0.7706***	***0.8757***
3	***373***	***0.7703***	***0.882***
4	***373***	***0.8143***	***0.8954***
5	***372***	***0.8134***	***0.8925***
6	***370***	***0.8038***	***0.8811***
7	***368***	***0.846***	***0.8835***
8	***369***	***0.7804***	***0.8835***
9	***366***	***0.7736***	***0.8907***
10	***364***	***0.802***	***0.8901***

The results of the experiments are favorable for validation of the proposed method. It could be observed that with one exception all sonnets from Shakespeare have lover complexity and normalized parameters $d_2(\mathbf{S}_1\mathbf{S}_2)$ and $d_4(\mathbf{S}_1\mathbf{S}_2)$, than the sonnets from other authors. One exception is the sonnet no. 8 from Shakespeare. The probable reason for it is that the reference sonnet is to short and not enough similar expressions (substrings) was found in order to make the method more selective. Therefore in following set of experiments we augmented the reference string by concatenating more sonnets

from Shakespeare. The corresponding results of experiments are in Tab. III. And Tab. IV.

Therefore we made another set of experiments with Shakespeare sonnets in which the reference string was composed of 2, 3 and 4 different concatenated sonnets. The results are given in Tab. III. And Tab. IV. Only $d_2(\mathbf{S}_1\mathbf{S}_2)$ is given there as it represents the founding, that using longer reference texts can improve the authorship detection also for the proposed method. One can observe that for the longest reference text string, there is no exception in correct assignment as was the case by previous experiments.

TABLE III. RESULTS OF EXPERIMENTS WITH SONNETS FROM SHAKESPEARE VERSUS MORE CONCATENATED SONNETS FROM SHAKESPEARE

Sonnet No.	$d_2(S_1, S_1)$ for sonnets from Shakespeare		
	2 sonnets in reference	*3 sonnets in reference*	*4 sonnets in reference*
1	***0.05***	***0.598***	***0.7752***
2	***0.756***	***0.8603***	***0.9150***
3	***0.7574***	***0.8643***	***0.8616***
4	***0.7633***	***0.8663***	***0.9116***
5	***0.7559***	***0.8543***	***0.9083***
6	***0.7723***	***0.8703***	***0.9150***
7	***0.7703***	***0.8683***	***0.9183***
8	***0.7840***	***0.8523***	***0.9060***

TABLE IV. RESULTS OF EXPERIMENTS WITH SONNETS FROM OTHER AUTHORS VERSUS MORE CONCATENATED SONNETS FROM SHAKESPEARE

Sonnet No.	$d_2(S_1, S_1)$ for sonnets from different authors		
	2 sonnets in reference	*3 sonnets in reference*	*4 sonnets in reference*
1	***0.8***	***0.8643***	***0.9083***
2	***0.7706***	***0.8723***	***0.9183***
3	***0.7574***	***0.8643***	***0.8616***
4	***0.8143***	***0.8982***	***0.9284***
5	***0.8134***	***0.8922***	***0.9340***
6	***0.8038***	***0.8663***	***0.9196***
7	***0.846***	***0.8723***	***0.9265***
8	***0.7804***	***0.8822***	***0.9239***

We can conclude that the experiments with sonnets support the hypothesis, that the proposed method could be used for author detection and that longer reference text strings could improve its performance as was expected.

B. The Federalist Papers

In order to make the proposed method verification by other than text strings of artists, in this chapter The Federalist Papers are used in similar experiments as in previous subchapter. These is text, which advocates the ratification of United States Constitution. There are altogether 85 Federalist Papers from 3 authors, namely John Jay, James Madison and Alexander Hamilton which published them under the same pseudonym. They were subject to authorship detection in past under usage of different methods. For papers no. 49, no. 50, no. 51, no. 52, no. 53, no. 54, no. 55, no. 56, no. 57, no. 58, no. 62 and no. 63. The detection of the author was most difficult. At the end it was discovered, that the author of all these texts is Madison. The results of the experiments are in Tab. V. and Tab. VI.

TABLE V. RESULTS OF EXPERIMENTS WITH THE FEDERALIST PAPERS FROM HAMILTON VERSUS REFERENCE STRING HAMILTON

No.	Hamilton				
	c	d_1	d_2	d_3	d_4
49	***25251***	***23390***	***0.9673***	***24462***	***0.9688***
50	***24959***	***23578***	***0.9751***	***24358***	***0.9759***
51	***25397***	***23322***	***0.9646***	***24540***	***0.9663***
52	***25396***	***23383***	***0.9671***	***24600***	***0.9687***
53	***25615***	***23275***	***0.9626***	***24711***	***0.9647***
54	***25463***	***23384***	***0.9671***	***24668***	***0.9688***
55	***25504***	***23332***	***0.9650***	***24657***	***0.9668***
56	***25194***	***23460***	***0.9703***	***24475***	***0.9715***
57	***25608***	***23259***	***0.9620***	***24688***	***0.9641***
58	***25547***	***23287***	***0.9631***	***24655***	***0.9651***
62	***25706***	***23079***	***0.9545***	***24606***	***0.9572***
63	***26162***	***22934***	***0.9485***	***24917***	***0.9524***

TABLE VI. RESULTS OF EXPERIMENTS WITH THE FEDERALIST PAPERS FROM MADISON VERSUS REFERENCE STRING HAMILTON

No.	Madison				
	c	d_1	d_2	d_3	d_4
49	***25009***	***23112***	***0.9632***	***24126***	***0.9647***
50	***24753***	***23352***	***0.9732***	***24110***	***0.9740***
51	***25156***	***23056***	***0.9609***	***24217***	***0.9627***
52	***25135***	***23068***	***0.9614***	***24208***	***0.9631***
53	***25341***	***22993***	***0.9582***	***24339***	***0.9605***
54	***25193***	***23069***	***0.9614***	***24267***	***0.9632***
55	***25275***	***23062***	***0.9611***	***24342***	***0.9631***
56	***24930***	***23184***	***0.9662***	***24119***	***0.9675***
57	***25382***	***23030***	***0.9598***	***24417***	***0.9620***
58	***25305***	***23001***	***0.9586***	***24311***	***0.9607***
62	***25511***	***22904***	***0.9545***	***24420***	***0.9572***
63	***25953***	***22732***	***0.9474***	***24690***	***0.9513***

From the results in Tab. V. and Tab VI. It could be seen that the correct author Madison of papers no.: 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 63 will be detected. Only the paper no. 62 will be attributed to Hamilton. The success rate of these experiments is more than 90%. This also confirms that the proposed method could be used for authorship detection.

IV. CONCLUSIONS

In this manuscript a very simple method for authorship detection was presented. Results of experiments verified its applicability, which makes it easy to be used by broader community of non-experts.

ACKNOWLEDGMENT

This work was supported by EU RTD Framework Program under ICT COST Action IC 1104 *Random Network Coding*

and Designs over GF(q) and by Visegrad Fund and National Scientific Council of Taiwan under IVF–NSC, Taiwan Joint Research Projects Program application no. 21280013 *The Smoke in the Chimney - An Intelligent Sensor - based TeleCare Solution for Homes* and by Scientific Grant Agency of Ministry of Education of Slovak Republic and Slovak Academy of Sciences under contract VEGA 1/0518/13, by Slovak Research and Development Agency under contracts SK-AT-0020-12 and SK-PT-0014-12.

REFERENCES

[1] I. Tokárová, L. Meluch, P. Farkaš, E. Ružický, Forensic author identification with help of compaction algorithm, (In Slovak), Slaboproudý obzor, ISSN 2336-5773, vol. 71, no. 3, pp. 1-5. Oct. 2015

[2] A. N. Kolmogorov On Tables of Random Numbers. In Sankhya: The Indian Journal of Statistics. Vol.25, no. 4, pp. 369-376, 1963

[3] A. Lempel, J. Ziv , . On the Complexity of Finite Sequences. In IEEE Transactions on Information Theory, vol. 22, no. 1, pp. 75-81, 1976.

Privacy Protection Performance of De-identified Face Images with and without Background

Zongji Sun, Li Meng* and Aladdin Ariyaeeinia
School of Engineering and Technology
University of Hertfordshire
Hatfield, AL10 9AB, UK
Email: {z.sun3, L.1.MENG*, a.m.ariyaeeinia}@herts.ac.uk

Xiaodong Duan, Zheng-Hua Tan
Department of Electronic Systems
Aalborg University, Denmark
Email: {xd, zt}@es.aau.dk

***Abstract*—This paper presents an approach to blending a de-identified face region with its original background, for the purpose of completing the process of face de-identification. The re-identification risk of the de-identified FERET face images has been evaluated for the *k*-Diff-furthest face de-identification method, using several face recognition benchmark methods including PCA, LBP, HOG and LPQ. The experimental results show that the *k*-Diff-furthest face de-identification delivers high privacy protection within the face region while blending the de-identified face region with its original background may significantly increases the re-identification risk, indicating that de-identification must also be applied to image areas beyond the face region.**

Keywords—face de-identification; privacy protection; face re-identification; seamless cloning

I. INTRODUCTION

With the advance of sensor technology and the reduced cost of data storage, a huge amount of personal data are being collected at an increasing pace due to their economic and social values. Face biometric, one of the easiest biometrics to collect, is widely used in various application scenarios. Its popularity has boosted due to the recent breakthrough in pattern recognition using deep learning. The latest benchmark results on the LFW (Labelled Faces in the Wild) dataset showed that with a massive training set the performance of face recognition based on a deep convolutional neural network (CNN) successfully achieved a pair-wise verification accuracy of 99.77% [1]. Zhang et al. [2] showed that the person recognition rate could achieve 83.05% on a photo album, where only half of the images in the album contain a frontal face.

However, such data often contain sensitive personal information, which has inevitably raised privacy concerns and been exploited in privacy attacks. The 1995 Data Protection Directive of the European Union (Directive 95/46/EC) [3] demands the deployment of appropriate technical and organizational measures to protect private information in the course of transferring or processing such data. This legal requirement along with ethical responsibilities has attracted intensive research effort in the field of de-identification [4], [5]. The goal of de-identification is twofold: privacy protection as well as data utility preservation. In such way, a de-identified dataset can be used as a replacement of its original dataset to facilitate further data analysis.

Masking, blurring and pixilation are the ad hoc face de-identification methods that were widely used in the past. However these methods do not achieve their goal of privacy protection as the blurring process is revisable while the identity in a pixelated or masked face image can be 100% recognised by parrot recognition [4]. Furthermore, all of these ad hoc methods are destructive and destroy data utility.

The *k*-Same-Pixel/Eigen face de-identification method was the first successful face de-identification method in terms of privacy protection. The *k*-Same methods guarantee a re-identification risk lower than 1/*k*. However, the original *k*-Same-Pixel/Eigen method has some limitations, including ghost artefacts in the de-identified image, *k*-dependant performance, loss of data utility and loss of data diversity, etc. Some of these limitations have been successfully addressed. For example, *k*-Same-M [5] face de-identification eliminated ghost artefacts by aligning faces in the Active Appearance Models. To further improve the privacy protection performance, *k*-Same-furthest [6] and *k*-Diff-furthest [7] face de-identification methods were proposed by the authors and their privacy protection performance with some moderate datasets were presented. This paper presents the work that has been carried out to close the loop of face de-identification, where the de-identified face region is blended back with the background of the original face image. Furthermore, the privacy protection performance of the previously proposed *k*-Diff-furthest methods has been evaluated with the benchmark dataset FERET, before and after background blending.

The rest of the paper is structured as follows. Section II explains the main techniques used in this work. Section III describes our work on closing the loop of face de-identification. Section IV evaluates the privacy protection performance of the proposed face de-identification methods. Finally, Section V summarises the key findings and proposes future work.

II. SUBJECT REVIEW

A. Face Modelling

Active Appearance Model (AAM) is a statistical approach to object modelling. In AAMs, each face image region is represented as a combination of a face shape model and a face texture model. The face shape is defined by the locations of a set of pre-defined facial landmarks and the shapeless texture is generated by warping the original face onto a common shape. An AAM is established through training, where Principal

Component Analysis (PCA) is applied to both the shapes and the textures of the faces in the training set for the purpose of dimensionality reduction. As a result, only the first n shape PCA eigenvectors $\{s_1, \dots, s_n\}$ and the first m texture PCA eigenvectors $\{A_1, \dots, A_m\}$ are used by the trained AAM to represent a given face image. For a given face image, the trained AAM generates its shape as $s = s_0 + \sum_{i=1}^{n} p_i s_i$, and its texture as $A = A_0 + \sum_{i=1}^{m} \lambda_i A_i$, where s_0 and A_0 are respectively the average shape and texture over the training set, shape parameters $\mathbf{p} \in \mathbb{R}^n$ and texture parameters $\boldsymbol{\lambda} \in \mathbb{R}^m$. Typically, when shape and texture parameters are combined together, a scaling factors is applied to one of the two parameter sets, e.g. in the format of $\{W_s p_1, \dots, W_s p_n, \lambda_1, \dots, \lambda_m\}$, to make sure they share a common range. AAMs have been used in face de-identification to not only extra shape and texture features from given face images but also to synthesise new face images. Some face features are not identity related and but will affect the visual quality of the de-identified images in forms of ghost artefacts (e.g. glasses, beard and moustache) or unnatural skin tone (e.g. illumination condition). To make sure good visual quality of the de-identified face images, some non-identity related factors are removed in this work by excluding face images with such non-identity features from the AAM training set. More details of this is given in Section III with example result images.

B. The k-Diff-furthest Face De-identification Method

The *k*-Same methods (*k*-Same-Pixel, *k*-Same-M, *k*-Same-furthest, etc.) all performs face de-identification by applying microaggregation to the original face images, where the set of original faces are divided into clusters of *k* and, for each cluster formed, all the original faces in the cluster are de-identified with the average of that cluster. As each de-identified face image appears in the de-identified image set *k* times and it can be matched, at best, with one of its *k* original faces, all *k*-Same method can guarantee a re-identification risk lower than 1/*k* for their de-identified faces.

To further reduce the re-identification risk, the *k*-Same-furthest method was proposed. It minimises the association between subjects and their face data by introducing data swapping after microaggragation. In each iteration of de-identification, *k*-Same-furthest forms two clusters, each of size *k,* among the remaining original faces and ensures the distance of the two formed clusters is maximised. Unlike *k*-Same-Pixel and *k*-Same-M, the *k*-Same-furthest method de-identifies the faces in one cluster using the average of the other cluster rather than its own. A similar clustering approach has been adopted in the MDAV-generic to form clusters of size *k* [8]. As stated in [8], optimal clustering is NP-hard. The clustering approach of [8] as well as the *k*-Same-furthest method is near-optimal. The main differences between the *k*-Same-furthest and the MDAV-generic method are: 1) MDAV-generic always uses the average of the remaining data to form the next pair of clusters while *k*-Same-furthest forms the next cluster pair based on a randomly selected face. As a result, any two runs of the *k*-Same-further method could hardly produce identical results. 2) Overlapping between each pair of clusters is prevented in *k*-Same-furthest to minimise re-identification risk (mathematical proof of this has been given in [7]). And 3) *k*-Same-furthest applies data swapping between each pair of clusters while MDAV-generic as a microaggregation method, it does not adopt data swapping.

In a set of *k*-Same de-identified face images, there are at least *k* subjects sharing the same de-identified face, making it impossible to distinguish let alone track individuals. In information theory, entropy of a data set is defined as the negative of the logarithm of its probability distribution. Repetition of data in *k*-Same face de-identification inevitably leads to entropy loss or information loss. Let $\mathbf{F}_{\text{orig}} = \{\Gamma_1, \dots, \Gamma_M\}$ be a set of *M* face images and be person-specific (i.e. one image per person). Let $\mathbf{F}_{deid}$ denote the de-identified version of $\mathbf{F}_{orig}$. The entropy of $\mathbf{F}_{orig}$ and $\mathbf{F}_{deid}$ is given in (1) and (2), respectively.

$$H(\mathbf{F}_{\text{orig}}) = -\sum_{i=1}^{M} \mathrm{P}(\Gamma_i) \log_2 \mathrm{P}(\Gamma_i) = -\log_2 \frac{1}{M} \quad (1)$$

$$H(\mathbf{F}_{deid}) = -\sum_{i=1}^{\left|\frac{M}{k}\right|} P(\Gamma_{di}) \log_2 P(\Gamma_{di}) \leq -\log_2 \frac{k}{M}$$

$$= -\log_2 \frac{1}{M} - \log_2 k = H(\mathbf{F}_{orig}) - \log_2 k \quad (2)$$

It is easy to see that $H(\mathbf{F}_{orig}) > H(\mathbf{F}_{deid})$ as $k > 1$, meaning that information has been lost through the microaggreation process in *k*-Same face de-identification. This information loss is due to a decrease in data diversity but it is the mechanism adopted in all *k*-Same face de-identification methods for achieving a guaranteed re-identification risk lower than $1/k$.

To prevent information loss, we have maintained data diversity in the de-identified face set and proposed the *k*-Diff-furthest face de-identification method [7]. Like the *k*-Same-furthest method, *k*-Diff-furthest forms two non-overlapping clusters in each iteration. However, it swaps the cluster centres without aggregating the faces in each cluster. There are two contributions of *k*-Diff-furthest. Firstly, the entropy of the de-identified dataset is the same as that of the original dataset, which means each subject gets a unique de-identified face. The privacy protection performance of the proposed *k*-Diff-furthest method is evaluated in Section IV where the re-identification risk of the *k*-Diff-furthest de-identified faces is always near-zero.

C. Re-identification Risk of the De-identified Faces

Three types of re-identification attack can be used to test the protection performance of the de-identification results [4]. The first type of attack is termed naïve recognition, in which the original face images are used as the gallery and de-identified face images as the probes. The second type of attack is termed reverse recognition, in which the de-identified face images are used as the gallery and the original face images as the probes. The third type of attack is termed parrot recognition. The word 'parrot' means that the de-identification technique is duplicated by the attacker and the set of de-identified face images generated by the attacker is used as probes to match with the published version of the de-identified face image set. As mentioned in Section II.B, in each iteration of *k*-Same/Diff-furthest face de-identification two clusters are formed based on a randomly selected face. This means it is highly unlikely to repeat the same random selections and produce the same set of de-identified faces. In other words, parrot recognition does not work on either *k*-Same-furthest or *k*-Diff-furthest method.

TABLE I. KEY PARAMETERS OF THE FACE RECOGNITION METHODS USED IN THE EVELUATION EXPERIMENTS

Feature	*Parameter values*	*Distance measurement*
PCA	–	Euclidean distance
LBP	*radius* = 1, *neighbours* = 8	Chi-squared distance
HOG	*cell* = 10 × 10, *orientations* = 16	Cosine distance
LPQ	*cell* = 10 × 10	Cosine distance

In our work, AAM is used as the face descriptor. The proposed face de-identification methods form clusters based on Euclidean distance in the AAM feature space. Experimental results from our previous work [6], [9] showed that the re-identification risk is near zero if the attacker uses AAM face representation and matches faces based on Euclidean distance as well. To fully evaluate the privacy protection performance of our proposed methods, further evaluation experiments have been conducted which used various face representation models and distance measures, including Eigenface (PCA) [10], Local Binary Patterns (LBPs) [11], Histogram of Oriented Gradient (HOG) [12] and Local Phase Quantization (LPQ) features [13]. TABLE I summarises the key parameters of these face recognition benchmark methods. In addition, *k*-Nearest Neighbours method has been used to find the top match; and the dimension of both HOG features and LPQ features were reduced to 500 by applying PCA.

III. BLENDING THE DE-IDENTIFIED FACE WITH ITS ORIGINAL IMAGE BACKGROUND

So far, all the published face de-identification methods focus on the isolated face region in the original images. The result images presented in the publications are composed of a de-identified face region and a blank background. The second column of Fig. 1 show the examples. As shown in the examples, the face region exclude hair, the ears, the forehead, the neck, the rest of the human body and the shooting environment. However, real life applications always prefer a face with a background. As stated, the background here means the rest of the human body and the shooting environment that are presented in the original image. In the case of face de-identification, this leads to the demand of blending the de-identified face region back onto its original image background.

One of the main challenges in this task is given by the noticeable differences between the original and the de-identified faces in terms of skin tone, illumination, direction of lighting, etc. Previous research in the field of face swapping and image editing has investigated similar problems and has provided several useful solutions to this challenge. The study on face swapping in [14] used one recolouring method followed by one relighting method to adjust the skin tone. Impett et al. [15] used histogram matching in the RGB space to allow real-time operation. As real-time operation is not a priority for our work at this stage, the more powerful but more time-consuming method of Poisson seamless cloning [16] has been used to achieve a better visual quality of the blended images.

As shown in Fig. 1, the face de-identification process changes not only the texture of a face but also the shape of the face. As a result, the shape of the new face might not fit within the original image background or the new shape may be too small to cover the area of the original face completely. This means a simple replacement would not generate satisfying results. One approach to this challenge could be warping the new face texture to the original face shape, where the shape of the original face would be recovered in the de-identified image and hence the face would fit perfectly with the original image background. However, the shape of a face contains rich personal identifiable information [17]. Bringing back the original face shape would significantly degrade the privacy protection performance of a face de-identification system. Therefore, the new shape of the de-identified face must be maintained after being blended with the original background.

To maintain the new shape of the de-identified face region, a different approach has been taken in this work, where the background of the original face image is deformed to make just enough room to fit the new shape of the face region. The deformation of the background is achieved using Moving Least Squares [18] by solving the best affine transformation $l_v(x)$ that minimizes $\sum_i w_i |l_v(p_i) - q_i|^2$ where $\{p_i\}$ is a set of original points and $\{q_i\}$ is the target deformed positions of $\{p_i\}$. In our work, $\{p_i\}$ and $\{q_i\}$ each are the set of facial contour landmarks as defined in [19]. $\{p_i\}$ are the contour landmarks of the original face in the original image. $\{q_i\}$ are the aligned landmark positions of the de-identified face region. In this work, the de-identified face region is aligned to the original face region using inner corners of the eyes and tip of the nose.

The employment of background deformation has two contributions to a face de-identification system. The first contribution is a better visual quality of the de-identified face image. Fig. 1 shows the blending results of two de-identified faces without and with background deformation. As shown in the second image of Fig. 1 (c), unexpected white spots has appeared on the sides of the face as this de-identified face is wider than its original face. In Poisson blending, the new colours (or the new skin tone) of the de-identified face is calculated pixel by pixel based on the colours of the neighbourhood pixels in the destination image (i.e. the original face image in our case). When the de-identified face region is more narrow than its original face region (as in the first row of Fig. 1), for the pixels on either side of the de-identified face, their neighbourhood pixels from the destination (original) image are pixels within the original face region carrying the original skin tone. Therefore, after Poisson blending, the side of the de-identified face would be colour-matched to the skin tone of the original face and hence the rest of the body (e.g. the ears, the forehead and the neck). The same applies to the inner pixels of the de-identified face (e.g. the redness on the cheeks). In contrast, when a de-identified face is wider than its original face (as in the second row of Fig. 1), for the pixels on either side of the de-identified face, their neighbourhood pixels from the destination (original) image are pixels outside the original face region carrying colours from the original background. As the original face has a white background, without background deformation white spots has appeared on both sides of the de-identified face after Poisson blending (as shown in the second image of Fig. 1 (c)).

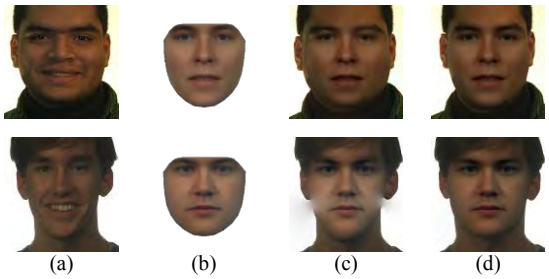


Fig. 1. Blending results of de-identified face regions. (a) Original face images. (b) De-identified face region. (c) De-identified face image without background deformation. (d) De-identified face image with background deformation. Without background deformation, original face shape of (a) is restored in (c); while the shape of (b) is kept in (d) through background deformation.

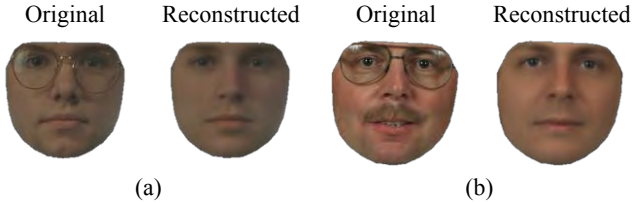


Fig. 2. Examples of reconstructed faces from face features

IV. Evaluation of Re-identification Risk

This section evaluates the re-identification risk of the de-identified face images generated by the *k*-Diff-furthest method.

A. Dataset

A subset of the FERET face dataset [20] containing 963 subjects has been used in the experiments. This subset was chosen from the available images of 994 subjects to ensure that each subject has two colour frontal face images ('fa' and 'fb'). All the 'fa' faces were used as the gallery in the re-identification tests to match against either the original version or the de-identified version of the 'fb' faces. Fig. 5 shows the original 'fa' (the first row), the original 'fb' (the second row) and the de-identified 'fb' (the last row) images for five subjects from the chosen FERET subset. As each pair of 'fa' and 'fb' images were taken at the same shooting session, they present a high degree of similarity and hence a challenge to our *k*-Diff-furthest face de-identification method.

B. Training of the AAM Face Representation Model

In this study, all the *k*-Diff-furthest de-identified faces were generated with the same AAM, which was trained on a subset of the FERET dataset. The AAM training set contains 1952 colour images of frontal faces taken from the FERET dataset and all of them are without glasses, beard or moustache. The exclusion of such features has enabled the automatic removal of such non-identity related features through AAM representation, as shown by the examples in Fig. 2. As a result, no faded (or averaged out) glasses frame, beard or moustache would appear on the resulting de-identified face images to degrade the visual quality of the image. In addition, it was observed that the most significant texture feature in our trained AAM describes merely the lighting condition of the images. To calibrate the lighting condition, the most significant texture feature is always set to its mean value (over the training set) for all the images represented in our face model.

C. Re-identification Tests

Before calculating face features with its face representation model, all face recognition systems isolate the face region from the background through cropping. Some systems crop the face region with a rectangular box while some systems define the face region with some facial landmarks (e.g. those on the eyebrows and the jawline as in our *k*-Diff-furthest method). In our re-identification tests, the face images were first aligned to the mean shape of the training set (approximately 200 × 200 pixels) and then cropped by a rectangular box co-centred with the mean shape. The size of the cropping box is either 200 × 200 or 300 × 300 pixels. The cropped images can be with or without a background. Fig. 3 shows the face images used in these re-identification tests for an example subject.

TABLE III presents the re-identification risk of the *k*-Diff-furthest method measured in naïve recognition attacks against four face recognition methods. All the recognition rates in TABLE III are averages over 10 runs and are presented in the format of accuracy ± standard deviation. TABLE III (a) shows that without merging with the original background the de-identified face regions generated by our *k*-diff-furthest method always present a near zero re-identification risk for all the benchmark face recognition methods tested.

TABLAE III (b) shows the re-identification risk increases when the de-identified face region is blended with its original background. This is mainly due to the fact that the Passion blending process has brought the skin-tone and illumination of the original face image back to the de-identified face image. Furthermore, the more the original background in a cropped face image, the higher the re-identification risk will be. The results are under 7% and still acceptable when the face images were cropped with a 200 × 200 square. However, the level of re-identification risk became unacceptably high when the face images were cropped with a 300 × 300 square, indicating that background areas around the face region may also contain personal identifiable information and de-identification must also be applied to these image areas to achieve complete privacy protection. Further experiments have been conducted to evaluate the re-identification risk presented by the background area alone. Results of these experiments are presented and discussed in the next sub-section.

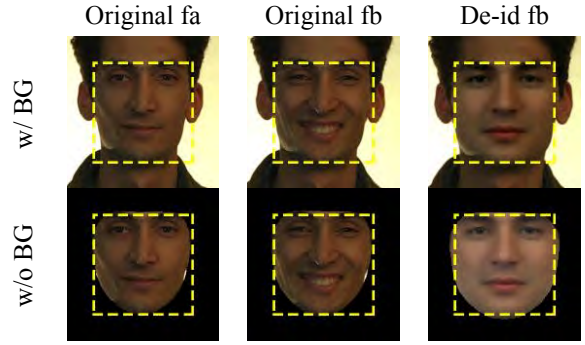


Fig. 3. Example of faces of one person with 200 pixels square (inside yellow boxes) and 300 pixels square cropping

TABLE IV shows the results of reverse recognition attacks, which are similar to those of naïve recognition attacks. All the de-identified face images yield a near zero re-identification risk when the background is excluded from the image but a much higher risk when the background is included.

D. Background Attack to Face De-identification System

As shown by TABLE III and TABLE IV the background area of an original face image presents personal identifiable information and can increase the re-identification risk when being blended with the de-identified face region. Here the background area may contain not only the background environment of the original image but also the hairstyle, the ear, the neck and the dressing style presented in the original image.

This experiment aimed to investigate the possibility of using merely the background area of the original face image to attack the face de-identification system. Fig. 4 shows some example images used in this experiment. Background attack is a generic attack to any face de-identification method that modifies the face region only. The idea and implementation of this attack is basic. All the face images were applied with an inverse crop based on their facial landmarks, so that only the image area outside the face region was kept. Inverse crop has been used in face recognition to compare the recognition performance between human and computer [21], [22]. After inverse crop, the images were cropped into the size of 300 × 300 for face recognition. The experimental results are shown in TABLE II Comparing with the results shown in TABLE III, it is clear that the background area of the original image was the main contributor to the increase in the re-identification risk.

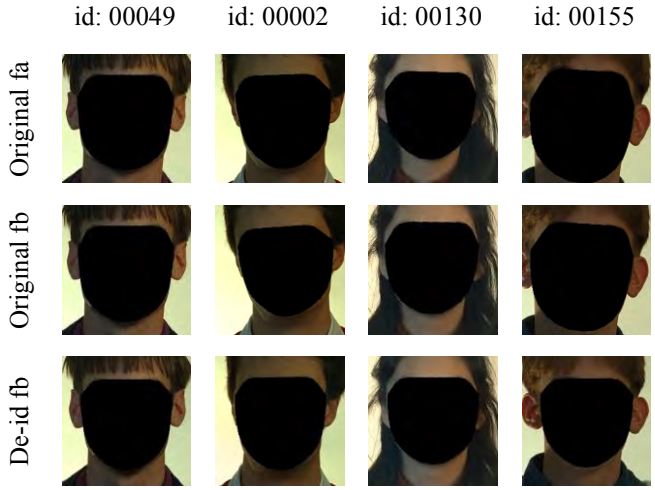


Fig. 4. Examples of inverse crop face images used in background attack experiment

TABLE II. RE-IDENTIFICATION RISK (%) OF INVERSE CROPPED FACE IMAGES

Method	*300 × 300 inverse crop*	
	Original	*De-id*
PCA	56.39	31.82 ± 0.89
LBP	**78.19**	**55.62 ± 0.68**
HOG	53.27	27.07 ± 1.09
LPQ	60.44	32.88 ± 1.06

V. CONCLUSIONS

The experiment results show that the *k*-Diff-furthest face de-identification method provides high privacy protection within the face region. However, blending the de-identified face to its original background increases the re-identification risk. Although a face recognition software focuses on the cropped face region, but information contained in the background area around the face region (e.g. hair colour, hairstyle, and dressing style) can also be used to identify a person. Our experiment of a background attack confirmed that face region is sufficient but not necessary for identifying a person. To protect privacy of the individuals captured in an image/video, de-identification must be applied to not only the face region but also all the image regions that contain personal identifiable information.

REFERENCES

[1] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang, "Targeting Ultimate Accuracy: Face Recognition via Deep Embedding," Jun. 2015.

[2] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev, "Beyond Frontal Faces : Improving Person Recognition Using Multiple Cues," in Computer Vision and Pattern Recognition (CVPR), 2015.

[3] European Parliament, "Directive 95/46/EC," Off. J. Eur. Communities, vol. L281, p. 31, 1995.

[4] E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," IEEE Trans. Knowl. Data Eng., vol. 17, no. 2, pp. 232–243, Feb. 2005.

[5] R. Gross, L. Sweeney, F. de la Torre, and S. Baker, "Model-Based Face De-Identification," in 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), 2006, pp. 161–161.

[6] L. Meng and Z. Sun, "Face De-identification with perfect privacy protection," in 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014, pp. 1234–1239.

[7] Z. Sun, L. Meng, and A. Ariyaeeinia, "Distinguishable de-identified faces," in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, vol. 04, pp. 1–6.

[8] J. Domingo-Ferrer and V. Torra, "Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation," Data Min. Knowl. Discov., vol. 11, no. 2, pp. 195–212, Sep. 2005.

[9] L. Meng, Z. Sun, A. Ariyaeeinia, and K. L. Bennett, "Retaining expressions on de-identified faces," in 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014, pp. 1252–1257.

[10] M. Turk and A. Pentland, "Eigenfaces for Recognition," J. Cogn. Neurosci., vol. 3, pp. 71–86, 1991.

[11] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[12] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, vol. 1, pp. 886–893.

[13] V. Ojansivu and J. Heikkilä, "Blur Insensitive Texture Classification Using Local Phase Quantization," in Image and Signal Processing, vol. 5099, A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 236–243.

[14] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: automatically replacing faces in photographs," ACM Trans. Graph., vol. 27, no. 3, pp. 39:1–39:8, Aug. 2008.

[15] L. Impett, P. Robinson, and T. Baltrusaitis, "A facial affect mapping engine," in Proceedings of the companion publication of the 19th international conference on Intelligent User Interfaces - IUI Companion '14, 2014, pp. 33–36.

[16] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," ACM Trans. Graph., vol. 22, no. 3, p. 313, Jul. 2003.

[17] M. Smiatacz, "Face Recognition: Shape versus Texture," in Advances in Intelligent Systems and Computing, vol. 313, R. S. Choraś, Ed. Cham: Springer International Publishing, 2015, pp. 211–218.

[18] S. Schaefer, T. McPhail, and J. Warren, "Image deformation using moving least squares," ACM Trans. Graph., vol. 25, no. 3, p. 533, Jul. 2006.

[19] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge," in 2013 IEEE International Conference on Computer Vision Workshops, 2013, pp. 397–403.

[20] P. J. Phillips, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 10, pp. 1090–1104, 2000.

[21] P. J. Phillips and A. J. O'Toole, "Comparison of human and computer performance across face recognition experiments," Image Vis. Comput., vol. 32, no. 1, pp. 74–85, Jan. 2014.

[22] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, "Describable Visual Attributes for Face Verification and Image Search.," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 10, pp. 1962–77, Oct. 2011.

TABLE III. NAIVE RECOGNITION RATES (%) OF ORIGINAL 'FB' FACES AND DE-IDENTIFIED 'FB' FACES AGAINST ORIGINAL 'FA'

Method	*Without background*				*With background*			
	200 × 200		*300 × 300*		*200 × 200*		*300 × 300*	
	Original	*De-id*	*Original*	*De-id*	*Original*	*De-id*	*Original*	*De-id*
PCA	47.25	0.13 ± 0.16	42.37	0.07 ± 0.09	54.83	4.10 ± 0.58	61.27	39.13 ± 0.55
LBP	**74.25**	0.11 ± 0.10	**63.03**	0.13 ± 0.11	**83.39**	1.30 ± 0.24	**87.23**	55.12 ± 0.56
HOG	47.14	0.21 ± 0.14	18.38	0.17 ± 0.20	74.87	**6.09 ± 0.48**	78.09	56.93 ± 0.70
LPQ	53.27	**0.25 ± 0.18**	47.04	**0.25 ± 0.11**	80.48	4.42 ± 0.61	82.66	**59.14 ± 0.81**
	(a)				(b)			

TABLE IV. REVERSE RECOGNITION RATES (%) OF ORIGINAL 'FB' FACES AND DE-IDENTIFIED 'FB' FACES AGAINST ORIGINAL 'FA'

Method	*Without background*				*With background*			
	200 × 200		*300 × 300*		*200 × 200*		*300 × 300*	
	Original	*De-id*	*Original*	*De-id*	*Original*	*De-id*	*Original*	*De-id*
PCA	45.38	**0.33 ± 0.19**	41.53	**0.30 ± 0.21**	53.48	**9.41 ± 0.53**	61.37	49.06 ± 0.61
LBP	**72.59**	0.23 ± 0.14	**63.14**	0.21 ± 0.11	**80.48**	1.59 ± 0.41	**86.60**	59.28 ± 1.63
HOG	46.52	0.24 ± 0.16	17.03	0.07 ± 0.09	74.77	5.92 ± 0.70	76.95	**59.81 ± 0.68**
LPQ	50.88	0.23 ± 0.15	46.31	0.20 ± 0.14	80.06	4.35 ± 0.58	82.87	58.36 ± 0.48
	(a)				(b)			

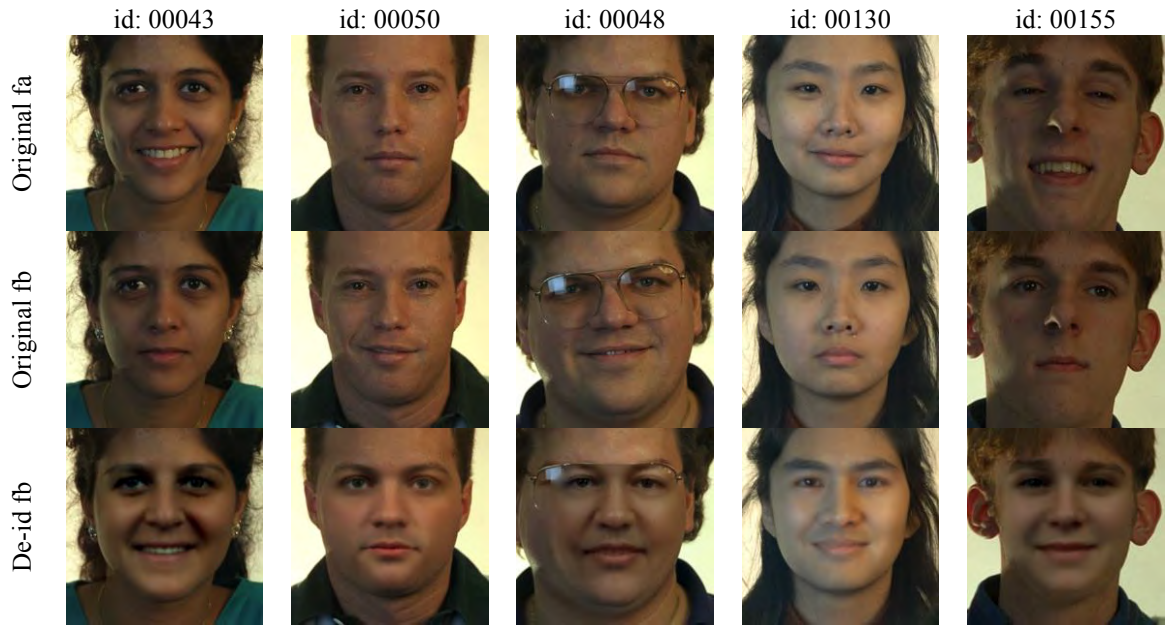


Fig. 5. Examples of the FERET face images used in the re-identification tests

Deep Metric Learning for Person Re-Identification and De-Identification

Ivan Filković, Zoran Kalafatić and Tomislav Hrkać
University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{ivan.filkovic, zoran.kalafatic, tomislav.hrkac}@fer.hr

***Abstract*—Large amounts of visual data are gathered from various surveillance systems across different places and times, and have to be processed in order to infer the current state of the world. One of the common problems in surveillance scenarios is person re-identification, the task of associating a person across different cameras. On the other hand, these scenarios raise privacy concerns, which lead to the need for person de-identification, i.e. concealing person identity. This task is related to the re-identification in two aspects: (i) multiple appearances of the same person could be de-identified in similar manner; and (ii) if we discover the features useful for re-identification, we could try to hide the identity by modifying those features. Re-identification can be addressed as a classification problem. The state-of-the-art classification methods are based on deep learning. In this paper we explore the applicability of the recently proposed *Triplet network* architecture to the person re-identification problem, by applying it on *VIPeR* dataset. We show that the network is able to learn useful feature-space embeddings, and analyze its benefits and limitations.**

I. Introduction

Large amounts of visual data from various surveillance systems across different places and time have to be processed so that conclusions about current state of the world could be obtained. This capability is now widely exploited in a variety of scenarios to obtain recordings of people, either for immediate inspection or for storage and subsequent analysis and sharing. One of the common problems in that kind of scenarios is person re-identification. Person re-identification is a task of associating persons across different cameras. Due to the diversity of real world cameras (different resolutions and aspect rations, mono/color) as well as their non-standardized placement resulting in various person perspectives (Figure 1), person re-identification is a challenging problem. Furthermore, humans are articulated objects and their pose can significantly change. Also, the usual problems in computer vision like changing lighting and weather conditions, partial occlusions and motion-blur make this problem even more difficult. Therefore, it is not surprising that person re-identification is getting significant attention in research community [1], [2], [3], [4], [5], [6], [7].

A related problem is person de-identification, the process of concealing the identity of persons recorded in surveillance video, motivated by the increasing privacy concerns. In general, privacy can be defined as the ability of an individual or group to have their personal information and affairs secluded from others, and to disclose them as they choose. Privacy concerns exist wherever personal data or identifiable information is collected and stored. Personal data mean any information related to a person that can be identified, directly or indirectly, with reference either to an identification number or to one or more factors specific to his/her physical, physiological, habitual, mental, economic, cultural or social identity. The protection of private data is required by European Union Data Protection Directive (95/46/EC).



Fig. 1. Examples of pedestrian images from *VIPeR* dataset (Section III-A). The same pedestrians are in succession of two. Different person perspectives, changing lighting and weather condition, various pose changes and small examples resolution (48×128 pixels) exemplify how hard is the task of person re-identification. For the last three image pairs it is nearly impossible even for a human to identify them as a examples of a same person.

De-identification in multimedia content is usually defined as the process of hiding identities of individuals captured in a dataset (images, video, audio, text). De-identification in images typically consists of the following phases: detection and tracking of persons, resulting in bounding boxes around persons; segmentation of relevant visual features from background in order to locate them more precisely; and applying a de-identifying transformation [8]. This task is related to the re-identification in two aspects: (i) multiple appearances of the same person could be de-identified in similar manner; and (ii) discovering the features useful for re-identification could help the de-identification. If we find and study the cues and/or features appropriate for robust and accurate person re-identification, we could use them in de-identification methods, by tracking and masking them so that the person's identity can be preserved.

The goal of this work is to test the *Triplet network* model capability of learning applicable person embeddings in the feature space, i.e. to test if model can implicitly learn visual soft-biometric features such as person gait, weight, height, gender, age, ethnicity, hair style, facial hair style, glasses, tattoos, scars, birth marks, dressing style and accessories (backpacks, suitcases). We approach the problem of person re-identification as a classification, where each person constitutes its own class (Fig. 3, 4, 5).

The remainder of this paper is organized as follows. In Section II an overview of previous work in person re-identification and deep metric learning is given. Publicly available person re-identification Viewpoint Invariant Pedestrian Recognition

(*VIPeR*) dataset is outlined in Subsection III-A. The used model for deep metric learning based on *Triplet network* is presented in the same section. The conducted experiments are described and the achieved qualitative and quantitative results are presented in Section IV. In the end we discuss the results and provide some conclusions about the proposed setup and future goals.

II. Related Work

Machine learning methods, including metric learning, can be divided into standard supervised and unsupervised categories. Supervised learning aims at minimizing the within-class and maximizing the between-class variance with respect to the labeled training data. On the other hand, unsupervised metric learning methods try to find a lower-dimensional vector space where maximum information is preserved. Common supervised methods are: Linear discriminant analysis (LDA) [9], Regularized linear discriminant analysis (R-LDA) [10], Representation based linear projection (SRLP) [11], Iterative nearest neighbors-based linear projections (INNLP) [12], Discriminative deep metric learning [13] and Adaptive density discrimination method [14]. Principal component analysis (PCA) [15] and deep Autoencoders [16] are examples of unsupervised learning methods.

In the person re-identification, most methods that tackle the given problem, can be divided into two representative classes. The first class focuses on learning representative example embeddings in feature space [4], [17], [18], [19], [20]. Learned representations are usually more robust and discriminative than standard Euclidean distance or its generalization Mahlanobis distance. The second class focuses on finding robust features that can successfully deal with camera problems, various person viewpoints, human pose changes, changing lighting and weather conditions and partial person occlusions [21], [22], [23], [24], [5], [6]. Huang et al. [4] proposed nonlinear local metric learning method (NLML) where the final distance metric is a linear combination of local metrics. Koestinger et al. [18] introduced distance metric from equivalence constraints based on a statistical inference perspective. Hirzer et al. [19] learned a metric from pairs of samples from different cameras. Earlier approaches ignored the transition from one camera to the other. Liu et al. [21] proposed an unsupervised approach for learning a bottom-up feature importance. Ma et al. [22] aggregated local descriptors with Fisher vectors to produce a global representation of the image.

III. Methodology

A. Viewpoint Invariant Pedestrian Recognition (VIPeR) Dataset

The Viewpoint Invariant Pedestrian Recognition (*VIPeR*) dataset [7], [25] contains images of 632 persons where each person is captured two times with different cameras i.e. there are two examples per person. Each example is scaled to 48×128 resolution. The dataset was created in a fairly controlled academic setting over the course of a few months. The *VIPeR* dataset is very difficult for the task of person re-identification because of strongly varied perspectives between two examples of the same person, excessive lighting changes in the scene, small resolution of example images and finally because of just two examples per person (Figure 1).

B. Model

For our deep metric embedding we use *Triplet network* model, proposed by Hoffer and Ailon [26], aiming to learn usable image representations by comparing distances between positive and negative examples. It embeds the data into a vector space such that the distance between two positive examples becomes larger than the distance between the reference positive example and a negative example. A *triplet network* consists of three convolutional neural network instances of the same architecture (Table I) which share all the network parameters.

The *Triplet network* model [26] builds on the work of Bromley et al. [27] and Chopra et al. [28] i.e. on the *Siamese network* model. The main difference is that *Siamese network* consists of two identical convolutional networks and receives a pair of images belonging to the same or different classes, whereas the *Triplet network* learns by receiving triplets of images, two of which belong to the same class and the third to a different class. The triplet input setup improves the robustness in the case of small between-class variation (such as in person re-identification task).

A triplet consisting of positive (x^+), reference (x) and negative (x^-) example is set as one input (x^+, x, x^-) to *Triplet network* (Figure 2). First, the sampling layer is responsible for randomly choosing x^+, x and x^- examples in each epoch. Then, each example from the input triplet is forward-passed through one of the three networks in the model to get a low-dimensional vector.

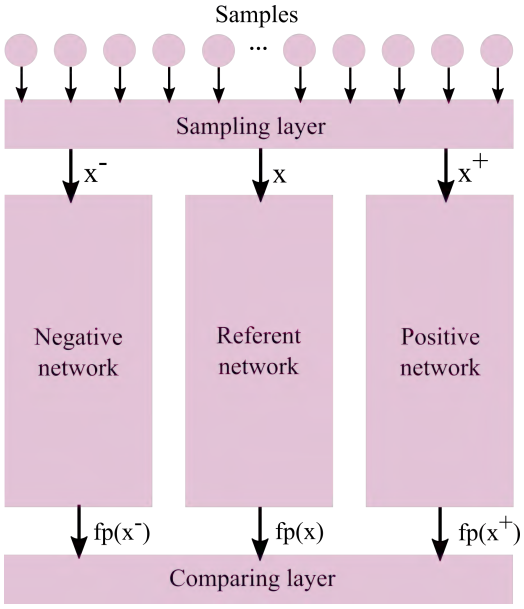


Fig. 2. The flow chart diagram of the *Triplet network* training process. The input samples are images of persons. The sampling layer is responsible for sampling reference, positive and negative samples from the dataset. Each of the triplet elements is forward-passed ($fp(\cdot)$) through the corresponding network (all the networks share same parameters). The resulting vectors are compared in the last layer.

The resulting vectors are compared in the last layer (Eq. 1), by applying the Soft-max function σ (Eq. 2) to the corresponding L_2 distances $d_+ = \|fp(x) - fp(x^+)\|_2$ and $d_- = \|fp(x) - fp(x^-)\|_2$. The fp stands for the forward pass

through the network.

$$C = \begin{cases} 1, & \text{if } \sigma(d_+) < \sigma(d_-) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The goal of the training is to satisfy the condition C given by Eq. 1 for all training examples. The distance between the reference example should be (in the embedded space) closer to the positive than to the negative example. As the loss function in the training, the standard mean squared error on the Softmax difference ($\sigma(d_+) - \sigma(d_-)$) is used.

$$\sigma(\|fp(x) - fp(x')\|_2) = \frac{e^{\|fp(x)-fp(x')\|_2}}{\sum\limits_{x^* \in (x^-, x^+)} e^{\|fp(x)-fp(x^*)\|_2}}, \quad (2)$$

where x' represents positive or negative example depending on the distance to be computed.

The architecture of a network instance used as a component in our *Triplet network* is specified in table I. Each of the three networks consists of five convolutional layers, five max pooling layers and nonlinearities (Rectified Linear Unit - ReLU [29]) between them. The number of maps, kernel sizes and neurons are listed for each layer. During training, we use dropout [30] (randomly turning off neurons), which is a form of model averaging. This enables us to get some of the benefits of a multiple network model while having a single network. The network output is a 128 dimensional vector embedded in feature space. The presented architecture is similar to the *Triplet network* proposed by Hoffer and Ailon [26] with main differences in convolutional kernel sizes to accommodate different input image size (48×128) so that the final dimension of feature vector remains the same.

Layer	type	dimensions	properties
0	input	3 (RGB) maps of 48x128	
1	convolutional	64 maps of 44x122 neurons	kernel 5x7 stride 1
2	ReLU & max pooling	64 maps of 22x61 neurons	kernel 2x2 stride 2
3	dropout		dropout ratio = 0.25
4	convolutional	128 maps of 20x56 neurons	kernel 3x6 stride 1
5	ReLU & max pooling	128 maps of 10x28 neurons	kernel 2x2 stride 2
6	dropout		dropout ratio = 0.25
7	convolutional	256 maps of 8x24 neurons	kernel 3x5 stride 1
8	ReLU & max pooling	256 maps of 4x12 neurons	kernel 2x2 stride 2
9	dropout		dropout ratio = 0.25
10	convolutional	256 maps of 4x10 neurons	kernel 1x3 stride 1
11	ReLU & max pooling	256 maps of 2x5 neurons	kernel 2x2 stride 2
12	convolutional	128 maps of 2x3 neurons	kernel 1x3 stride 1
13	ReLU & max pooling	128 maps of 1x1 neurons	kernel 2x2 stride 2

TABLE I. DESCRIPTION OF ONE NETWORK ARCHITECTURE INSTANCE USED IN *Triplet network* MODEL.

For the network evaluation, we use the following error metric on train, validation and test sets:

$$error = \frac{1}{M * N} \sum_{i=1}^{i=M} \sum_{j=1}^{j=N} C_{ij} \quad (3)$$

Here M represents the number of batches and N is the number of triplets $(x_{ij}^+, x_{ij}, x_{ij}^-)$ in a given batch, while C_{ij} for a particular triplet is defined in Eq. 1.

IV. EXPERIMENTS

We approach the problem of person re-identification as a classification, where each person constitutes its own class, and where the error is defined over triplet comparisons (Equation 3). For preliminary qualitative results we show the embedded examples (feature vector dimensionality of 128) projected onto two main components of Principal component analysis (PCA) [15].

In Figures 3, 4 and 5 train, validation and test examples, respectively, are presented in the two dimensional PCA vector space. The associated examples (presented in the same color in the same figure) represent a specific person. Please note that different figures show different sets of persons, although we reuse the same colors. Qualitative analysis shows that the projected examples of the same person are usually close to each other. There are some obvious outliers and those represent hard examples in *VIPeR* dataset (Figure 1).

We decided to conduct our experiment on really challenging *VIPeR* dataset because we wanted to approach real world conditions as much as possible. The goal was to analyze how will the small number of example per person affect person re-identification score when metric embedding is done by *Triplet network* [26]. For high speed training and evaluation we used the *Torch7* framework [31], using its CUDA and cuDNN implementation on a GeForce GTX 970. Training a *Triplet network* model takes about 16 minutes, without taking into account the time it takes to test at regular intervals to keep track of training and validation error.

In training process Stochastic gradient descent (SGD) as a learning algorithm, and standard negative log-likelihood as a loss function are utilized. The training takes 22 epochs (each epoch contains 6400 batches) and early stop is imposed if validation error increased 7 times consecutively. The values of parameters are as follows: learning rate is set to 0.1, learning rate decays for 10^{-6} after each epoch, L2 penalty on the weights is set to 10^{-4}, momentum is 0.9 and batch size (number of triplets in batch) is set to 128.

For evaluation, standard experimental protocol was used. *The VIPeR* dataset which has 632 persons was divided in train (64%; 404 persons), validation (16%; 101 persons) and test (20%; 127 persons) subsets, where special attention was given to ensure that examples of the same person go to the same subset, i.e. randomized stratified sampling was used. After the last training epoch of the used model the error on the train set was 2.453% (Figure 3), validation error was 8.125% (Figure 4) and test error was 13.125% (Figure 5).

We can conclude that deep metric learning conducted with *Triplet network* is promising direction when applied to person re-identification and de-identification tasks. It is important to observe that *VIPeR* dataset has just two examples per person and with that notion, the obtained results are acceptable. With more examples per person it is reasonable to expect that deep metric models would give even better results and implicitly learned soft-biometrics features could be more emphasized.



Fig. 3. Train examples (404 pairs) from *VIPeR* dataset projected onto two main components of PCA. Associated examples i.e. ones color coded with same color represent the specific person, where one example is captured with *Camera A* and other one with *Camera B*. When qualitatively analyzed, it can be seen that examples from the same person are close to each other in two dimensional vector space. Also, some obvious outliers can be noticed. N.B. In train, validation and test subset figures, same color does not mean the same person.



Fig. 4. Examples from validation subset (101 pairs) of *VIPeR* dataset.



Fig. 5. Examples from test subset (127 pairs) of *VIPeR* dataset.

V. Conclusion

Person re-identification and de-identification are daunting tasks that need to be properly addressed: person re-identification for security reasons and person de-identification because of privacy concerns. The large amounts of visual data from various surveillance systems lead this areas rapidly forward and in the future we expect significant progress.

In this paper we propose and discuss applying deep metric learning onto person re- and de-identification tasks. We show promising results on challenging *VIPeR* dataset when *Triplet network* as deep metric model is used. Triplet comparison error on test set of 13.125% shows that applicable person embeddings in the feature space based on soft-biometric features can be successfully learned even when there is small number of examples per person.

In the future, we would like to meticulously investigate which soft-biometric features such as person gait, weight, height, gender, age, ethnicity, hair style, facial hair style, glasses, tattoos, scars, birth marks, dressing style and accessories (backpacks, suitcases) have the most significant impact on metric learning performance. Also, more rigorous experiments on diverse person re-identification datasets are planned to have more profound analysis of this daunting task.

Acknowledgments

This work has been supported by the Croatian Science Foundation under the project UIP-11-2013-1544.

References

[1] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person re-identification*. Springer, 2014, vol. 1.

[2] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[3] X.-Y. Jing, X. Zhu, F. Wu, X. You, Q. Liu, D. Yue, R. Hu, and B. Xu, "Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[4] S. Huang, J. Lu, J. Zhou, and A. K. Jain, "Nonlinear local metric learning for person re-identification," *CoRR*, vol. abs/1511.05169, 2015.

[5] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2360–2367.

[6] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking." in *BMVC*, vol. 2, no. 5, 2010, p. 6.

[7] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro*, 2007.

[8] P. Agrawal and P. J. Narayanan, "Person de-identification in videos," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 21, no. 3, pp. 299–310, 2011.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley Interscience, 2000.

[10] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American statistical association*, vol. 84, no. 405, pp. 165–175, 1989.

[11] R. Timofte and L. J. V. Gool, "Sparse representation based projections," in *British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings*, 2011, pp. 1–12.

[12] ——, "Iterative nearest neighbors for classification and dimensionality reduction," in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, 2012, pp. 2456–2463.

[13] J. Hu, J. Lu, and Y. Tan, "Discriminative deep metric learning for face verification in the wild," in *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, 2014, pp. 1875–1882.

[14] O. Rippel, M. Paluri, P. Dollár, and L. D. Bourdev, "Metric learning with adaptive density discrimination," *CoRR*, vol. abs/1511.05939, 2015.

[15] I. Jolliffe, *Principal component analysis*. Wiley Online Library, 2002.

[16] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA, July 2, 2011*, 2012, pp. 37–50.

[17] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 1–16.

[18] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2288–2295.

[19] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 780–793.

[20] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*. IEEE, 2011, pp. 649–656.

[21] C. Liu, S. Gong, C. C. Loy, and X. Lin, "Person re-identification: What features are important?" in *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer, 2012, pp. 391–401.

[22] B. Ma, Y. Su, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer, 2012, pp. 413–422.

[23] M. Bauml and R. Stiefelhagen, "Evaluation of local features for person re-identification in image sequences," in *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*. IEEE, 2011, pp. 291–296.

[24] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*. IEEE, 2010, pp. 435–440.

[25] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I*, 2008, pp. 262–275.

[26] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Similarity-Based Pattern Recognition - Third International Workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015, Proceedings*, 2015, pp. 84–92.

[27] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," in *Advances in Neural Information Processing Systems 6, [7th NIPS Conference, Denver, Colorado, USA, 1993]*, 1993, pp. 737–744.

[28] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, 2005, pp. 539–546.

[29] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, 2010, pp. 807–814.

[30] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.

[31] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A matlab-like environment for machine learning," in *BigLearn, NIPS Workshop*, 2011.

Deformable Part-based Robust Face Detection under Occlusion by Using Face Decomposition into Face Components

Darijan Marčetić, Slobodan Ribarić
University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia
{darijan.marcetic, slobodan.ribaric}@fer.hr

Abstract - In this paper, we propose modifications of deformable part-based models in order to increase the robustness of face detection under occlusion. The modifications are: i) the tree, representing the deformable part-based model of the frontal face, which is partitioned into 11 subtrees representing face components; ii) the weight of each face component which is obtained based on the results of psychological experiments; iii) the introduction of new scoring functions and thresholds; and iv) a new procedure for robust face detection based on the valuation of scoring functions and thresholds. The experiment was performed only for frontal face images, and thus this work is used only as a proof of concept. Based on the encouraging experimental results, we conclude that the proposed method is suitable for extension to detect faces with different poses under occlusions.

Keywords: Robust Face Detection, Occlusion, Deformable Part-based Model, Face Components

I. INTRODUCTION

Face detection is the process of finding and localizing faces in an image. Finding and analysing faces in unconstrained real-world images is a very difficult computer vision task. Constraining factors are the ambiguities of the face, such as poses, occlusions, expressions, illumination, scale, colours and textures. Diverse methods have been developed based on scanning window classifiers [1], deep neural networks [2], part-based models [3-6] and regression [7]. Several papers have been published related to face detection under occlusion [8-11], but they are constrained to detect only frontal faces under occlusion. In [8], a support vector machine (SVM) with local kernels was used for robust frontal face detection under partial occlusion. A test was performed on HOIP and CMU face databases. The proposed method was demonstrated for the detection of faces wearing sunglasses or a scarf. It was confirmed that the proposed method was better than a method based on conventional SVM with a global kernel. In [9], a generalization of the work of Viola and Jones for the efficient detection of faces with occlusions was described. The main idea is based on boosting algorithms and cascade structures to efficiently detect faces with occlusions. Several improvements were presented: (i) a robust boosting scheme; (ii) reinforcement training; and (iii) cascading with evidence to extend the system to handle occlusions. Experimental results for detecting frontal faces with artificial occlusions were obtained on CMU and MIT face databases. The authors in [10] divided a face region into multiple patches, and mapped those weak classifiers to the patches. The weak classifiers, inspired by the Viola and Jones approach, belonging to each patch, were combined to create a new classifier to determine if this was a valid face patch without occlusion. All of the valid face patches, with different weights, were used to make a final decision whether the input subwindow was a face. The experimental results showed that the proposed method is promising for the detection of occluded faces. In [11], a novel approach for automatic and robust face detection was presented. It used a component-based approach that combined techniques from both the statistical and structural pattern recognition domain. The method relied on Haar-like features and an AdaBoost trained classifier cascade, and the topology verification was based on graph matching techniques. The system was applied to frontal face detection and the experiments showed improved performance compared to conventional face detection in the presence of partial occlusions, uneven illumination, and out-of-plane rotations, thus yielding higher robustness.

Initially, face detection, pose estimation, and landmark localization were handled separately. These three tasks can be simultaneously handled with a unified part-based approach that uses mixtures of trees with a shared pool of part templates for face landmarks, and utilizing an encoding scheme for modes of elastic deformation, and a view-based topology for face detection and pose estimation [5]. The parameters of part-based models are discriminatively trained in a max-margin framework. Facial landmarks are visible in multiple views and thus the corresponding parts can be shared among different views, which reduces the complexity of the learning procedure for the model. Most part-based models are sensitive to the presence of occlusions because they are learned on non-occluded faces. Facial landmark scores are processed one by one in a cascade, which also decreases robustness.

In this paper, we propose robust frontal face detection obtained by modifying the method presented in [5]. In our approach, a face is decomposed into distinct face components. These face components are selected by psychological experiments related to face recognition by humans [12]. Robustness is achieved by fusion with a

weighted voting scheme where the relative relations among weights are selected based on recommendations given in [12]. In this paper, we present a proof of concept for the detection of frontal faces under occlusions which is suitable for extension to multi-viewpoint occluded face detection.

II. Deformable Part-based Models

Deformable part-based models (DPMs) are one of the most popular face detection, pose estimation and landmark localization methods [5].

The DPM was used to find a visual object in an actual image in [4]. The authors describe the approach of breaking down visual objects into a number of "primitive parts" and specify a permitted range of spatial relations which these primitive parts must satisfy for the object to be present in an image. The object model is represented by primitive parts held together by "springs". The springs joining the primitive parts serve both to constrain relative movement and measure the "cost" of the movement by how much they are "stretched" [3].

The maximal matching between an object model and a visual object in the scene is expressed by the total cost of embedding parts at their locations. For total cost optimization, the authors proposed a method based on dynamic programming.

This initial idea was extended to a mixture of multi-scale deformable part models, represented by HOG features, and methods for discriminative training with partially labelled data [5].

In [6], the authors proposed a method that solved the speed bottleneck of DPMs without impairing the accuracy of detection.

The method proposed in [5] uses a model based on mixtures of trees with a shared pool of parts represented by HOG descriptors. Facial landmarks are modelled as parts. A global mixture of parts is used to capture topological changes due to face viewpoints. A set of 13 trees was used to represent different face viewpoints (Fig. 1). Each tree has only one root node. From 13 trees which represent different face viewpoints, we selected only the tree associated with the frontal face, thus simplifying the model description. This simplification was used for a proof of concept, related to the robust detection of faces under occlusion.

The tree associated with the frontal face $T = (V, E)$ has $|V| = 68$ vertices (or parts) and $|E| = 67$ edges (Fig. 1). Let us write I for an image and $l_i = (x_i, y_i)$ for the pixel location of part V_i. A configuration of parts $L = \{l_i : i = 1,..., 68\}$ is scored as follows [5]:

$$Score(I,L) = App(I,L) + Shape(L) + \alpha \quad (1)$$

$$App(I,L) = \sum_{i=1}^{68} w_i \cdot \phi(I,l_i) \quad (2)$$

$$Shape(L) = \sum_{ij \in E} a_{ij} dx^2 + b_{ij} dx + c_{ij} dy^2 + d_{ij} dy . \quad (3)$$

Equation 2 sums the appearance evidence for placing a HOG-based template w_i for part V_i at location l_i. A HOG-based feature vector $\phi(I,l_i)$ is extracted from pixel location l_i in image I. Equation 3 scores the mixture-specific spatial arrangement of parts L, where $dx = x_i - x_j$ and $dy = y_i - y_j$ are the displacement of the V_i part relative to the V_j part. Each term in the sum can be interpreted as a spring that introduces spatial constraints between a pair of parts, where a, b, c and d represent spring parameters (rest location and rigidity) and were learned on a set of training images [3]. The term α is a scalar bias associated with the front face view.

Inference corresponds to maximizing $Score(I,L)$ in Eq.1 over L:

$$Score^*(I,L) = \max_L [Score(I,L)]. \quad (4)$$

Since $T = (V, E)$ is a tree, the inner maximization can be done efficiently with dynamic programming [3].

Let us denote L for which the $Score$ (I, L) is maximal with L^*, i.e. $Score^*(I,L) = Score(I,L^*)$.

If $Score$ $(I,L^*) \geq \theta_1$, then it is assumed that a face is detected in the image (i.e. locations of facial landmarks are given by L^*). θ_1 is a threshold value applied in [5].

III. Robust Face Detection

In the context of this paper, robustness is related to the ability to detect frontal faces in an image I with different types of occlusions in cases where the original approach [5] fails due to false negative detection (i.e. a face has $Score(I,L^*) < \theta_1$, Fig. 2). In order to increase the robustness of frontal face detection, we modified the original approach described in Section 2 by introducing a new algorithm based on modifications as follows:

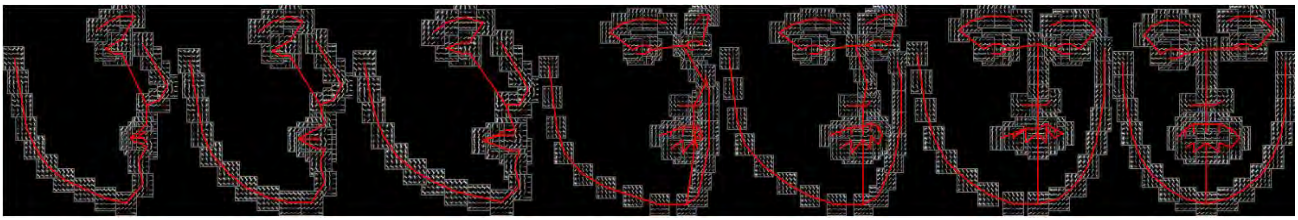


Figure 1. Mixture of trees that encode topological change due to viewpoint [5]. The red lines denote springs between pairs of parts.

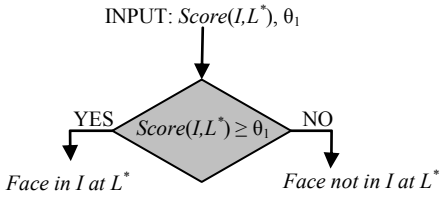


Figure 2. Flowchart of the decision step for the original face detection algorithm [3].

i) The first modification partitions the tree $T = (V, E)$, which corresponds to the frontal face (seventh tree in Fig. 1), into 11 subtrees $T^k = (V^k, E^k)$; $k = 1, 2, ...11$, by removing the edges that connect different face components (Fig. 3). Each subtree is associated with one face component: *left upper edge*, *left lower edge*, *right upper edge*, *right lower edge*, *left side of the lips*, *right side of the lips*, *nose*, *left eye*, *left eyebrow*, *right eye*, *right eyebrow*.

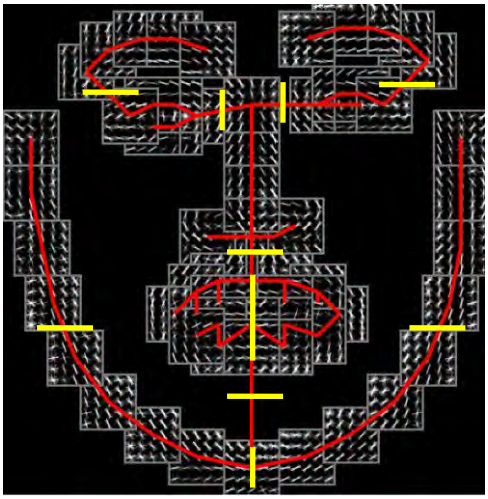


Figure 3. The face region is partitioned into 11 face components containing parts corresponding to the face components. Removed edges are marked with short yellow lines.

ii) The second modification introduces three new thresholds (θ_2, θ_3 and θ_4) and three new scoring functions ($Appearance_k(I,L^*)$, $ScoreAppearance(I,L^*)$ and $ScoreThreshold(I,L^*)$).

The original threshold θ_1 proposed in [5] is -0.65, and the threshold θ_2 is determined based on experiments (described in Section 4). These two thresholds define three intervals for $Score(I,L^*)$. Interpretation of these three intervals is clarified in Fig. 4 b).

The thresholds θ_3 and θ_4 divide the 2-dimensional decision space into face and non-face space (Fig. 4 b)).

Fig. 2 illustrates the original decision step [5]. Fig. 4 a) depicts a flowchart for the final decision step of our frontal face detection algorithm.

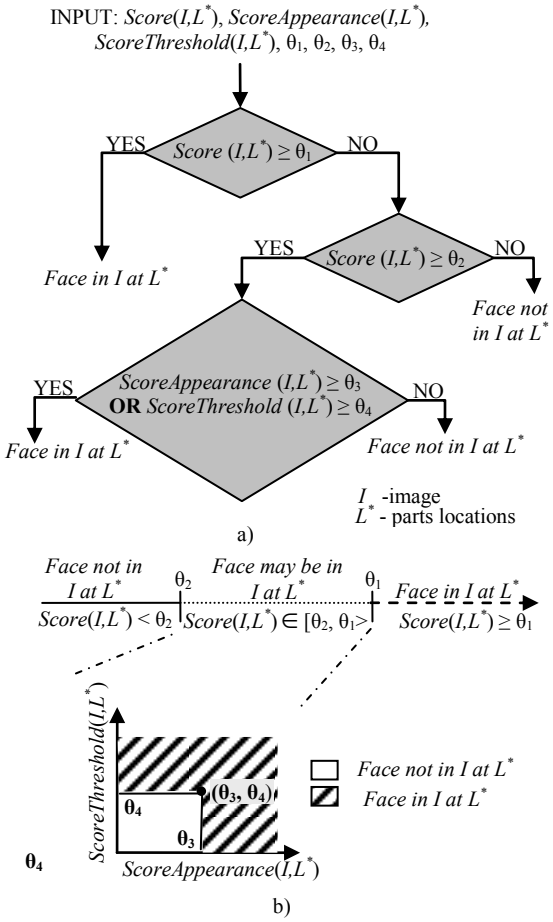


Figure 4. a) Flowchart of the decision step for the proposed face detection algorithm; b) Intervals

A. Final decision step for the face detection algorithm

The final decision step for the face detection algorithm is given as follows:

i) If $Score(I,L^*) \geq \theta_1$, then a face is detected in image I at location L^*.

ii) If $Score(I,L^*) < \theta_2$, then a face is not detected in image I at location L^*.

iii) If $Score(I,L^*)$ falls into the interval $[\theta_2, \theta_1>$, then $ScoreAppearance(I,L^*)$ and $ScoreThreshold(I,L^*)$ are used in conjunction with the newly introduced thresholds θ_3 and θ_4, respectively (Fig. 4).

If $ScoreAppearance(I,L^*) \geq \theta_3$ or $ScoreThreshold(I,L^*) \geq \theta_4$, then there is a face in an image I at location L^* (Fig. 4).

The $ScoreAppearance(I,L^*)$ and $ScoreThreshold(I,L^*)$ are described as follows.

A face appearance score is defined as:

$$ScoreAppearance\left(I, L^*\right) = \sum_{k=1}^{11} \mu_k \cdot Appearance_k\left(I, L^*\right), \quad (5)$$

where μ_k is the weight of a k-th face component selected based on the results of psychological experiments [12] which estimate the importance of the face components in the face recognition process performed by humans. The experimental results [12] indicate the following order of importance of facial components (in decreasing order): eyes, eyebrows, mouth, nose and contour of face (Table I).

TABLE I. THE EXPERIMENTALLY DETERMINED IMPORTANCE OF FACE COMPONENTS IN FACE RECOGNITION BY HUMANS

Face component	Weight μ_k
Left/right upper contour - μ_1/μ_2	1
Left/right lower contour - μ_3/μ_4	1
Left/right side of the lips - μ_5/μ_6	2
Nose - μ_7	1.5
Left/right eye - μ_8/μ_{10}	3
Left/right eyebrow - μ_9/μ_{11}	2.5

For each face component k = 1, 2, ...11, the appearance score is defined as:

$$Appearance_k\left(I, L^*\right) = \sum_i w_i \cdot \phi\left(I, l_i\right), \quad (6)$$

where index i determines the parts V_i from the set of parts V^k associated with a k-th face component, w_i is a HOG-based template for part V_i, and $\phi\left(I, l_i\right)$ represents the HOG-based feature vector extracted from pixel location l_i in an image I. For example, k = 8 corresponds to a left eye (Table I), and there are six parts $V^8 = \{V_{21}, V_{22}, ..., V_{26}\}$.

A face appearance threshold score is defined as:

$$ScoreThreshold\left(I, L^*\right) = \sum_{k=1}^{11} c_k \quad (7)$$

where

$$c_k = \begin{cases} 1 & \text{if } Appearance_k(I, L^*) \geq Threshold_k \\ 0 & \text{if } Appearance_k(I, L^*) < Threshold_k \end{cases} \quad (8)$$

and $Threshold_k$ defines the minimal value of the appearance score of a k-th face component for which it is supposed that this face component is correctly detected. Thresholds $Threshold_k$; k = 1, 2, …, 11 are selected based on the experiments conducted in the XM2VTS [13] and INRIAP [14] databases described in Section 4. The interpretation of $ScoreThreshold\,(I, L^*)$ is as follows. For example, if $ScoreThreshold\,(I, L^*)$ = 5, this means that 5 face components are visible and properly detected, while the remaining six (11 − 5) face components are not detected due to poor quality of an image or occlusion.

The main reason for introducing $ScoreAppearance\,(I, L^*)$ and $ScoreThreshold\,(I, L^*)$ is the fact that in the interval $[\theta_2, \theta_1>$ there are many false positive and false negative faces, so it is used to minimize the total error rate (TER). The thresholds θ_2, θ_3 and θ_4 define the operating point and they are selected to minimize the TER.

IV. EXPERIMENTAL RESULTS

Two databases were used to perform the experiments. From the XM2VTS database [13], 201 images of frontal faces without occlusions were selected (first column of Fig. 5). For each of these 201 images, six new images were created for simulating six typical configurations of face occlusions by covering different face regions with black rectangles (from the second to seventh column of Fig. 5).

Additionally, 73 images without faces were selected



Figure 5. Results of face detection for frontal face images from the XM2VTS database with three score values S1 for $Score(I,L^*)$, S2 for $ScoreAppearance(I,L^*)$ and S3 for $ScoreThreshold(I,L^*)$. The first column is faces without occlusions and the other columns represent faces with different modes of occlusion. For example, for the 5th face in the top row S1 = − 0.69887, which means that the original method fails to detect the face, while S3 = 3, which means that the modified method detects the face.

from the INRIA person database [14] (Fig. 6). In total, 1,480 (i.e. 7 × 201 + 73) images were used for testing. Let us now justify somewhat the unusual configuration of the testing. Typically, some "in the wild" datasets, such as AFW [15], which contain complex scenes with face occlusions, are used for testing face detection performance. All such databases violate our assumption about the presence of only frontal faces. In order to circumvent this shortcoming of the popular datasets used for face detection testing, we were forced to create an ad hoc face dataset just for the purpose of the proof of concept. Because the XM2VTS database contains only face images, the INRIA person database was used to enable an evaluation of false positive face detections.



Figure 6. False detections of faces for an image from the INRIA person database with $Score(I,L^*) > -1$. For each detection frame, the values of $Score(I,L^*)$, $ScoreApperance(I,L^*)$ and $ScoreThreshold(I,L^*)$ are given in the same line, respectively. For example, in the image at the lower left frame, the values are $Score(I,L^*) = -0.89707$, $ScoreApperance\ (I,L^*) = 0.52952$ and $ScoreThreshold(I,L^*) = 1$.

First, we ran the originally proposed algorithm with the predefined threshold $\theta_1 = -0.65$ proposed in [5]. The following results were obtained: 160 frontal faces were not detected (false negatives) among 1,407 faces (Fig 9.a)), and 1 image region was detected as a face false positive among 1,480 images (one image from INRIA had one image region with a $Score\ (I,L^*) > -0.65$, Fig. 7.), so we concluded that the robustness was relatively high compared with the state-of-the-art-method [7]. Among 160 false negatives, there were 145 frontal face images having the type of occlusion depicted in the 5th column in Fig. 5, and the remaining 15 having the type of occlusion depicted in the 2, 3, 6, or 7th column in Fig. 5. Note that there is no false negative detection for 201 images without occlusions, and a single false positive was in one of the 73 images from the INRIA person database (Fig 7.).

For our method, we searched for the values for thresholds θ_2, θ_3, and θ_4 to find an operating point that minimizes the total error rate (TER). The minimum TER = 0.087 was achieved for the following threshold values: $\theta_2 = -0.83$, $\theta_3 = 0.69$, and $\theta_4 = 3$. The threshold θ_1 is assumed to be -0.65, as in the original paper [5]. The values of the thresholds $Threshold_k$; $k = 1, 2, \ldots, 11$ are determined as the value of the first quartile of $Apperance_k(I, L^*)$ obtained on 201 frontal images without occlusions (Fig. 8).



Figure 7. The single false detections of faces for the image from the INRIA person database with $Score(I,L^*) \geq -0.65$.

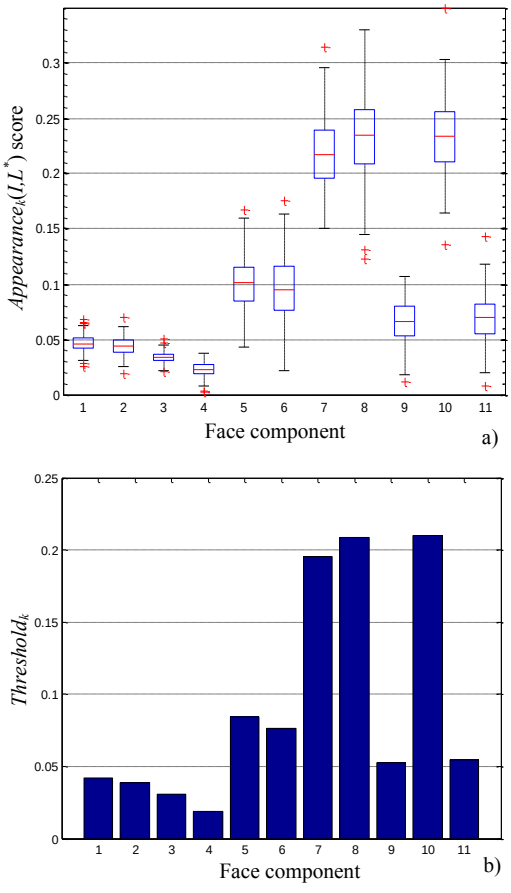


Figure 8. Appearance scores of 11 face components for 201 images of frontal faces without occlusions from the XM2VTS database: a) $Appearance_k(I,L^*)$ distribution, the red line represents the mean value, the box edges the first and the third quintile, the data range is represented with the dashed line, and the red crosses represent the data outliers; b) the values of $Threshold_k$ are obtained as the first quartile from the distribution in a).

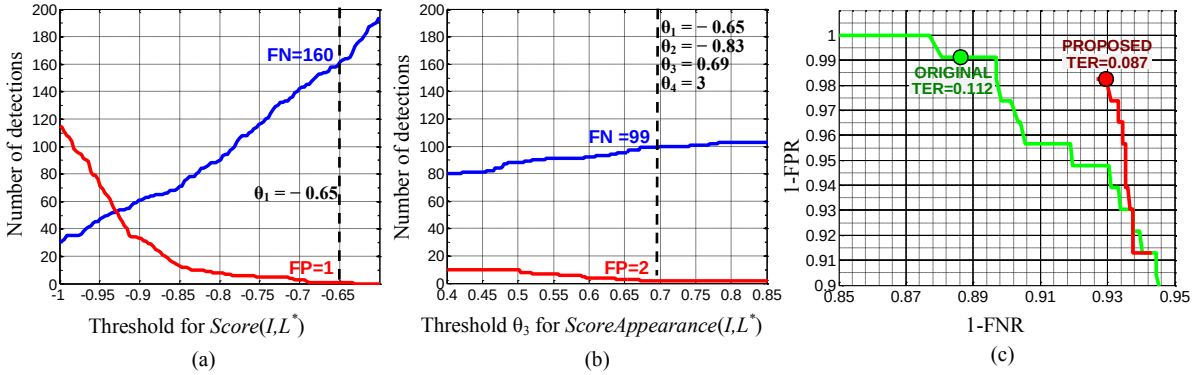


Figure 9. Comparison of the original and the proposed method: a) FN and FP for the original method; b) FP and FN for the proposed method where θ_1 = -0.65, θ_2 =-0.83, and θ_4 = 3; c) 1-FPR and 1-FNR for the original and the proposed method, the green circle represents the working point of the original algorithm for threshold $\theta_1 = -0.65$, and the red circle represents the working point for the proposed algorithm for $\theta_1 = -0.65$, $\theta_2 = -0.83$, $\theta_3 = 0.69$, and $\theta_4 = 3$ for which the TER is minimal.

By using the proposed method and optimal thresholds we achieved a reduction of false negative face detections from 160 to 99, while false positive face detections increased from 1 to 2. The TER was reduced from 0.112 for the original approach to 0.087, which corresponds to an improvement of 22.32% (Fig. 9 c).

V. CONCLUSION

In this paper, we have analysed the influence of face occlusion on the face detection method based on deformable part-based models as originally proposed by X. Zhu and D. Ramanan [5]. We proposed modifications of the method in order to increase the robustness of face detection and to minimize the total error rate. The main modifications are: i) the tree, representing the part-based model of the frontal face, which is partitioned into 11 subtrees representing the face components (left upper edge, left lower edge, right upper edge, right lower edge, left side of the lips, right side of the lips, nose, left eye, left eyebrow, right eye, right eyebrow); ii) the weight of each face component which is obtained based on the results of psychological experiments [12]; iii) the introduction of new scoring functions and thresholds; iv) a new procedure for the robust detection of faces under occlusion based on an evaluation of scoring functions and thresholds.

By using our method on a database consisting of 1,480 images (201 non-occluded face images, 1,206 face images with six types of occlusion and 73 non-face images), the TER was improved by 22.32 % compared to the original method [5]. Note that the experiment was performed only for frontal face images, and thus this work is used only as a proof of concept. Based on the encouraging results, we plan to extend the proposed method by applying the original mixture trees for 13 different face poses and in a total of 146 parts [5].

ACKNOWLEDGMENTS

This work has been supported by the Croatian Science Foundation under project 6733 De-identification for Privacy Protection in Surveillance Systems (DePPSS). It is also the result of activities in COST Action IC1206 "De-identification for Privacy Protection in Multimedia Content".

LITERATURE

[1] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” inProc. IEEE Comput. Soc. Conf. Comput.Vis. Pattern Recog., 2001, pp. I-511–I-518.[

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. “Deepface: Closing the gap to human-level performance in face verification,” In CVPR, 2014, pp. 1701–1708.

[3] M. Fischler and R. Elschlager, “The representation and matching of pictorial structures,” IEEE. Trans. Computers, C-22(1), 1973, pp.67–92.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” IEEE TPAMI, 32(9), 2010, pp. 1627–1645.

[5] X. Zhu and D. Ramanan, “Face detection, pose estimation, and landmark localization in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 2879–2886.

[6] J. Yan, Z. Lei, L. Wen, and S. Z. Li, “The fastest deformable part model for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 2497–2504.

[7] S. Liao, A. K. Jain, S. Z. Li, “A Fast and Accurate Unconstrained Face Detector,” IEEE TPAMI, vol. 38, Issue: 2, 2016, pp. 211–223.

[8] K. Hotta, “A robust face detector under partial occlusion,” in Proc. Int. Conf. Image Process., 2004, pp. 597–600.

[9] Y. Lin, T. Liu, and C. Fuh, “Fast object detection with occlusions,” in Proc. Eur. Conf. Comput. Vis., 2004, pp. 402–413.

[10] J. Chen, S. Shan, S. Yang, X. Chen, and W. Gao, “Modification of the adaboost-based detector for partially occluded faces,” in Proc. 18th Int. Conf. Pattern Recog., 2006, pp. 516–519.

[11] L. Goldmann, U. Monich, and T. Sikora, “Components and their topology for robust face detection in the presence of partial occlusions,” IEEE Trans. Inf. Forensics Security, vol. 2, no. 3, 2007, pp. 559–569.

[12] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russel, “Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About,” Proc. IEEE, vol. 94, no. 11, Nov 2006, pp. 1948–1962.

[13] K. Messer, J. Matas, J. Kittler, J. Luettin and G. Maitre; “XM2VTSbd: The Extended M2VTS Database, Proceedings 2nd Conference on Audio and Video-base Biometric Personal Verification (AVBPA99)” Springer Verlag, New York, 1999.

[14] INRIA Person Dataset, http://pascal.inrialpes.fr/data/human/

[15] The Annotated Faces in the Wild (AFW) testset, https://www.ics.uci.edu/~xzhu/face/

Creating a Face Database for Age Estimation and Classification

Petra Grd
Center for Forensics, Biometrics and Privacy
Faculty of Organization and Informatics
Varaždin, Croatia
petra.grd@foi.hr

Miroslav Bača
Center for Forensics, Biometrics and Privacy
Faculty of Organization and Informatics
Varaždin, Croatia
miroslav.baca@foi.hr

***Abstract* — Age classification and estimation is an important task in classifying face images. One way to classify humans by age is by using two-dimensional images of people's faces. In order to create an accurate algorithm for age classification, appropriate datasets for training and testing are required. A small number of datasets are available for age estimation and classification based on images of human faces. In order to develop a more accurate algorithm, this paper compares existing age estimation databases and describes the process of creating a private database ageCFBP (Center for Forensics, Biometrics and Privacy age database) with facial images useful for age estimation and classification.**

Keywords—age estimation, face database, biometrics, ageCFBP

I. INTRODUCTION

Age estimation can be defined as the determination of the age of the person or his/her age group [1], [2]. Human age classification and estimation can be done in many ways, but this paper is concerned with the age estimation and classification based on two-dimensional images of people's faces. Age estimation and age classification of humans using their face images are a part of a field of biometric. Biometric systems use behavioral and physiological characteristics to recognize individuals. Soft biometric traits like age, gender, ethnicity, height, weight, in combination with hard biometric traits can be used to enhance the performance of biometric systems.

In order to create an accurate algorithm for age classification and estimation, appropriate databases for training and testing are required. There are many available databases, some of them are public and some are private. The problem arises when images of children need to be used for age estimation. Most of the available databases do not offer images of people from birth to adulthood. The Fg-net database is most commonly used in age estimation research [3] and it is described in detail in this paper. In order to develop more accurate algorithms, more databases with facial images need to be created. This paper describes the process of creation of such a database.

II. AVAILABLE DATABASES

There is a large number of databases available for human age estimation research. The problem arises when research involving images of children needs to be done. Then, the number of databases drops significantly.

A. Databases Overview

Currently available databases that can be used for age estimation or classification are: Fg-net Aging Database [4], MORPH Database [5], YGA Database [6], WIT-DB Database [7], AI&R Asian Face Database [8], Burt's Caucasian Face Database [9], LHI Face Database [10], HOIP Face Database [11], Iranian Face Database [12], Gallagher's Web-Collected Database [13], Ni's Web-Collected Database [14] and PAL database [15]. An overview of these databases can be seen in Table I [16].

TABLE I. FACE AGE DATABASES [16]

Database	No. of subjects	No. of images	Age	Race
Fg-net	82	1002	0-69	C
MORPH	13000	55000	16-77	C, N, M
YGA	1600	8000	0-93	M
WIT-DB	5500	12008	3-85	M
AI&R	17	34	22-61	M
BURT'S	147	147	20-62	C
LHI	8000	8000	9-89	M
HOIP	300	306600	15-64	-
IRANIAN	616	3600	2-85	M
GALLAGHER'S	N/A	28231	-	-
NI'S	-	219892	1-80	-
PAL	575	-	18-93	C, N, M

[a] *C - Caucasoid, N - Negroid, M - Mongoloid*

It can be concluded that some of these databases are not appropriate for research on age estimation and some cannot be used for age classification. YGA, WIT-DB, AI&R, and LHI database are databases of only Asian subjects. HOIP database has age labels distributed in 10 groups with five year intervals and Gallagher's Web-collected database has age labels in seven categories, which is not detailed enough. Burt's Caucasian Face Database and MORPH database have subjects

aged respectively, between 20 and 62 and between 16 and 77, which is not the entire age range (children are not included).



Fig. 1. Example of images from Fg-net database

The majority of state of the art research use Fg-net database for algorithm training and testing (Table II). In order to compare the ageCFBP database with Fg-net database, next chapter gives a description of Fg-net database.

TABLE II. FACE AGE DATABASES BY NUMBER OF PAPERS

Database	No. of papers
Fg-net	60
MORPH	16
PAL	9
Private databases	17
Other	20

B. Fg-net database

The Fg-net (Face and Gesture Recognition Research Network) ageing database [4] is a publicly available ageing database that has been extensively used for evaluation by researchers. The database is composed of 1,002 images of 82 subjects. There are 607 images of male subjects and 395 images of female subjects from 34 female subjects and 48 male subjects (Table III) in the age range 0–69 years [17].

TABLE III. SUBJECTS AND IMAGES BY GENDER IN FG-NET DATABASE

	No of subjects	No of images
Female	34	395
Male	48	607
Total	82	1002

The database predominantly includes images in the age range 0–18 years. The images are manually annotated with 68-landmark features and include information on image size, age, gender, presence of facial hair and eyeglasses, horizontal pose, and vertical pose. Many of the images are scanned from photographs and these images exhibit common noise errors associated with old photographs [17] (Fig.1.).

Most subjects provided from 10 to 13 images of themselves, 6 subjects provided less than 10 images, and 20 subjects provided more than 13 images (Table IV).

TABLE IV. NUMBER OF IMAGES PER PERSON IN FG-NET DATABASE

	6	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**18**
Male	1	3	1	7	8	7	11	6	1	3	0
Female	0	0	1	3	7	7	6	3	3	3	1
Total	1	3	2	10	15	14	17	9	4	6	1

Regarding number of images per age, the largest number of images is at the age of 18, the smallest number is of age 46 and above. Some ages don't have images associated with them (age 56, 57 59, 64, 65, 66 and 68) (Table V).

TABLE V. IMAGES PER AGE IN FG-NET DATABASE

Age	No of images	Age	No of images	Age	No of images	Age	No of images
0	43	**16**	37	**32**	4	**48**	3
1	27	**17**	28	**33**	9	**49**	3
2	39	**18**	47	**34**	8	**50**	2
3	42	**19**	23	**35**	11	**51**	3
4	42	**20**	20	**36**	8	**52**	3
5	40	**21**	16	**37**	3	**53**	2
6	41	**22**	17	**38**	5	**54**	2
7	41	**23**	22	**39**	6	**55**	2
8	31	**24**	9	**40**	9	**58**	1
9	25	**25**	17	**41**	6	**60**	1
10	40	**26**	11	**42**	5	**61**	3
11	33	**27**	11	**43**	4	**62**	1
12	37	**28**	12	**44**	4	**63**	1
13	32	**29**	9	**45**	7	**67**	1
14	32	**30**	19	**46**	3	**69**	1
15	30	**31**	6	**47**	2		

III. LEGAL ASPECTS

Since personal data is collected and used, collection and processing of such data must be in accordance with the Croatian Law on Personal Data Protection. Some of the basic concepts recognized by this Law are [18]:

- Personal information is any information relating to an identified natural person or a natural person who can be identified (respondent); a person who can be identified is the person whose identity can be determined directly or indirectly, in particular on the basis of one or more factors specific to his/her physical, physiological, mental, economic, cultural or social identity.
- Personal data processing is any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise made available, alignment or combination, blocking, erasure or destruction, and the

implementation of logical, mathematical and other operations on the data.

- Personal data collection is each set of personal data available by specific criteria, whether centralized, decentralized or dispersed on a functional or geographical basis and irrespective of whether it is contained in computer databases or other technical aids or manually.
- Personal data collection director is natural or legal person, public or other body which determines the purposes and means of processing personal data. Once the purpose and method of treatment is prescribed by law, the same law also determines the personal data collection director.
- User is a natural or legal person, public or other body which may use personal data to perform regular activities within its statutory activity.
- Subjects consent is freely given and specific manifestation of the will of the subjects with which he/she expresses consent to the processing of their personal data for specific purposes.

During data collection, respondent was aware of the fact that he/she is giving personal information, aware of the purpose for which the information is collected, and that they can withdraw their data at any time [16].

Respondent received a Privacy policy for review, which describes the duration of the Privacy policy, all the data collected, the purpose for which the data is collected, and whom to contact in case of any questions. In addition to this



Fig.2. Landmarks annotated using a script

policy, all participants signed a Declaration of conformity, with which they give their consent to use their personal data for the research, and confirm that they received for review the Privacy policy mentioned earlier. Since this research collects personal data of minors, their parents/guardians gave consent for the use of their data [16].

IV. Database Description

As mentioned earlier, the main problem is lack of images of children, which results in lowered accuracy of age estimation and classification algorithms. The ageCFBP database concentrates on collecting images of person's age 0 to 25.

A. Image Acquisition

Images were collected during two years. Subjects were asked to provide images of themselves from childhood to adulthood. Every subject had to provide a minimum of four images. Images of all subjects have been collect according to Privacy policy described earlier, and all subjects signed the Declaration of conformity, with which they stated that their images may be used for the research. The images were acquired by scanning analog images and collecting digital images. A person providing images needed to provide basic information about every image. The basic information collected is shown in Table VI.

TABLE VI. Information collected for the database

Field	Field definition
ID	Identifier of the subject
IM_ID	Image identifier
IM_AGE	Age of subject in image
DOB	Date of birth
DOA	Date of Acquisition
Gender	Male or female

The subjects were asked to provide a minimum of four images. Most of the subjects provided five images, but some of them provided more than eleven. Only six subjects provided a minimum of four images.

B. Demographics

During data collection, images of 287 subjects have been collected, 151 female and 136 male subjects. In total, 1655 images have been collected.

Regarding number of images per age, the largest number of images is of age of twenty, the smallest number is of age 21 and above. All other age have similar number of images.

All of the images provided cannot be used for age classification and estimation. Some of the images are of low quality, some have face images from different angles and some have bad lighting. So selection of these images needed to be done. The next step was to crop only faces from the images. After that, images with at least 100 pixels width and height were selected.

After the selection, 867 images from 242 subjects (121 male subjects and 121 female subjects) are left and 788 images have been left out (Table VII).

TABLE VII. Subjects per gender in the ageCFBP database

	No of subjects
Female	121
Male	121
Total	242

Regarding number of images per age, the largest number of images is still of age of 20 and the smallest number is of age 21 and above (Table VIII).

TABLE VIII. NUMBER OF IMAGES PER AGE IN THE AGECFBP DATABASE

Age	No of images	Age	No of images	Age	No of images
0	11	9	43	18	50
1	54	10	54	19	43
2	51	11	26	20	65
3	47	12	28	21	1
4	25	13	15	22	3
5	47	14	41	23	0
6	40	15	43	24	0
7	59	16	59	25	2
8	31	17	29		

C. Landmark annotation

In order to apply algorithms to the new database, positions of landmark points need to be determined. The positions are selected manually using a Python script that records mouse click positions. The script records the position of 26 landmarks on human face. These 26 landmarks are identified as landmarks with the highest impact on age estimation by [16]. The position of the landmarks can be seen in Fig. 2.

Input to the script is an image, and output is a file with coordinates of selected landmarks.

V. CONCLUSION

This paper gave an overview of the process of creation a new database for age estimation and classification. The ageCFBP database can also be used for face recognition, or gender classification. If the ageCFBP database is compared with the Fg-net database, it can be seen that the Fg-net database has a wider age range, but the ageCFBP database has more subjects (Table IX).

TABLE IX. COMPARISON OF SUBJECTS BY GENDER IN FG-NET AND AGECFBP DATABASE

	No. of subjects in Fg-net database	No. of subjects in ageCFBP database
Female	34	121
Male	48	121
Total	82	242

Future work will concentrate on collecting images of subjects of different age, not only from age 0 to 25. Also, future work will utilize the ageCFBP database to train the algorithms for age estimation and classification.

ACKNOWLEDGMENT

Authors would like to thank all the subjects that took part in the image collection for the creation of the ageCFBP database for age estimation.

REFERENCES

[1] Scholarpedia, "Facial Age Estimation", Available from: http://www.scholarpedia.org/article/Facial_Age_Estimation. [Accessed 13th February 2013]

[2] Grd, P., "Introduction to Human Age Estimation Using Face Images", Research papers Faculty of Materials Science and Technology Slovak University of Technology in Trnava, Special number. pp. 35-41, 2013

[3] Lanitis, A., "Comparative Evaluation of Automatic Age Progression Methodologies", EURASIP Journal on Advances in Signal Processing. 2008. pp. 1-10, 2008

[4] Fg-net. The Fg-net Aging Database. Available from: http://www-prima.inrialpes.fr/FGnet/html/benchmarks.html. [Accessed 19th March 2014]

[5] Morph. Morph Face Database. Available from: http://www.faceaginggroup.com/morph/. [Accessed 19th March 2014]

[6] Fu, Y. and Huang, T.S., "Human Age Estimation with Regression on Discriminative Aging Manifold", IEEE Transactions on Multimedia. 10 (4). pp. 578-584, 2008

[7] Ueki, K., Hayashida, T., Kobayashi, T., "Subspace-based Age Group classification Using Facial Images under Various Lighting Conditions", Proceedings of IEEE Conference on Automatic Face and Gesture Recognition. Southampton, April 2nd to April 6th 2006. pp.43-48, 2006

[8] Fu, Y. and Zheng, N., "M-Face: An Appearance-Based Photorealistic Model for Multiple Facial Attributes Rendering", IEEE Transactions on Circuits and Systems for Video technology. 16 (7). pp. 830-842, 2006

[9] Burt, D.M. and Perrett, D.I., "Perception of Age in Adult Caucasian Male Faces: Computer Graphic Manipulation of Shape and Colour Information" Proceedings of the Royal Society of London Series B Biological Sciences, 259 (1355). pp. 137-143, 1995

[10] Yao, B., Yang, X., Zhu, S.C., "Introduction to a large scale general purpose ground truth dataset: methodology, annotation tool, and benchmarks", Proceedings of the 6th international conference on Energy minimization methods in computer vision and pattern recognition. pp. 169-183, 2007

[11] Nakano, M. and Fukumi, M., "Age and Gender Recognition Using Facial Edge Information", Proceedings of the First World Congress of the International Federation for Systems Research: The New Roles of Systems Sciences For a Knowledge-based Society, Kobe, November 14th to November 17th 2005. JAIST Press, 2005

[12] Bastanfard, A., Nik, M.A., Deshibi, M.M., "Iranian Face Database with Age, Pose and Expression", In Proceedings of the International Conference on Machine Vision. Islamabad, December 28th to December 29th 2007. IEEE. pp. 50-55, 2007

[13] Gallagher, A. and Chen, T., "Clothing Cosegmentation for Recognizing People", Conference on Computer Vision and Pattern Recognition, Anchorage, June 23rd to June 28th. IEEE. pp. 1-8, 2008

[14] Ni, B., Song, Z., Yan, S., "Web Image Mining towards Universal Age Estimator", Proceedings of the 17th ACM international conference on Multimedia, Beijing, October 19th to October 24th 2009. pp. 85-94, 2009

[15] PAL, "Pal database", Available from: http://agingmind.utdallas.edu/facedb. [Accessed 19th March 2014]

[16] Grd, P., "Two-dimensional face image clasification for distinguishing children from adults based on anthropometry", Doctoral Thesis, Faculty of organization and informatics, 2015

[17] Ricanek, K. JR., Mahalingam, G., Albert, A.M., Vorder Bruegge, R.W., "Human face ageing: a perspective analysis from anthropometry and biometrics", Book Chapter in Age Factors in Biometric Processing Edited by Michael Fairhurst

[18] Narodne Novine, "Zakon o zaštiti osobnih podataka", Available from: http://narodne-novine.nn.hr/clanci/sluzbeni/305952.html. [Accessed 12th February 2013]

Forensic anthropometry from voice: an articulatory-phonetic approach

Rita Singh [†], Bhiksha Raj [‡], Deniz Gencaga [†‡]
[†] Computer Science Department, Carnegie Mellon University, Qatar
[†‡] Robotics Institute and [‡]Language Technologies Institute, Carnegie Mellon University, USA

***Abstract*—This paper addresses a problem that is of paramount importance in solving crimes wherein voice may be key evidence, or the *only* evidence: that of describing the perpetrator. The term *Forensic anthropometry from voice* refers to the deduction of the speaker's physical dimensions from voice. There are multiple studies in the literature that approach this problem in different ways, many of which depend on the availability of sufficient volumes of speech for analysis. However, in the case of many voice-based crimes, the voice evidence available may be limited. In such cases it is especially advantageous to regard the recorded signal as comprising multiple pieces of evidence. In this paper, we show how this can be done. We explain why, for any anthropometric measurement from speech, it makes sense to consider the contributions of each articulatory-phonetic unit independently of others, and to aggregate the deductions from them only in the aftermath. This approach is based on the hypothesis that the *relative* evidence given by different compositional units of speech can be more indicative of the anthropometric factor being deduced, than the evidence derived from the aggregate voice signal. We explain the applicability of this approach through experiments on standard speech databases.**

I. Introduction

Voice-based crimes comprise a significant fraction of crimes committed in the world today. Such crimes include those in which voice may be *key* evidence (such as a security camera video footage of a store robbery where the perpetrator may be masked but may demand something of the victims), and those in which voice is the *only* evidence. Examples of the latter include crimes committed over phone or internet, such as harassment, blackmail, threats, ransom demands in kidnappings, impersonation with intention to defraud in banking and other scenarios, voice-based phishing, hoax emergency calls, false reporting such as bomb threats in public areas, "swatting" calls to the police etc. All of these and more included in this category of crimes are often faced with the investigative challenge of finding the perpetrator(s) through the analysis of the voice evidence.

Forensic analysis of voice for this purpose benefits from a multitude of studies in different areas of science that study voice. These have demonstrated that the human voice carries a wealth of information about the speaker, including the speaker's physical characteristics such as height, weight, physiological characteristics and age, physical and mental state of health, social status, geographical origins etc., and a plethora of other information including that about their immediate physical surroundings. However, deriving such information is currently a goal that is challenged by many scientific problems. It is hinged on the understanding of the *signatures* of all of the speaker's personal characteristics and environmental parameters (at the time of recording) that are embedded in the speech signal, and using these to measure different characteristics of the speaker.

In this context, it is important to find out what must be done to identify the signatures alluded to above, e.g. what feature representations might best capture different signatures, what techniques might help identify them etc. Where our paper becomes relevant is that in addition to all of the above, it is also important to know *where* in the signal to look for such signatures, i.e. which parts of the signal are informative and which are not in the *expression* of the parameter in question and its signature. This may be tied to the actual feature representation(s) being used, but a framework is nevertheless needed to specify these informative locations in the signal. The goal of this paper is to provide such a framework.

In the forensic context, terms such as anthropometry (the measurement of body parameters), psychometry (the measurement of psychological parameters or state of mind), sociometry (the measurement of social parameters) etc. refer to the processes of deducing different categories of speaker characteristics that may help generate a reasonable description of the speaker and may thereby help locate him/her. To deduce these person-descriptive parameters, our framework comprises an approach based on considerations of the human speech production mechanism. In this paper we primarily present the reasoning behind, and evidence in support of, this *articulatory-phonetic* approach to anthropometry.

Although some studies have used information derived from phonemes for biometric applications, such as speaker matching [1], the choice of specific phonemes has largely been based on heuristic decisions. In contrast, the framework we present outlines a generic methodology based on well-established articulatory-phonetic guidelines, for the deduction of any person-descriptive parameter from voice. The key elements of our approach also involve a novel method for the sub-phonetic segmentation of speech in order to derive features that are compatible with this approach, and the demonstration of useful ways to visualize, interpret and utilize the information derived from the articulatory-phonetic categories. We build our arguments in favor of this approach through a brief review of the manner in which the human speech production process relates to the speaker's biometric parameters.

A. The speech production process and biometric parameters

The human vocal tract can be viewed as a system of dynamically configurable resonance chambers. *Voice* is the acoustic signal we hear when the mechanical vibrations of the vocal folds transform the aerodynamic energy of the air expelled from the lungs into acoustic energy in the form of sound waves. This *excitation* signal is further modulated into the sound patterns characteristic of speech by the physical movements of the vocal tract. The movements change the shape and dimensions of the various resonant chambers of the vocal tract, causing time-varying resonance patterns in the acoustic signal. This sequence of resonance patterns in the acoustic signal is perceived as (often) intelligible speech by the listener. Each distinct pattern, supported by the articulatory configuration of the vocal tract that produces it, is considered to be a unique compositional unit of speech, or a phoneme.

In continuous intelligible speech, the articulators are required to move continuously as the speaker forms words and sentences. During the production of continuous speech, the vocal tract attempts to "flow" from the canonical configuration for one phoneme to that of the next. The resonant characteristics of the phoneme-specific configurations are governed by the dynamics of the movement between different configurations, the degree to which the articulators achieve the canonical configuration for any phoneme, the excitation of the vocal tract, and all of the other articulatory and acoustic phenomena that affect the production of the phonemes.

All of these factors are known to be influenced by the speaker's physical and mental (biological) factors. Anthropometric characteristics such as skeletal proportions, race, height, body size etc. largely influence the voice by playing a role in the placement of the glottis, length of vocal cords, relative sizes and proportions of the resonance chambers in the vocal tract etc. When a speaker enunciates different phonemes, all of these structures act in concert, and the final speech signal produced carries the signatures of the specific vocal tract proportions, movements and configurations that the speaker is able to produce for each phoneme. Each phoneme therefore carries some evidence of all of these characteristics, except that the evidence is reasonably expected to be expressed differently for each phoneme.

The advantage of this reasoning is that it can be easily extended to apply to other categories of speaker characteristics, such as the speaker's mental state. Factors that relate to a person's mental state affect the *movement* and *locus* of the articulator configurations. This relationship is evident from several older studies that show that different mental states affect the body's muscle agility and response times, including that of the facial muscles, and by direct association, that of the articulators e.g [2], [3]. In one of his early expositions, Charles Darwin noted the relationship between emotion and specific patterns of muscle activity, particularly in the face [4]. Currently there is a large body of literature on skeletal muscle activity associated with psychological illnesses. Examples include muscle agility changes with anxiety and depression [5], with personality traits [6], etc. All of these effects are expected to carry over to the articulators. Following the same reasoning that we apply to a speaker's physical state, we expect different phonemes to also carry the signatures of the speaker's psychological state, and to express them differently from other phonemes.

Based on this reasoning we expect that estimates of a speaker's person-specific parameters may presumably be recovered more reliably from appropriate phoneme-specific analysis of the individual phonemes.

The rest of this paper is arranged as follows: In Section II we discuss some basic categorizations of speech from an articulatory-phonetic perspective. With this in context, in Section III we describe our approach for deriving anthropometric evidence from speech recordings. In Section IV we present experimental results in support of the proposed methodology. This is followed by conclusions in Section V.

II. A REVIEW OF PHONEME CATEGORIZATIONS

Based on the commonalities and differences between the articulator motions and configurations that produce them, articulatory phonetics differentiates speech into *phonemes*, its constituent compositional units, and further into several categories grouped by specific articulator locations and vocal fold activity. At the broadest level, phonemes are divided into *consonants*, which include some kind of airflow obstruction in the vocal tract, and *vowels*, which do not. These are briefly described below.

A. Articulatory-phonetic categorization of consonants

Depending on the voicing, place and manner of articulation, consonants are divided into several categories. These are named based on the key articulators involved. Fig. 4 (which also doubles as a template for representing results in the experimental section of this paper) lists these categories. Articulators that are considered in this categorization include the teeth, lips, hard palate, soft palate (velum), alveolar ridge, tongue (front, back or middle/sides, i.e. apex, dorsum and laminus respectively), uvula, glottis and pharynx. The list of phonemes in Fig. 4 is limited to those found in North American English, and also confined to the set of phonemes we analyze for the work presented in this paper. Consonants are further divided into two broad categories (not shown in the table). These are the *Obstruents*, which include all Stops, Affricates and Fricatives and are characterized by a significant constriction or obstruction in the vocal tract; and the *Sonorants or Approximants* which include the rest of the consonants, and are characterized by a slight constriction of the vocal tract. The key characteristics of the divisions named in Fig. 4 are described below.

Phonemes that involve the active vibration of the vocal cords are called *Voiced* phonemes, while those in which the vocal cords do not vibrate are termed as *Unvoiced*. For the purpose of this study, we focus on five categories of consonants based on the manner of articulation: namely *Plosives, Fricatives, Affricates, Nasals, Liquids and Glides*. The key characteristics

associated with each of these are: *Plosives*: complete stoppage of airflow, followed by sudden release of air; *Fricatives*: creation of turbulent airstream; *Affricates*: contain the characteristics of both plosives and fricatives; *Nasals*: complete airflow obstruction and release through the nose; *Liquids*: airflow along the sides or top of the tongue; *Glides*: stricture between the roof of the mouth and the tongue. If the stricture occurs such that air flows along the sides of the tongue, the glide is called a *Lateral* glide. If the sound is more "r"-like, the glide is called a *Rhotic* glide.

Depending on where in the vocal tract these key characteristics are generated (e.g the location of the airflow obstruction for a Plosive), the five categories above are further divided into the following subcategories. The articulators that are involved are indicated in parentheses: *Bilabial* (both lips), *Labiodental* (Lips and teeth), *Interdental* (upper and lower teeth), *Alveolar* (alveolar ridge), *Palatal* (hard and soft palate), *Velar* (Velum) and *Glottal* (glottis).

B. Articulatory-phonetic categorization of vowels

The primary difference between vowels and consonants is that in vowels there is no constriction of airflow in the vocal tract, whereas in consonants there is some degree of constriction somewhere in the vocal tract. Vowels are categorized based on their height (how high the tongue is), backness (which part of the vocal tract is pivotal in its production), laxness (rigidness or relaxedness of configuration) and roundedness (whether or not the lips are rounded). Fig. 5 in the experimental section of this paper shows the standard vowel categorization for North American English.

III. Proposed approach to anthropometry

Our methodology for the recovery of anthropometric attributes from voice is based on training a bank of phoneme-based predictors for each attribute measured, selecting a subset of them based on statistical criteria, and combining *their* decisions for a final prediction. Such "combination-of-predictors" approaches for the prediction of attributes are fairly standard in many different contexts in the machine learning literature, including multimedia processing [7] and audio processing [8]. What is novel about our approach is the utilization of articulatory-phonetic criteria to create the predictors in the mixture, and the specific mechanism for locating the right segments of speech for feature extraction. We describe this mechanism below.

A. A case for sub-phonetic *features*

Over the course of an utterance, spectral patterns vary continuously as the vocal tract transitions from the configuration for one phoneme to that for the next, often blurring the boundaries between them, resulting in a continuous, highly variable signal with complex, never-exactly-repeated spectral patterns. However, since the biophysical parameters of the speaker are manifested in *every* part of the speech signal, every section of these complex, never-repeated patterns is affected by the speaker's current biophysical state. Due to the complex nature of the speech signal itself, therefore, it is often difficult to distinguish between a *biophysically-affected* pattern seen in one phonetic context, and a naturally occurring pattern in another. As a result, signal measurements based on overall characterizations of the signal will often show weak, or no statistical relation to the speaker's state, although these relations may be locally evident in different portions of the signal.

In order to effectively characterize the expression of biophysical parameters on the speech signal, it therefore becomes necessary to focus on relatively *stable* sound structures that so typify the underlying phoneme that their absence or modification may change the perceived phoneme itself. Since the nature or state of the speaker also affects the articulation of these structures, the effect of the speaker's biophysical state/parameters on their expression can be isolated with relatively lower ambiguity than from other, more variable parts of the speech.

Such stable structures are generally phoneme-internal or *sub-phonetic* features. Candidate sub-phonetic features include voicing-onset time (VOT) [9], voicing offsets [10], onset of pitch, phonetic loci, etc. Indeed, each of these features is affected by different biophysical factors. For instance it is well known that VOT is affected by neurological disorders [11] and age [10], anomalies in onset of pitch are characteristic of vocal cord paralysis, formant positions in loci are related to age, body parameters [12], vocal tract shape [13] etc. The challenge, however, is that not all sub-phonetic features are affected by all biophysical factors. Possibly the most "universally" affected feature is the phonetic locus, which we briefly describe in the following subsection.

B. An HMM-based midstate-sharing technique using entropic regularizers for deriving stable sub-phonetic measures

As noted above, phoneme expression tends to be affected both by adjacent phonetic context [14] and longer-term prosodic and expressive trends in the speech signal. In order to isolate our measurements of the phonemes from the variability introduced by these contextual and longer-term effects, we must identify regions of each phoneme that are most invariant to context or longer-term trends.

The locus theory of phonemes states that every phoneme has a "locus", corresponding to a canonical arrangement of the vocal tract for that phoneme, and that the articulators move towards it in the production of the phoneme [15]. In continuous speech, the loci of phonemes may not be fully reached as the articulators move continuously from one set of configurations to another. Fig. 1 shows an example. While the locus theory does not explain all the variations observed in different instantiations of a phoneme, a key, valid insight that may be derived from it is that the interior regions of the phoneme that are representative of the locus are much more invariant to contextual and longer-term effects than the boundaries. The exact interior region that represents the locus may, however, vary with the instance of the phoneme. It need

not be at the center of the phoneme, and its position must be carefully estimated.

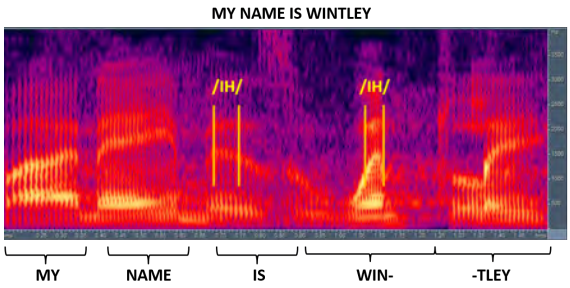


Fig. 1. The sentence "'My name is Wintley"' spoken by an adult male. Note the long-term spectral patterns that flow continuously across the entire utterance while moving towards different phoneme loci in the process. The phoneme IH is marked . It coccurs in two different phonemic contexts, and the influences can be seen at the beginning and end of the phoneme in each instance.

Since we wish to extract context-invariant features from the phoneme, we must identify this central "locus" segment and extract features from it. However, since the actual position and duration of this segment can vary significantly within each instance of a phoneme, actually identifying this segment is not a trivial task, and requires a sophisticated procedure.

We employ an HMM-based automatic speech recognition system trained using a modified version of the Baum-Welch algorithm for this segmentation. HMM-based large vocabulary continuous speech recognition systems model speech through context-dependent phonetic units called *triphones*. Each of these is modeled by an HMM with multiple states. As is well known, triphones are phonemes in context, many of which may be similar across multiple triphones. In conventional HMM-based speech recognition systems, the states of all HMMs are therefore *tied*, i.e. the state-output probability distributions of the HMMs are shared among the triphones corresponding to any phoneme.

The HMMs must be trained on speech data for which word-level, although not necessarily phonetic-level transcriptions are available. The phoneme-level (if not provided) and state-level segmentations are automatically derived during the training process. However, in the absence of other constraints, there is no assurance that any of the derived states will capture the locus regions of the phonemes; additional constraints are required to achieve consistent locus-region segmentations.

To achieve effective segmentation of locus regions, we incorporate additional constraints into our model. We model each triphone using a 3-state HMM. The "boundary" states of these models are intended to capture context-dependent variations of the phoneme, while the central state is intended to model the locus segments. Since the locus segments of different instances of a phoneme are expected to be very similar in structure (independently of context), this conformity is enforced by making the central states of all triphones of any phoneme share a common distribution, eliminating context-dependencies in the model for this state. We call this a *midstate-sharing* technique.

In order to minimize the variance of the distribution of this central "locus" state, we train the HMMs with a modified Baum-Welch algorithm that incorporates an entropic regularizer [16], which attempts to minimize the entropy of the distribution representing the central state. The effect of this regularization is to maximize the statistical similarity between the feature vectors assigned to the central state of every triphone. The details of the regularized training algorithm are omitted here, and largely follow the developments in [16]. The CMU-sphinx speech recognition system was modified for our experiments to include this. Figure 2 shows typical segmentations that are achieved by the algorithm. Note the similarity in the structure of the data in this state across the different instances of the phoneme.

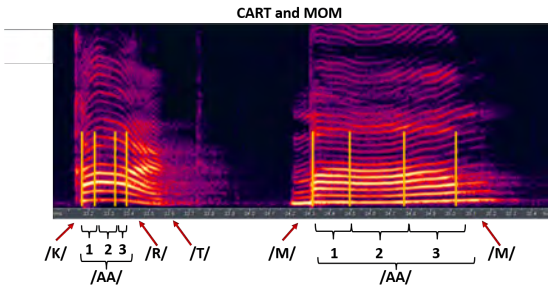


Fig. 2. State level segmentations for the phoneme AA in the American English pronunciation of the word CART (K AA R T) and MOM (M AA M).

IV. Experiments and Results

To evaluate the usefulness of the proposed approach, we apply it the deduction of height of speakers within the widely used and publicly available TIMIT continuous speech database [17], and to the estimation of age from the TIDigits database [18]. Note that our experiments are only for illustrative purposes, and we did not optimize the components used in them, such as feature types and models, to obtain the best possible performance. Nevertheless, we point out at the outset that the results we obtain for height are the best reported for the TIMIT database so far. The results we obtain from TIDigits have not been reported in the literature, and we have no points of comparison. We therefore only state them to the extent that they are illustrative of our procedure.

The first step in both cases was to segment the databases into their phonemic units. For this, we used the technique discussed in Section III-B. We used 3-state left-to-right Bakis topology HMMs for segmentation. These were trained on 5000 hours of clean speech data from a collection of standard speech databases in English available from the Linguistic Data Consortium. The training databases were parametrized using high-temporal-resolution MFCC vectors [19]. These were computed over 20ms analysis frames 200 times a second, to achieve a temporal resolution of 5ms. The trained HMMs were finally used to derive phoneme- and state-level segmentations of the recordings. The region of the middle state of each phoneme HMM was then taken to represent the locus region from which features for the deduction of height and age could be derived.

In this context, we note that the features we use for segmentation are not necessarily the same as the features we use to derive the anthropometric parameters of the speakers. Once the state-level segmentations are obtained, we extract the latter features from the central locus segment, as explained earlier in Section III-B. For the TIMIT data, our features were spectral features that capture resonance and fine variations in the envelope of the spectrum, but do not capture voluntary pitch-level detail. In this representation, we computed a 64-point *liftered* periodogram using Welch's method [20] over the locus segment. We used a 20-point lifter; this retains envelope-level detail in the spectrum, while smoothing out pitch-level structure. For the TIDigits database, we extracted the first five formants using the Burg's method [21] from each relevant segment.

Subsequent to the extraction of features, we followed two strategies. In the case of TIMIT, we conducted a 10-fold cross-validation experiment, partitioned across speakers. In each fold, all instances of a phoneme from 90% of the speakers were used to train a regression, which was used to predict heights from each instance of the phoneme spoken by the remaining 10% of speakers. The overall phoneme-specific predicted height for any speaker was the average of the predictions made by the individual instances of the phoneme by that speaker.

In the case of TIDigits, the phoneme-specific predictors were random forest regressions [22] that predict age, gender and accent from the formant features derived from the TIDigits sub-phonetic segments. Our experiments employed forests with 32 trees. Increasing this number had no significant effect on accuracy.

As described in Section II, a given speaker characteristic may heavily influence the sub-phonetic features of some phonemes, while not affecting others to the same degree. The specific phonemes affected may also vary with the feature. Accordingly, in both cases the next step was to identify the phonemes that resulted in the most statistically significant ($P - value < 0.0001$ using a t-test) predictions.

To demonstrate that the articulatory-phonetic conjectures we present earlier in this paper indeed hold, the predictions from the set of the most predictive phonemes were then combined using simple weighted interpolation. The weights for any predictor were the inverse training error standard deviations. The weighted combination resulted in a single aggregate prediction for the speaker. Note that in practice, for estimation of speaker parameters from new data, the knowledge of *which* phonemes are most predictive of the parameter (obtained from a different database, such as TIMIT in this paper) can still be effectively used in the same manner.

Figure 3 shows the Root Mean Squared Error (RMSE) of the height predictions obtained from each phoneme. The error ranges from 2.75 inches for the phoneme IY to 3.5 inches for ZH. We observe that all phonemes are not equivalent in their ability to predict height. Except for ZH (159 instances) and UH (762) instances, all phonemes are well represented in the set. The results are hence largely reflective of the innate

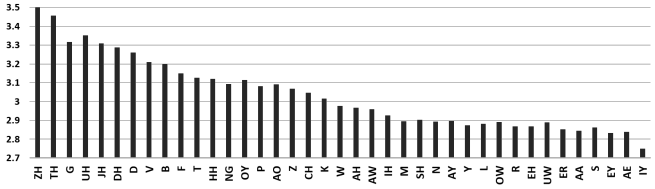


Fig. 3. RMS Error of the predicted values of height for each phoneme.

information about speaker height encoded in the phoneme. While not plotted here, the correlation between the predicted height and the actual height follows the trend shown by prediction accuracies and ranges between 0.1 (for ZH) to 0.7 (for IY). All correlation values are statistically significant with P-values of 0.001 or better. The combined prediction from all the phonemes with P-values less than 0.0001 achieves an overall RMSE of 2.54 inches.

To test our original hypothesis that the predictive ability of the phonemes varies, and reflects the articulator configuration of the phoneme, we evaluate the R^2 value of the individual phonemes, which represents how much of the variance in the dependent variable (height) can be predicted from the variations in the features for the phoneme. Phonemes with greater R^2 are more predictive of height. Fig. 5 shows the ten vowels with the greatest ability to predict height, color coded to show the rank ordering. Fig. 4 shows the top ten consonants.

We note that the phonemes that predict height best cluster closely in terms of articulator configurations. In particular, front vowels, and velar, alveolar and palatal consonants, all of which capture the acoustics of the *back* of the vocal tract, which relates more to body size, are seen to have the highest ability to predict height. Combining the predictions of just these phonemes results in an overall RMSE of 2.48 inches, which is among the best reported results for this data set [10].

Manner of articulation	Voicing	Place of articulation						
		Bilabial	Labio-dental	Inter-dental	Alveolar	Palatal	Velar	Glottal
Stop	Voiced	b			d		g	
	Unvoiced	p			t		k	
Fricative	Voiced		v	dh	z	zh		
	Unvoiced		f	th	s	sh		hh
Affricate	Voiced					jh		
	Unvoiced					ch		
Nasal	Voiced	m			n		ng	
Liquid(L)	Voiced				l			
Liquid(R)	Voiced					r		
Glide	Voiced	w				y	w	

Fig. 4. Classification of consonants based on the involvement of different parts of the vocal tract (place of articulation) and the manner of their movements (presence of absence of voicing, and the manner or articulation). This figure also shows consonants with the highest R^2 value for height. Shades of decreasing color intensities (black to light grey) depict decreasing R^2.

Note that the final combination was done only in the case of height estimation with TIMIT. We present the TIDigits case for illustration of the same technique for age prediction. Fig. 6 shows the capacity of the different phoneme to predict age, accent and gender from the TIDigits database. Of these, we

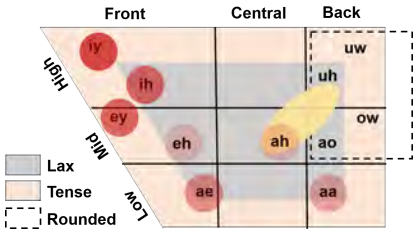


Fig. 5. Vowels in American English, modified from [23] to represent the specific vowels modeled by by the CMU Sphinx ASR system [24] used in our experiments. There are four additional phonemes used in our ASR system. These are the three *Diphthongs*: *ay* as in *side*, *aw* as in *how* and *oy* as in *toy*, and the semivowel *er* as in *surround*. This figure also shows the vowels with the highest R^2 for height. The semivowel ER is not shown here but is also as highly correlated to height as IY, which exhibits the greatest R^2. The figure shows decreasing R^2 values as red circles of decreasing color intensity.

only wish to highlight the case of age. We find that age is highly correlated to only a few phonemes, all of which turn out to be vowels.

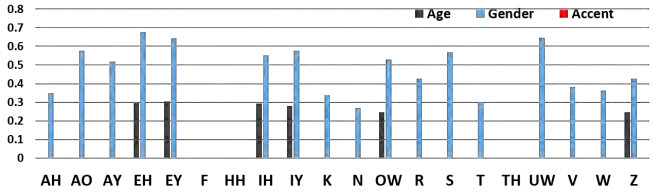


Fig. 6. The predictability of age, accent and gender from different phonemes in the TIDigits database, computed using sub-phonetic formant features over the entire database. Accent is not predictable from individual phonemes at all. This makes sense since prior studies have shown that accent is encoded *jointly* in phonemes, and entire formant charts are needed to identify them.

V. Conclusions

The results clearly demonstrate the validity of an articulatory-phonetic-based approach to forensic analysis of voice. The proposed methodology is of particular use in cases where the available voice sample may be of short duration, e.g. in the word *Mayday* from a hoax call, comprising just a few phonemes. Estimating speaker parameters based on only the most appropriate phonemes can provide useful results in these scenarios.

The articulatory-phonetic approach presented in this paper is exemplified in the context of deducing height and age, but can be applied to any other anthropometric, psychometric, sociometric and suchlike measurements from the voice. While we have not addressed robustness issues and analysis of noise-corrupted recordings in this paper, we have found in practice that once we are able to generate accurate sub-phonetic segmentations in these cases (using appropriate robustness techniques and specially modified state-of-art automatic speech recognition systems), our methodology applies well to the estimation of both anthropometric and psychometric parameters. We are in the process of publishing these results.

Acknowledgment

This material is based upon work supported by the U.S. Department of Homeland Security under Award Number 2009-ST-061-CCI002-07, via the Command, Control and Interoperability Center for Advanced Data Analysis (CCICADA).

References

[1] D. Mendes and A. Ferreira, "Speaker identification using phonetic segmentation and normalized relative delays of source harmonics," in *Proc. 46th Audio Engineering Society Conference on Audio Forensics: Recording, Recovery, Analysis and Interpretation*, Denver, Colorado, USA, 2012, pp. 215–222.

[2] G. D. Burrows, "Skeletal and facial muscle psychophysiology," *in Handbook of studies on depression*, 2013.

[3] R. D. Kent and J. C. Rosenbek, "Acoustic patterns of apraxia of speech," *Journal of Speech, Language, and Hearing Research*, vol. 26, no. 2, pp. 231–249, 1983.

[4] P. Ekman, *Darwin and facial expression: A century of research in review*. Ishk, 2006.

[5] I. B. Goldstein, "The relationship of muscle tension and autonomic activity to psychiatric disorders." *Psychosomatic Medicine*, vol. 27, no. 1, pp. 39–52, 1965.

[6] ——, "Role of muscle tension in personality theory," *Psychological Bulletin*, vol. 61, no. 6, p. 413, 1964.

[7] S. Gutta, J. R. J. Huang, P. Jonathon, and H. Wechsler, "Mixture of experts for classification of gender, ethnic origin, and pose of human faces," *IEEE Transactions on Neural Networks*, vol. 11, no. 4, pp. 948–960, 2000.

[8] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1177–1193, 2012.

[9] M. Sonderegger and J. Keshet, "Automatic discriminative measurement of voice onset time." in *INTERSPEECH*, 2010, pp. 2242–2245.

[10] R. Singh, J. Keshet, D. Gencaga, and B. Raj, "The relationship of voice onset time and voice offset time to physical age," in *Proc. ICASSP*, 2016.

[11] P. Auzou, C. Ozsancak, R. J. Morris, M. Jan, F. Eustache, and D. Hannequin, "Voice onset time in aphasia, apraxia of speech and dysarthria: a review," *Clinical Linguistics & Phonetics*, vol. 14, no. 2, pp. 131–150, 2000.

[12] R. Greisbach, "Estimation of speaker height from formant frequencies," *International Journal of Speech Language and the Law*, vol. 6, no. 2, pp. 265–277, 2007.

[13] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE TSAP.*, vol. 2, no. 1, pp. 133–150, 1994.

[14] C. T. Ferrand, *Speech Science: An Integrated Approach to Theory and Clinical Practice*. Allyn & Bacon, 2006.

[15] P. Delattre, "Coarticulation and the locus theory," *Studia Linguistica*, vol. 23, no. 1, pp. 1–26, 1969.

[16] M. Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction," *Neural Computation*, vol. 11, no. 5, pp. 1155–1182, 1999.

[17] L. D. Consortium, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," https://catalog.ldc.upenn.edu/LDC93S1, 1993.

[18] ——, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," https://catalog.ldc.upenn.edu/LDC93S10, 1993.

[19] R. Singh, B. Raj, and J. Baker, "Short-term analysis for estimating physical parameters of speakers," in *Proc. International Workshop on Biometrics and Forensics*. Limassol, Cyprus: IEEE, March 2016.

[20] P. D. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE TAE*, vol. 15, no. 2, pp. 70–73, 1967.

[21] J. P. Burg, "A new analysis technique for time series data," *NATO advanced study institute on signal processing with emphasis on underwater acoustics*, vol. 1, 1968.

[22] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[23] T. L. S. Project, "Sounds of standard American English," *University of Arizona*, 2001.

[24] "The cmu sphinx suite of speech recognition systems," http://cmusphinx.sourceforge.net/, 2013.

INTELLIGENT SYSTEMS

PAPERS

Computer vision for the blind: a dataset for experiments on face detection and recognition

S. Carrato, S. Marsi, E. Medvet, F. A. Pellegrino, G. Ramponi and M. Vittori
DIA, University of Trieste, Italy
e-mail: {carrato, marsi, emedvet, fapellegrino, ramponi}@units.it, michele.vittori@gmail.com

Abstract – We present a video dataset created for the needs of a research project that aims at developing vision-based techniques that assist the social interaction of a blind person. Two totally blind users have acquired the sequences, using at the same time a glasses-mounted camera and a necklace-mounted one. The video sequences were acquired in different environments, selecting conditions in which a user could be interested in detecting the presence of some of his/her acquaintances, to approach them in a most natural way. The sequences have been temporally cropped to extract video shots that, by inspection, were deemed to contain events valuable for the goals of the project. The shots are presently being annotated, and some preliminary experiments on face detection have been performed on the annotated data. We also present some observations about the specific application that is being considered.

I. INTRODUCTION

We present some results from an ongoing research project devoted to image processing and computer vision techniques that can facilitate a blind user in his/her social relations [1,2]. We aim at enabling the user to detect and recognize one of his/her acquaintances among the people that are present, as a normally-sighted person would do. A fundamental characteristic of the project is the involvement, from its early stages to its end, of a Users' Group that includes people affected by visual impairments and personnel whose professional role is to take charge of the assistance of people with impairments. For example, video data that are necessary to perform the experiments in the detection of faces and the recognition of the expressions are being acquired directly by a blind person, who in this way is able to manifest needs and preferences that will make the final devices suitable for a practical usage.

This paper presents a video dataset that was acquired with the purpose of performing experiments of face detection and recognition in realistic (and thus difficult) conditions. We make some observations about the filmed scenes, and we illustrate the annotation procedure by which we generate a ground truth for the analysis of the sequences. A section of the paper is then devoted to the hardware requirements imposed by our goals, deducing some solutions for a suitable architecture. On these bases, a few algorithms for face detection that can satisfy our needs are briefly described, and some preliminary results we have obtained on our sequences are presented.

II. RELATED WORK

We consider a system which may assist the social interaction of a blind person. The assistance may tailor several use cases. The issues that are perceived as the most important, according to the project "Social Interaction Assistant" [3], are due to the lack of some non-verbal cues that are present in the personal interaction; they are especially related to person recognition, eye gaze direction detection, and facial expression recognition. To approach these problems, the detection of faces in the frame is a fundamental preliminary step: to this end, we annotated the video sequences in order to enable insightful assessment of face detection algorithms.

Several datasets have been proposed for face detection: they differ in the level of detail of the annotation (e.g., a simple bounding box or several facial landmarks) and in the kind of data which is annotated (e.g., acquired in studio or collected from "the wild"), but most of them are built on images rather than video sequences. The use cases considered in our project require working on video sequences and hence we needed to cope with a much larger amount of data, resulting in a trade-off between how much data to annotate and the detail of annotation. We were driven by the goal of providing a dataset which can foster research beyond the mere detection: for example, including eye-related landmarks and considering consecutive frames permits to reason on pose modifications which may suggest an intention to communicate.

The MALF dataset proposed in [4] consists of 5,250 images and approximately 12,000 faces. The aim of the authors was to provide a large dataset for fine-grained evaluation of face detectors on so-called "in the wild" images. Beyond a well-defined bounding box, the annotations consist, for each face, of several categorical attributes (e.g., w/ or w/o glasses). As in our dataset, the annotations contain information about face relative size and pose, which allows for an assessment tailored to specific subset of faces (e.g., profile faces, far faces, and so on).

A frequently used dataset was presented in [5] which consists of 2,846 grey and colour images with approximately 5,100 faces: the annotations are given in form of elliptical regions which enclose faces. According to the authors, elliptical regions better capture faces; yet, they require special care when used to assess detectors which output rectangular bounding boxes.

Finally, in [6] both a model and a dataset for face detection are proposed which are tailored specifically to detailed landmark localization. The dataset contains 205 cluttered images with 468 faces. For each face, 6 landmarks, a discretized viewpoint and a bounding box form the annotation.

III. THE VIDEO DATASET

Many video sequences have been acquired by members of the Users' Group of the project. We show some examples recorded by two of them; these users are fully blind from birth, but they are determined to behave as far as possible in an autonomous way. They suffer only slightly from head- and body-posture modifications and from mannerisms, like body rocking, that are typically acquired at an early age by visually impaired people [7,8]. It is obvious that such mannerisms drastically affect the quality of the acquired video, and we expect that further sequences we will add to our dataset will provide other interesting case studies. Also for this reason, we opted for acquiring the scenes in two modalities: the final choice will be user-dependent.

Two commercial devices have been used to record the scene at the same time: in one case, which we will refer to as GL, the camera was mounted on the bridge of a pair of sunglasses; in the other case, labelled as NL, on a light support held by a short necklace. The glasses-mounted camera had a resolution of 1280 x 720p pixel and an angle of view of 135 deg.; the resolution of the necklace-mounted camera was 1920 x 1080p pixel and its angle of view was 124 deg. The GL sensor proved to provide slightly poorer images in terms of detail sharpness. The positioning of the videocameras is shown in Fig. 1. In a real deployment, these devices might be replaced by similar equipment able to stream video via wire, WiFi or Bluetooth; however, the results discussed here represent influential factors in the continuation of the project.

The video sequences were acquired in different environments, selecting conditions in which the user could be interested in detecting the presence of some of his/her acquaintances, to approach them in a most natural way. The selected locations are a university library, a coffee shop, the hall of a public building, the neighbourhood of a bus stop. In all cases, proper procedures were followed to comply with the normative about privacy protection. More precisely, signs providing reference to specific national laws were posted in the area in advance and during the acquisition, to inform passersby about the shooting of video sequences for scientific purposes. Written informed consent was obtained from all the people who actively participated in the operations.



Figure 1. Positioning of the two videocameras for the acquisitions

Some sample frames of the sequences are shown in Figs. 2 and 3, which permit to notice their typical characteristics. Several observations can be made.

- Since the user of course lacks any feedback about the subjects in the field of view, faces can be partially occluded or can be partially outside the frame.
- The wide angle of view of the acquisition devices is a necessity: acquiring with standard optics (e.g. with a camera such as the ones typically mounted on smartphones) would make worse the above mentioned problem of faces outside the frame. However, wide angle causes geometrical distortions to appear. Compensating for them is theoretically simple, but implies computational costs which may make an actual realization quite complex, and may require proper dedicated resources.
- The scene conditions are very different in the different contexts, and can change abruptly with time. In particular, the automatic exposure control of the camera can be unable to comply with the range of the illumination. Back-lighting of the people in the scene is particularly critical, even if many face detection and recognition algorithms are by design relatively robust to this disturbance.
- People in the scene often tend not to look straight towards the user; this is an instinctive behaviour, due to politeness.
- The field of view of both the GL and the NL cameras can easily be partially occluded, by a tuft of hair or by a lapel of the dress respectively. A firmly placed camera, especially the NL one, or tightly held hair and dress can be unpleasant to wear.
- Sudden, fast and wide subjective movements are present, especially in the GL sequences. Some of the users move their body and in particular their head towards perceived sounds; other users are much more static.

The available sequences were temporally cropped to

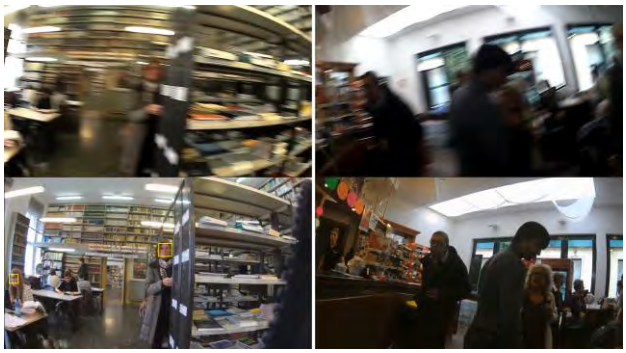


Figure 2. Two examples of motion blur induced by a sudden movement of the head (GL camera, top), and the corresponding frame filmed with the NL camera (bottom)

extract video shots that, by inspection, were deemed to contain events that are valuable for the goals of the project. They are a significant sample, which comprises 6,652 frames containing a grand total of 11,513 faces.

Details about the sequences can be seen in Tab. 1.

IV. ANNOTATION OF THE DATASET

Faces in the selected shots were then annotated using a freely available tool, ViPER-GT [9]. Particular care was placed in selecting which features to annotate, and exactly how. Indeed, it is known that the way in which annotation is performed can modify the outcome of an experimental comparison of face detection methods [10].

We annotated, in each frame, every face whose estimated distance from the user was less than 5 meters, which resulted in a bounding box whose longest side was at least 20 pixel long: the rationale for this criterion was to annotate all faces that indicated the possibility of an immediate social interaction with the user. A rectangle was traced, vertically delimited by chin and forehead (normal hairline, independent of the actual presence of hair in the subject), and horizontally delimited by the ears or, for rotated faces, one ear and the opposite foremost point between the tip of the nose and the profile of the cheek. Yaw (rotation of the head around a central vertical axis) was constrained to +/-90 deg.; pitch (horizontal, left-right axis) and roll (horizontal, antero-posterior axis) were not constrained; however, the presence of significant yaw (possibly also with pitch and roll contributions) was denoted by a dedicated flag, set if the farthest eye was not clearly visible. A further flag was set if the face was partially occluded. Beyond the rectangle, the positions of the centres of the two eyes and of the mouth were also annotated.

V. REQUIREMENTS FOR THE IMPLEMENTATION PLATFORM

The particular characteristics of the project require special features of the platform to be used in the system implementation: the system must be wearable and have a low power consumption, even if an autonomy of several hours can be considered sufficient for a typical usage. A further constraint is related to the different type of processing to be performed: indeed, the complete system requires both low-level video pre-processing to condition the input stream, and high-level analysis to extract and process the significant information from the video sequence.

The critical conditions in which the sequences are acquired, such as the presence of uncontrolled light dynamics, the distortion introduced by the wide-angle lenses and the possible shaking of the camera, require the adoption of suitable pre-processing algorithms to minimize these drawbacks and to control the input frames characteristics. These algorithms must be able to work in real time directly on the input video stream, thus they can greatly benefit from a suitable parallel processing that is available in processing systems like GPUs or FPGAs.

The selection of the appropriate hardware device must also take into account constraints related to portability and to power consumption; these features lead to a significant limitation in the type of suitable devices. It should be noted in fact that most of the GPU devices are employed in systems like console or desktop computers where low power consumption is not a priority, while very few GPU families [11] have been designed to be used inside mobile systems. Moreover, even within the same low-power GPU family, only the devices with more limited performances are integrated in mobile devices such as smartphones or tablets, while devices with higher performances are typically available only in consoles or notebooks.

TABLE I. DETAILS ABOUT THE SEQUENCES CHOSEN TO BE ANNOTATED

	Length (frames)	**Camera**	**Location**	**Features**
A	379	NL	Outside a coffee shop	High contrast. Partially occluded faces. Tilted camera.
B	1138	NL	Bus stop	Many faces. Good soft light.
C	1380	GL	Inside a library	*Motion blur* induced by head movements. Pretty good light.
D	697	NL	Outside	Partially occluded faces. Very tilted camera.
E	1200	NL	Inside a coffee shop	Low light. Fast movements. Many partially occluded or lateral faces.
F	838	NL	Inside a cafeteria	High contrast. Pretty good light. Slow movements.
G	360	GL	Inside a coffee shop	High contrast. Pretty good light.
H	300	GL	Inside a coffee shop	Low light. Fast movements. Many lateral faces.
I	360	GL	Bus stop	Good light. Hard shadows.

With these premises, using an FPGA seems the most promising solution. In fact, current FPGAs make available, beyond programmable hardware resources such as memories, logic blocks, interconnection networks etc., also some hardware processors as well as communication systems, together with many other functional blocks [12]. These complex architectures allow the designer to realize in a single chip a highly advanced processing system with many customizable functions.

A feasible solution for the realization of the present project is therefore an embedded a system that adopts a high performance FPGA as processing unit. A quite low-priced board with characteristics consistent with the system constraints could be [13]. This board centralizes all

the processing inside a single chip of the Xilinx Family Zinq-7000. This chip is equipped with a Dual Core ARM Cortex-A9 processor, several embedded memory blocks, interface blocks for an easily access to any external memories and many other functional blocks. It includes programmable logic that can be exploited for the realization of custom blocks to speed up the specific algorithm.

The complete processing system will implement a real-time operating system on one of the two ARM processing core. This core will handle the acquisition of images from the camera, together with any other data that may be sourced from other sensors that may be integrated in the system such as gyroscopes and accelerometers, suitable to improve the image stabilizer algorithms. These data will be transferred in real time to a shared memory. The video stream, using the FPGA resources, will be processed in real time through suitable image correction algorithms to minimize the drawbacks and the artefacts introduced by the camera motion, the geometric distortion of the lens and the high dynamic range of the input signal. The processed images will then be made available to the second ARM core through an appropriate shared memory. In the second ARM processor core a Linux operating system will be installed. This part of the system will be devoted to high-level signal processing operations such as face detection and face recognition, as well as any access to communication systems. This processor could even benefit from the FPGA hardware to speed up repetitive procedure exploiting the parallel processing facilities [14,15]. Eventually, this processor can also be used to perform the task of estimating the quality of the final results; thanks to these data, through a feedback control it can tune the parameters of the pre-processing algorithms to optimize the performances.

VI. SELECTED FACE DETECTION ALGORITHMS

The development of new algorithms for face detection is out of the scope of this project, which is more focused on our specific application (i.e., to develop an effective assistive technology for the blind); consequently, we analysed the state of the art in face detection algorithms and selected, based on the characteristics of the acquired video and on the computational performances that can be provided by a mobile platform, those which better fit our needs; they are briefly described in this section.

We point out that, at least in the first phase of our project, we aim at detecting faces that are almost frontal with respect to the user. This of course makes life simpler for the detection tools, in terms of both performances and computational load, and it is also a sensible first choice to detect those people in the scene who may be willing to interact with the user. Further developments of the algorithms and hardware will permit to include in the detection also subjects whose head is turned away.

The Normalized Pixel Difference (NPD) algorithm [16], very recently proposed by Liao et al., is based on the difference to sum ratio between couples of pixels, and uses a deep quadratic tree for the learning algorithm. It should be very effective in detecting faces with arbitrary size and pose, also in presence of occlusions and in cluttered scenes; a Matlab implementation has been made available by the authors in [17].

The so-called Viola-Jones (VJ) algorithm is older, as it has been proposed in 2001 [18] for generic object detection and in 2004 [19] specifically for face detection. It is based on Haar-like features and a cascade of simple classifiers, and should be very fast, also thanks to the use of a new image representation, the *integral image*, which permits the computation of the features at any scale in constant time; part of its wide popularity is probably due to its availability in the OpenCV framework [20].

Pixel intensity comparison is also used in the PICO algorithm [21]. Similarly to the VJ case, a cascade of rejecters is used as a decision tree; no integral image is used, and high processing speed is given by the simple structure of the tree nodes, which are just binary tests. The full source code has been made available by the authors in [22].

The three mentioned tools for face detection have been used in our experiments as provided by their authors; no attempt has been made to re-train them and make them more suitable to our dataset.

VII. ESTIMATION OF THE QUALITY OF THE FACE DETECTION ALGORTITHMS

The evaluation of the performances of the face detection algorithms is not trivial. From one side, their behaviour typically depends on several parameters, and it may be difficult to find the best combination of their values. To this extent, we decided (with one exception, described below) to use the default values suggested by the authors, assuming they performed an exhaustive series of tests. Actually, the best parameters combination can significantly vary according to the characteristics of the video data (e.g. blurriness, lighting conditions and backlighting, dimensions of the faces…), so that a more careful algorithm comparison should include the tuning of the parameters using our video database.

Performances cannot be evaluated based only on the number of detected faces in each frame. If for a certain frame the ground truth reports e.g. two faces, an algorithm may correctly detect only one face (this is a *true positive*, TP), miss the second one (this is a *false negative* (FN) error) and classify as face another object (a *false positive* (FP) error). Thus, it is common to use two quality parameters, *precision* and *recall*, defined as

$$precision = \frac{tp}{tp + fp} \quad (1)$$

$$recall = \frac{tp}{tp + fn} \quad (2)$$

being *tp, fp* and *fn* respectively the number of TPs, FPs and FNs.

It has also to be noted that the evaluation of *tp, fp* and *fn* requires some care, for two reasons. The first is that, since these algorithms typically provide the *x* and *y* coordinates of the detected face bounding box (BB), some

tolerance has to be allowed with respect to the coordinates of the BB in the ground truth: a misplacement of e.g. a couple of pixels in a 50 pixel size face should not be interpreted as a wrong detection. The second is related to the fact that, in case multiple faces are present in the image, the algorithm will likely *not* find the faces in the same order in which they appear in the ground truth; an automatic procedure based on a mere one-by-one position matching would erroneously yield only few, if any, TPs. It is then necessary to apply a minimization algorithm, which solves the assignment problem of finding the best match between the positions found by the algorithm under test and those reported in the ground truth: we used the Hungarian algorithm [23].

VIII. PRELIMINARY RESULTS

Trials are still in progress, but we can provide some first numerical results and some preliminary considerations derived from the direct observation of the comparison results.

First of all, we observe from Tab. 2 that the NL videocamera permitted to obtain far better results than those of the GL camera. This is due to several factors, including the slightly poorer quality of the GL camera in terms of resolution and sharpness, and the better stability of the NL solution.

TABLE II. AVERAGE *PRECISION* AND *RECALL* FOR EACH CAMERA TYPE

	Average Recall (%)	**Average Precision** (%)
NL	45.69	20.06
GL	29.85	13.92

TABLE III. AVERAGE *PRECISION* AND *RECALL* FOR EACH ALGORITHM

	Average Recall (%)	**Average Precision** (%)
NPD	33.29	27.27
VJ	28.98	17.25
PICO	43.70	24.87
PICO alt	53.77	12.62

Then, Tab. 3 shows that the algorithm PICO [21] offered the best results when considering both precision and recall. Its recall values have a +10% margin on the second best algorithm and a modest -3% drop in precision with respect to NPD [16], which is the best algorithm precision-wise.

It is very interesting to analyze the worst average results, as seen in Tab. 4. Tests on videos A, D, E and H gave significantly worse results, which can be explained by some common characteristics in the videos. As seen in Tab. 1, A, E and H present harsh light conditions, being shot in fairly dark and highly backlit areas, while videos A and D are both shot with a tilted camera. It has to be noted that in those four worst cases Viola-Jones [19] had better results than NPD, while still being outperformed by PICO both in the recall and in the precision results. Actually, the same four videos coincide with the cases where PICO had the best precision values.

Our preliminary experiments then indicate that in good conditions, good light, straight necklace-mounted camera, NPD has the best precision values and stands almost on a par with PICO's recall values. In difficult conditions, bad light, high contrast, and tilted camera, PICO performs significantly better than the other tested methods.

Comparing the positions of the TPs with those of the FPs, it is evident that the former are relatively stationary in adjacent frames, while the latter tend to appear in different positions of the image with no apparent continuity. Thus, it has been decided to focus on the recall results and give less importance to the precision of the tests. It is indeed reasonable that simple methods that exploit temporal correlation can be devised, which will be able to detect and reject many FP cases.

We then decided to modify a parameter in the PICO options to further improve its recall values, at the cost of worse precision and longer computing times. The results of these alternative tests can also be seen under the name "PICO alt" in the previous tables. They highlight how, at the cost of a halved precision, the recall values further improve: about 10% more faces are recognized. Even if presently we have no quantitative data, it should be noticed that this alternative method will be significantly affected in terms of the time taken to make the detection.

TABLE IV. *PRECISION* PERCENTAGES FOR THE DIFFERENT TESTED METHODS, ON EACH VIDEO

	NPD	**VJ**	**PICO**	**PICO alt**
A	4.65	2.01	4.91	2.27
B	55.25	40.20	51.14	27.35
C	10.58	17.07	12.31	6.77
D	8.47	3.58	14.48	7.82
E	12.31	8.43	13.04	7.76
F	35.40	13.39	29.11	18.43
G	22.70	27.92	22.43	11.47
H	13.21	8.78	23.36	17.19
I	47.21	46.29	54.83	29.88

IX. CONCLUSIONS

The previous analysis of the face detection results on our dataset, and specifically the consistency in results variation among algorithms on the different video sequences, confirm that the dataset is a valid benchmark for testing face detection algorithms. We can observe that even in the best case scenario the recall results obtained never pass the 80% mark, and average at little more than 50% even when considering the algorithm which performed best in this area.

This confirms the usefulness of a specific dataset to measure algorithm performances relative to the scope of a social aid to visually impaired people, instead of evaluating such performances on generic datasets such as FDDB [5] or "Faces in the wild" [6]. A comparison of the figures of merit available for the latter databases and the results we obtained on our dataset clearly indicates that the notion of data acquired "in the wild" has to be modified: real world problems such as the one we are addressing imply the need to cope with even harsher environments.

ACKNOWLEDGMENTS

This work has been supported by the University of Trieste - Finanziamento di Ateneo per progetti di ricerca scientifica - FRA 2014, and by a private donation in memory of Angelo Soranzo (1939-2012).

The authors are grateful to the Users' Group for their precious collaboration. They also thank Igor Pandžić, Nenad Markus and Visage Technologies for providing the PICO software, and Anil K. Jain and Shengcai Liao for the NPD software.

REFERENCES

[1] M. Bonetto, S. Carrato, G. Fenu, E. Medvet, E. Mumolo, F.A. Pellegrino, G. Ramponi, "Image Processing Issues in a Social Assistive System for the Blind," in 9th International Symposium on Image and Signal Processing and Analysis, ISPA 2015, Zagreb, Croatia, September 7-9, 2015.

[2] S. Carrato, G. Fenu, E. Medvet, E. Mumolo, F.A. Pellegrino, G. Ramponi, "Towards More Natural Social Interactions of Visually Impaired Persons," in Int. Conf. on Advanced Concepts for Intelligent Vision Systems, ACIVS 2015, Catania, Italy, Oct. 26-29, 2015.

[3] S. Panchanathan, S. Chakraborty, T. McDaniel, "Social Interaction Assistant: A Person-Centered Approach to Enrich Social Interactions for Individuals with Visual Impairments", IEEE Journal of Selected Topics in Signal Processing, to be published, 2016 (IEEE Early Access Articles).

[4] B. Yang et al. "Fine-grained evaluation on face detection in the wild," Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. Vol. 1. IEEE, 2015.

[5] V. Jain and E. Learned-Miller, "FDDB: A Benchmark for Face Detection in Unconstrained Settings," Technical Report UM-CS-2010-009, Dept. of Computer Science, University of Massachusetts, Amherst. 2010.

[6] X. Zhu and D. Ramanan. "Face detection, pose estimation, and landmark localization in the wild," Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012.

[7] A. Molloy and F. J. Rowe, "Manneristic behaviors of visually impaired children," Strabismus, Volume 19, Issue 3, 2011, pp.77–84.

[8] E. Fazzi, J. Lanners, S. Danova, O. Ferrarri-Ginevra, C. Gheza, A. Luparia, U. Balottin, and G. Lanzi, "Stereotyped behaviours in blind children," Brain and Development, vol. 21, no. 8, 1999, pp. 522–528.

[9] Language and Media Processing Laboratory, University of Maryland: "A Video Metadata Markup Tool: ViPER-GT," 2005. (v4.0) (2016, Feb 19) [Online] Available: http://viper-toolkit.sourceforge.net/products/gt/

[10] M. Mathias, R. Benenson, M. Pedersoli and L. Van Gool, "Face detection without bells and whistles," ECCV 2014, European Conference on Computer Vision, Zurich, CH, Sept. 6-12, 2014.

[11] Whitepaper NVIDIA Tegra 4 Family GPU Architecture (v1.0) (2013, Feb 22) [Online] Available: https://www.nvidia.cn/content/PDF/tegra_white_papers/Tegra_4_GPU_Whitepaper_FINALv2.pdf

[12] Xilinx inc. "Zynq-7000 All Programmable SoC Overview," datasheet DS190 (v1.9) (2016, Jan 20) [Online] Available: http://www.xilinx.com/support/documentation/data_sheets/ds190-Zynq-7000-Overview.pdf

[13] Dave Embedded System: "Bora: Xilinx Zinq XZ7Z010/ZC7Z020 CPU Module," (2016, Feb 20) [Online] Available: http://www.dave.eu/sites/default/files/files/bora-leaflet.pdf

[14] S. Jin, D. Kim, T. T. Nguyen, D. Kim, M. Kim, J. W. Jeon, "Design and Implementation of a Pipelined Datapath for High-Speed Face Detection Using FPGA," in Industrial Informatics, IEEE Transactions on, vol. 8, no. 1, 2012, pp. 158-167.

[15] C. Kyrkou, C. S. Bouganis, T. Theocharides, M. M. Polycarpou, "Embedded Hardware-Efficient Real-Time Classification With Cascade Support Vector Machines," in Neural Networks and Learning Systems, IEEE Transactions on , vol. 27, no.1, 2016, pp. 99-112.

[16] S. Liao, A. K. Jain and S. Z. Li, "A Fast and Accurate Unconstrained Face Detector," in IEEE Trans. on PAMI, vol. 38, n. 2, 2016.

[17] S. Liao, A. K. Jain and S. Z. Li, "A Fast and Accurate Unconstrained Face Detector," (2016, Feb 19) [Online] Available: http://www.openpr.org.cn/index.php/107-NPD-Face-Detector/View-details.html

[18] P. A.Viola and M. J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," accepted conference on computer vision and pattern recognition, 2001.

[19] P. A. Viola and M. J. Jones, "Robust real-time face detection," in International Journal of Computer Vision, vol. 57, no. 2, 2004, pp. 137-154.

[20] Open Source Computer Vision: "Face Detection using Haar Cascades," (2016, Feb 19) [Online] Available: http://docs.opencv.org/master/d7/d8b/tutorial_py_face_detection.html

[21] N. Markus, M. Frljak, I. S. Pandzic, J. Ahlberg and R. Forchheimer, "A method for object detection based on pixel intensity comparisons," unpublished.

[22] Nenad Markus: "pico," (2016, Feb 19) [Online] Available: https://github.com/nenadmarkus/pico

[23] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," in Naval Research Logistics Quarterly 2, 1955, pp. 83–97.

Impact of Light Conditions on the Vertical Traffic Signs Detection in Vertical Traffic Signs Recognition System

D. Solus*, Ľ. Ovseník * and J. Turán*
*Technical University of Košice/Department of Electronics and Multimedia Communications, Košice, Slovakia
david.solus@tuke.sk, lubos.ovsenik@tuke.sk, jan.turan@tuke.sk

Abstract – This paper is focused on impact of light conditions on the vertical traffic signs detection in Vertical Traffic Signs Recognition System (VTSRS). VTSRS contains two main phases – detection phase and recognition phase. The main task of detection phase is located in the area of the vertical traffic sign (ROI - Region of Interest) from the captured image of the real traffic. Detection phase might be based on three different methods – colour-based, shape-based and combination of these mentioned methods. The light conditions affect the success of each detection method. The detected ROIs are compared with the reference traffic sign by Optical Correlator in recognition phase. Several experiments have been done with each detection method in different light conditions. Results and conclusion are discussed.

I. INTRODUCTION

The traffic signs are an important part of road infrastructure, which perform a significant role in traffic regulation, visual guidance of vehicles or other road users on the traffic and providing of service information. They are installed at specific locations and designed to be easily recognized by drivers because their colour and shapes. Recognition of traffic signs has been a challenging problem for many years. Nowadays, there are several systems for the recognition of the traffic signs. In these systems digital image processing or intelligent techniques like neutral networks are used. The images are analyzed and processed by various techniques like colour segmentation, shape detection or edge detection. The procedure of the traffic signs recognition has usually two phases. The first is the localization of the Region of Interest - detection phase. The second phase is the traffic sign recognition – recognition phase.

Goal of this paper is to present Vertical Traffic Signs Recognition System, more specifically impact of light conditions on the vertical traffic detection. There are three ways of the vertical traffic detection – colour-based, shape-based and combination of both methods. So, the detection phase in VTSRS is based on these methods. Optical Correlator is used as a comparator in the recognition phase. This system was tested on the real traffic of Slovakia in different light conditions [6-13].

In Chapter II, the proposed system is presented. The particular procedure of VTSRS is shown (Figure 1.) and the individual steps are described. Experiments and results are involved in Chapter III. Experiments were done with the static images captured from video of real traffic in Slovakia. Conclusions are summarized in Chapter IV.

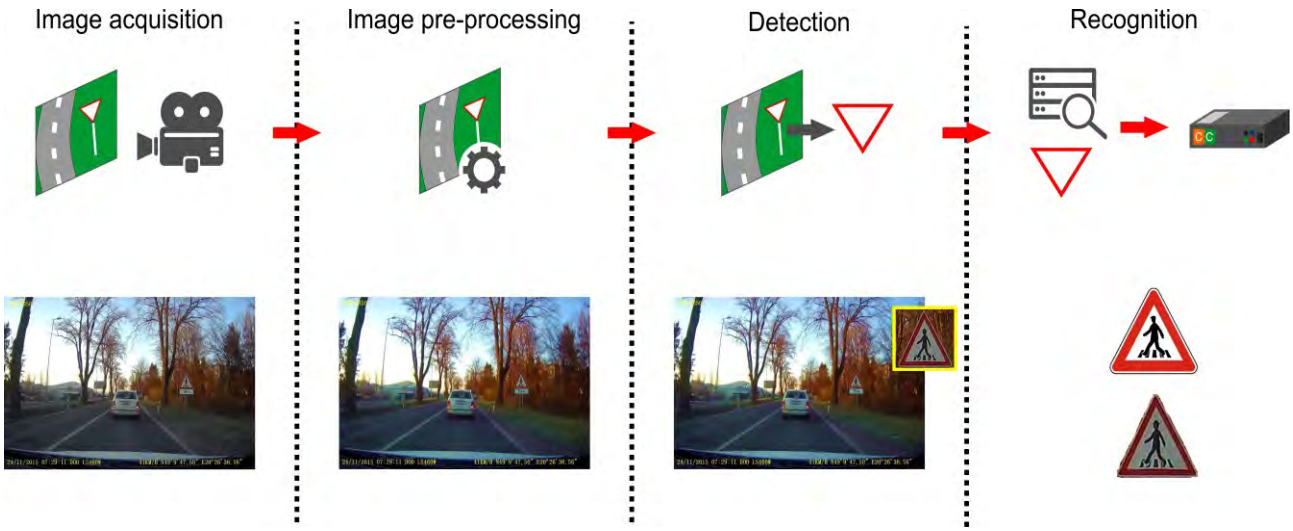


Figure 1. Procedure of VTSRS

II. VERTICAL TRAFFIC SIGNS RECOGNITION SYSTEM

The Vertical Traffic Signs Recognition System is a smart system designed to help drivers react properly in dangerous traffic situations. In our research, a system to detect and recognize vertical traffic signs was designed. VTSRS is mounted directly in vehicles and interprets traffic signs for driver. This system analyzed captured images of real traffic and provided recognizable vertical traffic sign to the driver. The VTSRS configuration is illustrates on Figure 1 [5-10].

A. Image Acquisition

The first section in VTSRS is the *Image Acquisition*. It is process of acquiring an image and to do this, the colour HD video camera is used. The colour HD video camera is installed into vehicle and it captures images of the scene front of the vehicle. Example of the captured image is shown on Figure 2.



Figure 2. Example of the captured image

B. Image Pre-processing

The next section after the *Image Acquisition* is the *Image Pre-processing*. The *Image Pre-processing* section is not a mandatory one, but it may increase the efficiency of the following sections. The fundamental function of the *Image Pre-processing* is to modify the captured image to easier obtain the necessary information from it (Figure 3.). The properties such as brightness, a contrast and a gamma are changed and to remove the noise from the image, a Gaussian filter is used. Thus the pre-processed image is prepared to the next section – *Detection* [10,11].



Figure 3. a) Original image, b) Pre-processed image

C. Detection

The main task of the *Detection* is to find the area of the traffic sign from the complicated captured image background. This area represents the *Region of Interest*, e.g. a region in which the traffic sign should be located. The output of this section represents real candidate of vertical traffic sign.

The traffic signs are designed to be principally distinguishable from natural or human-made backgrounds. They are characterized by many features which make them recognizable. They are designed in the standard geometrical shapes like triangles, circles, rectangles and octagons. The main colours of the traffic signs are chosen to be far away from the environment, so they are easily recognizable by the drivers. They use colours to represent the main information to the drivers. The typical colours used for the traffic signs are red, blue and yellow. So, information about shape and colour can be used in the detection. With this in mind, the detection can be based on shape, colour or both of them [5-10,13].

VTSRS can use three different detection methods – colour-based, shape-based and hybrid, which combine these mentioned methods together. Colour-based detection method is based on colour segmentation, e.g. process of partitioning image into multiple sets of pixels that have similar colour properties. In VTSRS, HSL colour space is used. The base of shape-based detection method are algorithms that are used to search some well-known traffic shapes of traffic signs. Of course the colour segmentation in this method is omitted. Hybrid detection method combines above mentioned methods together [12,13].

In all of these methods, digital image processing is used. We can divide digital image processing to simple operations that process image in different way. Each method uses different number of operations in different orders (Figure 4.) [10-15]. In VTSRS these following operations are used:

- *Grayscale* – process of the image transformation to a grayscale image. A grayscale image carries brightness information only. There is no information about a colour.
- *Thresholding* – the simplest method of image segmentation that converted gray scale image to binary image by a thresholding process. White pixels represent the pixels of the image which value is within the threshold range. Black pixels represent out of the threshold range value.
- *Dilatation* – the morphological operation that allows objects to expand, thus potentially filling in small holes and connecting disjoint objects.
- *Edge detection* – process of finding complex object boundaries by marking potential edge points corresponding to place in an image where rapid changes in brightness occur. After these edge points have been marked, they can be merged to form lines and object outlines. In our case, Canny edge detector is used.

- *Blobs filtration* – process of blobs removing, e.g. objects, which are smaller or bigger then specified limits.
- *Colour segmentation* – process of using colour segmentation based on red, blue and yellow colour for obtaining three simples images. These images contain only pixels, which refer to that colour.
- *Shape detection* – process of searching same well-known traffic sign's shapes.
- *Extraction of ROI* – process of selecting and bounding the Region of Interest. The largest object is bounded and subsequently extracted from image.

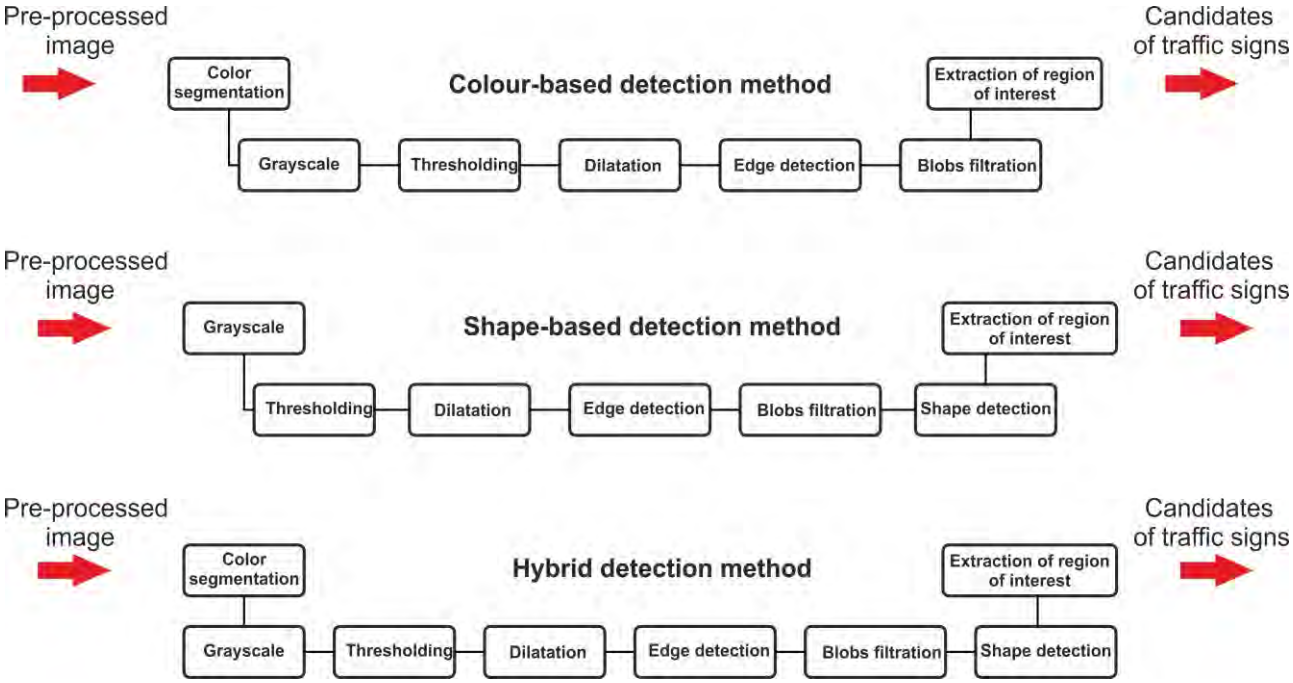


Figure 4. Procedure of detection methods

D. Recognition

The last section is *Recognition*. The output of the *Detection* is list of candidates that might represent the vertical traffic signs. This list is forwarded to the recognizer for further evaluation, and then based on the result of recognition this recognizer decided whether the candidates in the list are traffic signs or either rejected objects.

In VTSRS the *Cambridge Optical Correlator* is used as a recognizer. The *Cambridge Optical Correlator* is a basic device that uses the optical processing technology. It is a type of Joint Transform Correlator where process of optical correlation is formed by two consecutive Fourier transforms. The *Cambridge OC* is used to compare the input images with the reference images based on their similarities. The input scene is created by these images and then process of optical correlation is done. The optical output of the *Cambridge OC* contains highly localized intensities – correlation peaks, which size reflects measure of similarity of the images at the input scene. The position of these peaks is the same as the position of the input images [1-4].

III. Experiments and Results

Firstly, we created videos of real traffic in Slovakia. These videos capture the same road section, but in different light and weather conditions. We created two videos in early hours with ideal (Figure 5(a)) and low light conditions, mostly with fog (Figure 5(b)). The next videos were created in cloudy weather (Figure 5(c)) and at night (Figure 5(d)). Our experiments were realized with static images captured from created videos. Each captured image had to contain least one vertical traffic sign. We obtained about 38 static images from each video. After that, the database of reference vertical traffic signs was designed. The database includes only these vertical traffic signs that were found in created videos. Our experiments consist of two parts – *Detection* and *Recognition*.

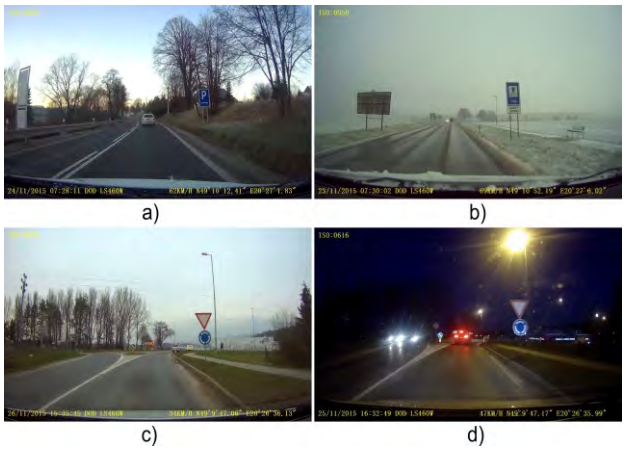


Figure 5. a) Early hours with ideal light conditions, b) Early hours with low light conditions (fog), c) Cloudy, d) Night

The main task of the *Detection* is to find real candidates of vertical traffic signs from captured images. These candidates represent the *Region of Interest*, e.g. an area in which the traffic sign should be located. As it mentioned above, VTSRS can use three different detection methods – colour-based, shape-based and hybrid. So, these detection methods were used to find candidates from 38 obtained static images of each videos. In this part of our experiments were obtained 38 candidates from each video by each detection method. Figure 6 contains reference traffic signs and some obtained candidates of traffic signs from video created in cloudy weather.



Figure 6. Example of traffic signs used in experiments

In the second part of our experiments, the *Recognition* is used to determine the content of the obtained candidates (38 images from each video by each method) which possibly contains one or more vertical traffic signs. The recognition process is carried out by the *Cambridge Optical Correlator* which function is to compare input and reference images situated in the input scene. In our case, the candidate represents the input image (Figure 6) and it is compared with the reference vertical traffic signs. The reference vertical traffic signs are stored in created database.

Cambridge Optical Correlator is coupled with simulation software, "*Fourier Optics Experimenter* (FOE)", to learning and easy understanding Fourier Optics, especially optical correlation. FOE allows make process of optical correlation based on *Cambridge Optical Correlator* between two or more images (Figure 7).

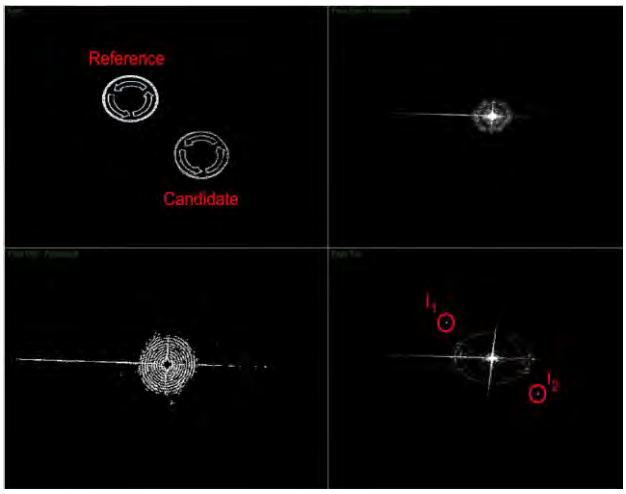


Figure 7. Process of optical correlation between two same images

In Figure 7 we can see process of optical correlation between candidate and reference traffic sign. The input scene (Figure 7(a)) is created by images mentioned above. Joint Power Spectrum (JPS) we can see in Figure 7(b) and this JPS binary or threshold processed in Figure 7(c). The optical output is showed in Figure 7(d). As was mentioned above, the optical output contains highly localized intensities and their value might be in within range <0;255> where value "255" refers to total match and value "0" refers to mismatch [1-4]. The equation (1) means percentage match between images situated in the input scene, where I_1 and I_2 are intensities of the correlation peaks

$$\text{Match} = (I_1 + I_2)/510 * 100. \quad (1)$$

We decided that if value of percentage match is greater than 70 %, the candidate is considered as vertical traffic sign. Values of intensities of correlation peaks of experiments shown in Figure 7 are $I_1 = 229$ and $I_2 = 214$. So according to equation (1) match of input images is 86,9%. From that fact match is 86,9% that candidate refers to reference traffic sign. The resulting values of the average intensities and percentage match between candidates obtained from each video by each method and reference traffic signs are shown in TABLE I.

TABLE I. RESULTING VALUES OF ANALYSIS

Video created in early hours with ideal light conditions						
Detection method	**Colour-based**		**Shape-based**		**Hybrid**	
Number of recognized traffic signs	26		26		31	
Average intensity value (all candidates)	146,1	57,3%	148,2	58,1%	164,5	64,5%
Percentage intensity value (only recognized)	188,5	73,9%	187,2	73,4%	190	74,5%
Video created in early hours with low light conditions (fog)						
Detection method	**Colour-based**		**Shape-based**		**Hybrid**	
Number of recognized traffic signs	23		25		28	
Average intensity value (all candidates)	140,8	55,2%	151	59,2%	148,9	58,4%
Percentage intensity value (only recognized)	182,6	71,6%	184,6	72,4%	184,6	72,4%

Video created in cloudy weather						
Detection method	**Colour-based**		**Shape-based**		**Hybrid**	
Number of recognized traffic signs	24		26		30	
Average intensity value (all candidates)	152,5	59,8%	153,8	60,3%	157,1	61,6%
Percentage intensity value (only recognized)	184,9	72,5%	183,9	72,1%	198,5	74,3%
Video created at night						
Detection method	**Colour-based**		**Shape-based**		**Hybrid**	
Number of recognized traffic signs	11		15		13	
Average intensity value (all candidates)	76,8	30,1%	80,8	31,7%	80,1	31,4
Percentage intensity value (only recognized)	187,9	73,7%	190	74,9%	191,8	75,2

IV. CONCLUSION

Vertical traffic signs recognition system is a system for vertical traffic signs detection and recognition. In this paper this VTSRS have been describes in previous chapters in detail. The lists of candidates of vertical traffic signs obtained from each video have been founded by three different methods. Each list contains 38 real candidates of vertical traffic sign. These candidates were compared with the reference vertical traffic signs stored in database by *Cambridge Optical Correlator*. The average values of intensities and percentage match of compared images were obtained. If the threshold was set to lower value than 70%, the number of recognized traffic signs will increase.

Each of detection method (colour-based, shape-based and hybrid) have different advantages and disadvantages. Because of the complex environment of the traffic and the scene around them, traffic signs can be founded in different conditions. The colour of the traffic signs fades with time by exposure of sun beams. The visibility of the traffic sign depends on the weather condition like rain, clouds, fog or snow and time of day or season. The objects in the scene like vehicles, buildings, billboards or others might be similar to traffic signs by colour, shape or booths. In addition to this, the surface of the traffic signs might be polluted, distorted, content unreadable or damaged. All of that had significant impact on detection and recognition.

The videos were obtained by HD colour video camera in different light and weather conditions. These conditions affected the success of each detection method in different way. So, each method might have different success rate in different light and weather conditions.

ACKNOWLEDGMENT

This work was supported by Cultural and Educational Grant Agency (KEGA) of the Ministry of Education, Science, Research and Sport of the Slovak Republic under the project no. „006TUKE-4/2014 - The Use of TUKE PON Experimental Model in Teaching. This work is also the result of the contract no. “APVV-0025-12 - Mitigation of Stochastic Effects in High-Bitrate All Optical Networks” supported by the Slovak Research and Development Agency.

REFERENCES

[1] P.Ambs, "Optical Computing, A 60-Year Adventure," in Advances in Optical Technologies Journal, 2010.

[2] Cambridge Correlators, http://www.cambridgecorrelators.com, october 2014.

[3] Cambridge Correlators, "Fourier Optics Experimenter," User Guide.

[4] J. W. Goodman, Introduction to Fourier Optics, Colorado: Roberts & Company.

[5] B. Cyganek, "Circular Road Signs Recognition with Soft Classifiers," Integrated Computer-Aided Engineering, Vol. 14, No. 4, pp. 323-343, 2007.

[6] B. V. K. V. Kumar, A. Mahalanobisbis, R. D. Juday, "Correlation Pattern Recognition," Cambridge University Press, ISBN 9780521571036, 2005.

[7] T. Harasthy, J. Turán, Ľ. Ovseník and K. Fazekas, "Optical correlator based Traffic Signs Recognition," International Conference on System, Signal and Image Processing, Vienna, Austria, 2012.

[8] S. Toth, E. Krsak, "Traffic Sign Recognition and Localization for Databases of Traffic Signs," Acta Electrotechnica et Informatica, vol. 11, no. 4 (2011), pp. 31-35.

[9] A. Broggi, P. Cerri, P. Medici, P. P. Porta and G. Ghisio, "Real Time Road Signs Recognition," Proceedings of the IEEE Intelligent Vehicles Symposium, Istanbul, Turkey, June 13-15, 2007.

[10] D. Soendroro, I. Supriana, “Traffic sign recognition with Colour-based Method, Shape-arc estimation on SVM,” in International Conference on Electrical Engineering and Informatics (ICEEI), pp. 1-6, 17-19 July 2011.

[11] C. Bahlmann, Y. Zhu, R. Visvanathan, M. Pellkofer, T. Koehler, “A system for traffic sign detection, tracking, and recognition using colour, shape, and motion information,” Proceedings. IEEE Intelligent Vehicles Symposium, 2005, pp. 255–260, 6–8 June 2005.

[12] Z. Song, B. Zhao, Z. Zhu, M. Wang, E. Mao, “Research on Recognition Methods for Traffic Signs,” FGCN '08. Second International Conference on Future Generation Communication and Networking, 2008, pp. 387–390, 13–15 Dec. 2008.

[13] S. Vitabile, A. Gentile, and F. Sorbello, "A neural network based automatic road sign recognizer," presented at The 2002 Inter. Joint Conf. on Neural Networks, Honolulu, HI, USA, 2002.

Wound detection and reconstruction using RGB-D camera

Damir Filko, Emmanuel Karlo Nyarko and Robert Cupec
Faculty of Electrical Engineering, Osijek, Croatia
damir.filko@etfos.hr, nyarko@etfos.hr, robert.cupec@etfos.hr

Abstract - The advent of inexpensive RGB-D sensors pioneered by the original Kinect sensor, has paved the way for a lot of innovations in computer and robot vision applications. In this article, we propose a system which uses the new Kinect 2 sensor in a medical application for the purpose of detection and 3D reconstruction of chronic wounds. Wound detection is based on a per block classification of wound tissue using color histograms and the nearest neighbor approach. The 3D reconstruction is similar to KinectFusion where ICP is used for determining the rigid body transformation, color enhanced TSDF is applied for scene fusion, while the marching cubes algorithm is used for creating a surface mesh. The entire system is implemented in CUDA which enables real-time operation. The end result of the developed system is a precise 3D colored model which can be used for determining a correct therapy and treatment of chronic wounds.

I. INTRODUCTION

Chronic wounds heal very slowly and the healing process may further be prolonged if an ineffective treatment is used. Clinicians need an objective wound characterization method to decide if the current treatment is adequate or require adjustments. Accurate wound measurement is an important task in chronic wound treatment, because changes in the physical parameters of the wound are indicators of the healing progress.

Wound measurement methods currently in use are mostly based on simple approaches for area measurement, using rulers and transparency tracing, or photography based non-contact systems. Such methods give a rough estimate of the total wound area and the evaluation depends on human experience; it is not objective and not eligible for validation of the wound healing process.

The first step in creating a precise non-contact measurement system is having the ability to reconstruct the wound as a 3D model. This would facilitate complex analysis or measurement of wounds which is otherwise awkward and painful for the patient or inconvenient for clinicians. Furthermore, the reconstructed 3D model could be used for collaboration between clinicians and telemedicine.

Non-contact wound measurement and 3D reconstruction systems have been developed before, utilizing lasers or photography with multiple view geometry algorithms. This results in very expensive systems in the case of the former, or less precise and noise prone systems in the case of the latter. With the advent of inexpensive RGB-D sensors, the computer vision community has gained an easier way to innovate and create applications in many fields including medicine.

In this paper, we propose a system for the detection and 3D reconstruction of chronic wounds using a Kinect 2 sensor. The developed system automatically detects wounds by analyzing image blocks and utilizing color histogram similarity and nearest neighbor approach. The 3D reconstruction is similar to the KinectFusion approach [1] where Iterative Closest Point algorithm is used for determining rigid body transformation. Color enhanced Truncated Signed Distance Function is applied for scene fusion and the Marching cubes algorithm is used for creating a surface mesh. Furthermore, the entire system is implemented in CUDA which enables real-time operation. The proposed system provides a high resolution 3D colored model of the wound and its surrounding tissue suitable for further analysis. This could result in more responsive therapy and treatment for chronic wounds. The 3D models could also be efficiently used for telemedicine, thereby improving clinician collaboration and enhancing the treatment of patients.

The rest of the paper is structured as follows. In Section 2 a short survey of related research is given. Section 3 provides an overview of the detection subsystem, while Section 4 elaborates the 3D reconstruction subsystem. Finally, this paper is concluded with Section 5, which comments on the results and future work.

II. RELATED RESEARCH

The majority of published research focuses on wound assessment by measuring physical attributes such as depth, perimeter, area and volume. However, before the measurement and evaluation phase, wound capture or reconstruction needs to be performed, which is the topic of the research presented in this paper.

Wound measurement is typically divided into two main approaches: contact and non-contact. Contact based methods require the measurements to be taken directly with wound contact [2, 3]. Rulers or transparent films are used in such situations for measuring depth, perimeter and area, while fluid or similar material is used for volume measurement. The main disadvantages of using contact based methods are imprecision and subjective interpretation. Contact based methods also increase the

possibility of wound infections and are typically not comfortable for patients.

Non-contact methods require a recording device, whereby a standard camera is typically used. For example, in [4] the wound is manually marked and tissue type percentages calculated, while area and perimeter measurements require that the ruler be visible in the image. Newer systems which use standard camera automatically segment the wound [5, 6]. It is generally accepted that image based methods have a maximum error of area measurement of about 10% depending on the wound location and the camera angle. This fact is experimentally proven in several studies including [7] where the authors also conclude that image based measurement is a viable alternative to contact based methods such as VisiTrak.

Other wound assessment systems use 3D reconstruction to enhance image based methods and gain more precise measurement. Therefore, using multiple view geometry algorithms with standard cameras is a typical approach. In [8], the authors use two wound images taken from different angles to generate a 3D mesh model. Because of the technology and algorithms used, the resulting 3D mesh has a low resolution. Eykona, a system similar in principle and commercially available, used in the study by [9] generates more dense 3D representation. Besides being very expensive, costing about $18000, Eykona has limitations regarding the size and position of the wound it can reconstruct because the whole wound needs to be visible in the single frame. Eykona also requires the use of a disposable optical marker which must be located in the image. Besides Eykona, there are other commercially available 3D wound measurement systems based on stereo vision such as MAVIS and 3D LifeViz, which are equally expensive and cost about $17000.

Lasers are also commonly used for 3D reconstruction in medical research, where a laser line projection sensor calibrated with an RGB camera can produce precise and colored 3D reconstructions. One of the first studies which used such a system was Derma [10], where the authors used the expensive Minolta VI910 scanner. Other similar research also used a combination of laser and RGB camera [11, 12]. Such systems have proven to be very precise, but are also not easy to use and very expensive. Furthermore, they also share the limitation that stereo vision systems have: the whole wound must be visible in one frame. One commercially available system for wound measurement which uses laser technology is SilhouetteStar by Aranz Medical which costs about $8000. Unfortunately, the system only provides measurement results and does not provide a reconstructed 3D model which can further be analyzed.

RGB-D camera is also used in one commercial system for wound measurement but is not yet available to the general public. InSight by eKare uses a technology similar to the original Microsoft Kinect and is used in the study by [13] for comparison in measurement performance with SilhouetteStar. The measurement shows comparable results except in wound analysis time which takes 45s for InSight and approximately 2 minutes for SilhouetteStar to produce results. The new InSight system also has the limitation regarding the visibility of the whole wound in a single frame.

As it can be seen, most of the systems have limitations regarding precision, ease of use and price, the latter of which limits the widespread use of most systems. However, the biggest limitation of all the presented systems is that the whole wound must be visible from a single view. This constraint limits the wound's size and location. The 3D wound reconstruction system proposed in this paper doesn't have such a constraint. Since it's based on scene fusion, the wound can be located on a larger area and on curved surfaces. Furthermore, an inexpensive Microsoft Kinect 2 sensor is used which costs about $150 which should make the system available to a much wider user base.

The system was fully developed in C++ and CUDA v7.0. Furthermore, all experiments were run on a laptop computer with Intel Core i7 4710HQ CPU, Nvidia 860M GPU and Windows 8.1 Pro x64 operating system.

III. WOUND DETECTON

The main goal of the wound detection subsystem is to find the center of the wound the user wishes to reconstruct. This is to facilitate the generation of high resolution local 3D models of the selected wound without the need to reconstruct the whole scene in front of the RGB-D sensor.

A basic requirement of this subsystem is that it should be fast and robust. The system needs to detect and mark regions on the image as potential wounds, with the user confirming the position of the wound. Thus, even though the system might detect several possible regions and mark them as wounds, some of which might be falsely classified, the system *should always* detect the region where the wound is located.

The system is based on color histograms and *k*-Nearest Neighbour algorithm (kNN): a color histogram of a potential region is created and the region classified with the aid of kNN algorithm. To this end, an initial database of classified regions is created and used in the classification process. This database is referred to herein as the reference database. Using an initial set of images containing wounds, differently sized regions of the images are cropped. Two classes of cropped images are generated: those that consist of wounds and those without wounds. Color histograms are then created for these cropped images. Thus, for any region of a test image that needs to be classified, a color histogram of the region is created, and then classified against the reference database using kNN.

Figure 1. shows an example image with cropped regions. 12 regions are cropped: 6 regions not containing wounds (marked in black) and 6 regions containing wounds (marked in green).

A total of 5 images were used in creating the reference database. Using these 5 images, 97 cropped regions containing wounds were generated while 124 cropped regions without wounds were generated. These 221

cropped images served as the reference database against which all test images were compared.

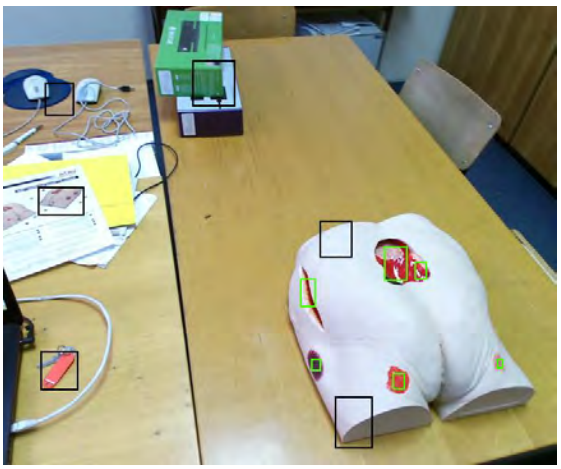


Figure 1. Example image with marked regions used in creating the reference database.

Two color spaces were tested for the purpose of the application. The Lab and HSV color spaces, in particular, the *a* and *b* channels of the Lab color space and the *H* and *S* channels of the HSV color space. Since 64 bins were used per channel, a 128 valued vector (histogram) was generated for each cropped image and stored in the reference database.

A flowchart of the wound detection process is shown in Figure 2. The image is divided into a uniform grid of 16 x 16 square cells or blocks. A color histogram is generated for each block and with the aid of the kNN algorithm (k = 7) and the reference database, each image block is classified.

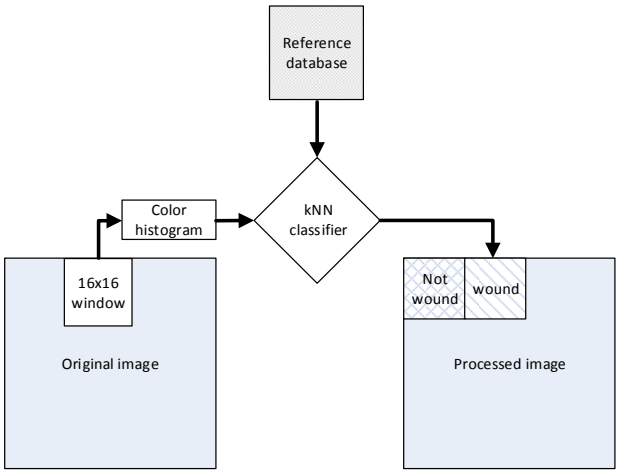
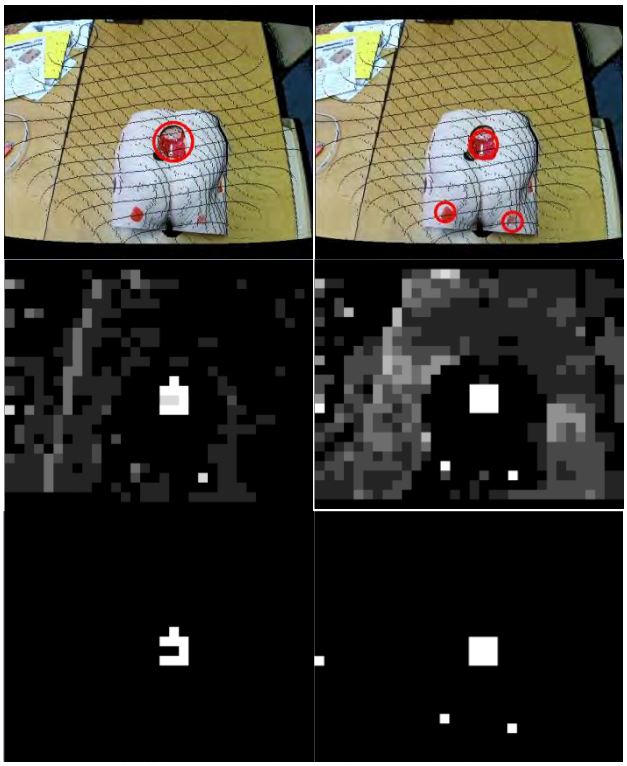


Figure 2. Flow chart of the wound detection process.

For a given test histogram (or image block), the kNN search algorithm returns the *k* vectors in the reference database that are closest to the test histogram. The Bhattacharyya distance is used to calculate the similarity between the test and reference histograms. The probability of the test histogram being classified as a wound can be defined as the ratio of the number, *n*, of *k*-nearest vectors classified as wound divided by *k*. As a result, each image block (or sub-image) of the original image is assigned a probability of being a wound. At the end of the classification process, a threshold is applied and only those regions with a probability greater than the threshold are marked as wounds. A threshold of 0.89 was used in all the experiments. Contours are then drawn around the regions marked as wounds in order to ease the selection of a region.



a) *Lab* color histogram used b) *HSV* color histogram used

Figure 3. Test image; (top row) images with regions marked as wounds; (middle row) result of per block classification obtained using kNN and reference database showing the probability of a block being a wound; (bottom row) result of thresholding the image obtained in the middle row.

Figure 3. shows an example test image and the results obtained using the Lab and HSV color histograms. Experimental results indicate that using HSV color histogram gives better classification results. The CUDA implementation of the wound detection subsystem has an average performance of 8.68ms per frame.

IV. WOUND RECONSTRUCTION

The methodology used in the proposed system for 3D reconstruction of wounds is basically the same as the one used in the KinectFusion system. However, for the considered application, models in a higher resolution than that provided by KinectFusion are required. Since the source code of KinectFusion is not publically available, we could not adapt it for our purpose and therefore we developed our own 3D reconstruction subsystem. The 3D reconstruction subsystem is divided into two phases: scene fusion and mesh generation.

A. Scene fusion

Scene fusion is the process where each captured RGB-D frame is integrated into a reconstructed volume. This process is further divided into two sub-phases: registration and integration. Registration determines the rigid body

transformation between frames and is based on the Iterative Closest Point (ICP) algorithm. ICP is a widely used algorithm for determining rigid body transformation of objects comprised of point clouds or meshes. Input to ICP are typically two point clouds or mesh samples and the initial guess of their relative rigid body transformation. The ICP then iteratively refines the transformation by repeatedly generating corresponding pairs of closest points and minimizing an error metric.

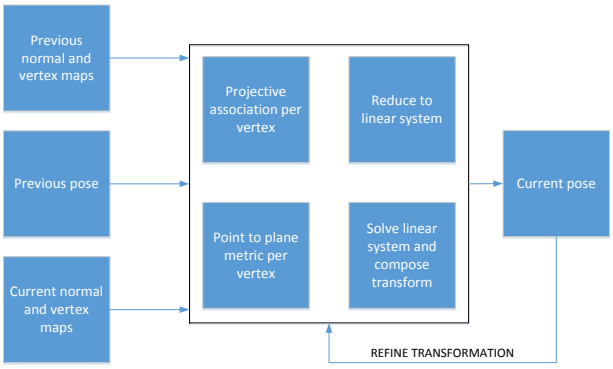


Figure 4. ICPCUDA work diagram

A slightly modified ICPCUDA implementation [14] is used for frame to frame registration. A modification to the implementation was introduced only to enable the use of Kinect 2 data properly. The ICPCUDA work diagram can be seen in Figure 4. The registration sub-phase performance was on average 8.73ms per frame.

The integration sub-phase is based on a TSDF. Signed distance function (SDF) is a non-parametric surface representation often used in computer graphics. It was introduced in [15], primarily for fusing depth maps. It is based on discretizing the volume into voxels of the same size. Each voxel is assigned an SDF value. Voxels with a positive SDF value are considered to be outside the object surface, while those with negative values are considered to be inside the object. Therefore, the object surface is the contour with the iso-value of 0. By truncating the SDF value to the interval [-1, 1], the truncated SDF (TSDF) is obtained. Figure 5 shows raw TSDF values from -1 (red) to 1 (blue) at voxel centers with a generated iso-surface of value 0.

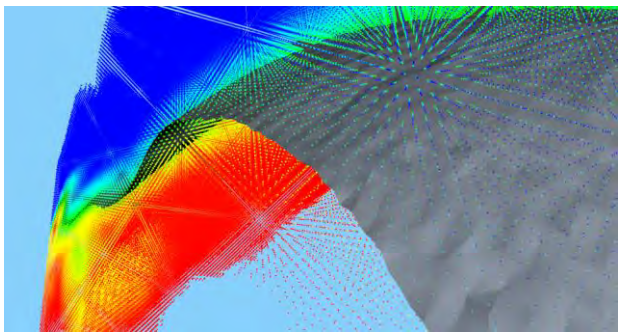
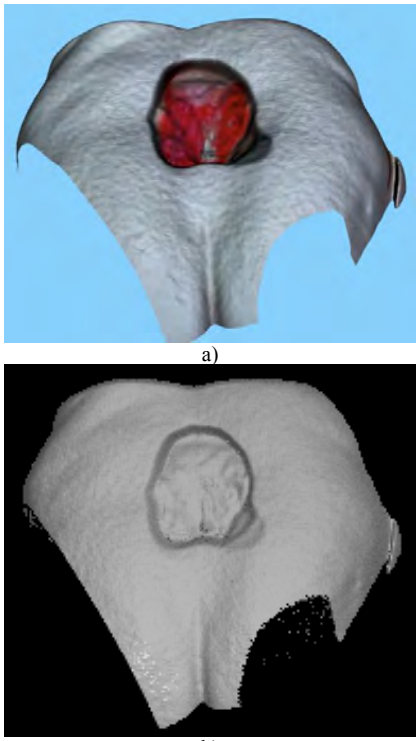


Figure 5. Raw TSDF with iso-surface generated by marching cubes algorithm

The integration sub-phase performance was on average 17.86ms per frame for a 256x256x256 volume of voxels.

B. *Mesh generation*

Given the TSDF volume gained by the scene fusion phase, there are two ways of obtaining a view or a rendering of the surface. The first option is to use an iso-surface generation algorithm such as the Marching cubes algorithm which returns a connected surface comprised of vertices and triangles (Figure 6a.). The other option is to use ray casting from some camera position directly onto the volume i.e. calculating the depth values of the zero crossing for rays originating from the camera position for each pixel of the rendered image (Figure 6b.).



a)

b)

Figure 6. TSDF surface rendering from similar viewpoint at iso-value 0 by a)marching cubes algorithm, b)ray casting

Images rendered by ray casing are very useful for determining the state of wound reconstruction and is therefore used during recording. TSDF ray cast rendering performance was on average 3.47ms per frame. The Marching cubes algorithm is therefore used for generating a high resolution 3D model which is the result of the system presented in this paper.

Marching cubes is the algorithm for polygonization of 3D scalar field at a certain level i.e. iso-surface construction. It was originally developed by Lorensen and Cline for visualization of CT and MRI scans [16]. Today, it is a standard algorithm for surface reconstruction. The algorithm is based on analyzing each voxel (element of a 3D scalar field) independently and determining which voxel vertex is within the predefined surface border value. As a result, for a given voxel, the algorithm determines which triangles to create from a finite number of combinations. Since there are 8 vertices per voxel, there are 2^8=256 possible triangle combinations. All 256 combinations are predefined in the implementation code which means that determining which triangle to create is very fast. Some of the possible triangle combinations

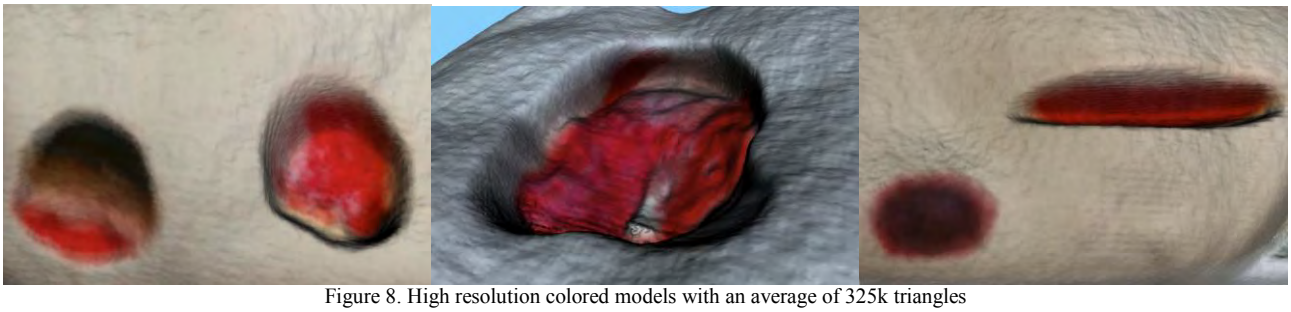


Figure 8. High resolution colored models with an average of 325k triangles

within a voxel can be seen in Figure 7, where the orange spheres are vertices having values which places them inside the surface.

Marching cubes CUDA implementation performance was on average 125ms for a 256x256x256 TSDF volume which typically resulted in 3D models with about 325k triangles. Figure 8 displays reconstructed high resolution colored 3D models for all the wounds on the Saymour II wound care model by VATA Inc.

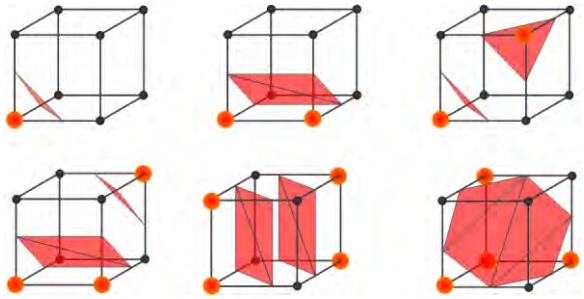


Figure 7. 6 out of 256 possible triangle combinations for marching cubes algorithm

V. Conlusion and Future Work

Chronic wounds are a big problem today in national health systems primarily due to expensive and prolonged treatment of patients which are often caused by incorrect or ineffective therapy. In this paper, a system for automatic wound detection and 3D reconstruction is presented which enables detailed wound analysis and, indirectly, improved treatment for patients. The system consists of two subsystems: wound detection and wound reconstruction. The main goal of the detection subsystem is to find the center of the wound in order to facilitate the generation of high resolution local 3D models. Unfortunately, due to a block-wise analysis the detected center can be unstable on a frame to frame basis. Improvement could be made by using a sliding window approach and a per-pixel score representation instead of block-wise classification. Furthermore, a larger reference database and a more sophisticated classifier such as support vector machine (SVM) or neural network could be used. The reconstruction subsystem results in well-defined and watertight colored 3D mesh models which are suitable for telemedicine and further analysis and measurement. The reconstruction subsystem can further be improved by increasing the precision of the ICP algorithm implementation. Optimization improvement might also open the possibilities of even higher resolution models. Future work on this system includes the development of a fully automatic wound segmentation subsystem which would permit automatic measurement of the perimeter, area, volume and depth of chronic wounds.

Acknowledgment

This work was supported by the Josip Juraj Strossmayer University of Osijek, under grant No. IZIP-2014-70.

References

[1] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, “KinectFusion: Real-time dense surface mapping and tracking,” in Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on, 2011, pp. 127–136.

[2] G. Gethin and S. Cowman, "Wound measurement comparing the use of acetate tracings and Visitrak digital planimetry", Jornal of Clinical Nursing, vol. 15, no. 4, pp. 422-427, 2005.

[3] T. Gilman, "Wounds Outcomes: The Utility of surface Measures", Lower Extremity Wounds, vol. 3, no. 3, pp. 125-132, 2004.

[4] D. Filko, D. Antonic and D. Huljev: “WITA - Application for wound analysis and management”, 12th IEEE International Conference on e-Health Networking Applications and Services, pp. 68-73, Lyon, France, 2010.

[5] R. Mukherjee, D. D. Manohar, D. K. Das, A. Achar, A. Mitra, and C. Chakraborty, “Automated Tissue Classification Framework for Reproducible Chronic Wound Assessment,” BioMed Research International, vol. 2014, pp. 1–9, 2014.

[6] C. Wang, X. Yan, M. Smith, K. Kochhar, M. Rubin, S. M. Warren, J. Wrobel, and H. Lee, “A unified framework for automatic wound segmentation and analysis with deep convolutional neural networks,” in Engineering in Medicine and Biology Society (EMBC), 37th Annual International Conference of the IEEE, pp. 2415–2418., 2015.

[7] A. C. Chang, B. Dearman, and J. E. Greenwood, “A comparison of wound area measurement techniques: visitrak versus photography,” Eplasty, vol. 11, 2011.

[8] S. Treuillet, B. Albouy and Y. Lucas, “Three-dimensional assesment of skin wounds using a standard digital camera”, IEEE Transactions in Medical Imaging, vol. 28, no. 5, pp. 752-762, May 2009.

[9] F. L. Bowling, L. King, J. A. Paterson, J. Hu, B. A. Lipsky, D. R. Matthews, and A. J. M. Boulton, “Remote assessment of diabetic foot ulcers using a novel wound imaging system: Remote foot ulcer assessment using a wound imaging system,” Wound Repair and Regeneration, vol. 19, no. 1, pp. 25–30, Jan. 2011.

[10] M. Callieri, P. Cignoni, P. Pingi, R. Scopigno, M. Coluccia, G. Gaggio, and M. N. Romanelli, “Derma: Monitoring the Evolution of Skin Lesions with a 3D System.,” in VMV, pp. 167–174., 2003.

[11] F. Zvietcovich, B. Castañeda, B. Valencia, and A. Llanos-Cuentas, “A 3D assessment tool for accurate volume measurement for monitoring the evolution of cutaneous leishmaniasis wounds,” in Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, pp. 2025–2028., 2012.

[12] U. Pavlovčič, J. Diaci, J. Možina, and M. Jezeršek, “Wound perimeter, area, and volume measurement based on laser 3D and

color acquisition," BioMedical Engineering OnLine, vol. 14, no. 1, Dec. 2015.

[13] J. D. Bills, S. J. Berriman, D. L. Noble, L. A. Lavery, and K. E. Davis, "Pilot study to evaluate a novel three-dimensional wound measurement device: Three-dimensional wound assessment tool," International Wound Journal, Nov. 2015., In press

[14] T. Whelan, https://github.com/mp3guy/ICPCUDA, downloaded on 28. April 2015.

[15] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pp. 303-312., 1996.

[16] W. E. Lorensen and H. E. Cline, „Marching Cubes: A high resolution 3D surface construction algorithm", SIGGRAPH Computer Graphics, Vol. 21, No. 4, July 1987.

Clustering of Affective Dimensions in Pictures: An exploratory analysis of the NAPS database

Marko Horvat*, Katarzyna Jednoróg** and Artur Marchewka***
* Zagreb University of Applied Sciences, Department of Computer Science and Information Technology, Zagreb, Croatia
** Nencki Institute of Experimental Biology, Laboratory of Psychophysiology, Warsaw, Poland
*** Nencki Institute of Experimental Biology, Laboratory of Brain Imaging, Warsaw, Poland
marko.horvat@tvz.hr

Abstract - Photographs and other multimedia documents can easily elicit strong emotional reactions. While knowledge discovery based on semantics and visual properties of pictures is considerably explored, little attention is given to mining of affective features of visual stimuli. We report result from clustering of 1356 stimuli from the Nencki Affective Picture System database using two-dimensional (valence/arousal) model of affect and *k*-means unsupervised learning algorithm. Clustering with $k = 1 - 94$ centroids is reported, together with dominant picture keywords and descriptive statistical parameters. Optimal number of centroids was estimated using minimum cumulative error rule. A custom Java application integrated with WEKA machine learning software was developed for the study. A practical goal of this research is to develop an intelligent expert system that could infer emotion content of multimedia from evaluation of its semantics, and vice versa, estimate dominant semantics from affect if such information is available. The system could potentially have many useful applications such as supported construction of affective multimedia databases, video recommendation or emotion estimation. Reported information on clustering of emotions is essential for success of such a system. The results are freely available by contacting the first author.

I. INTRODUCTION

Emotions are ubiquitously inherent to multimedia whether in a form of brief and intensive sentiments about a particular topic expressed in a written format, a longer lasting mood caused by watching an impressionable movie, hearing vocal expressions in a video-clip, or by eliciting emotions by looking at a picture. Any multimedia file can evoke emotions of varying polarity, variability, intensity and duration [1].

Multimedia documents with prior semantic and emotion content description are usually stored in affective multimedia databases and are explicitly intended for inducing emotional reactions. Because of their purpose such affective multimedia documents are also referred to as stimuli. Importantly, affective multimedia databases are standardized which allows stimuli to be used in a controllable and predictable manner [2]. The affective multimedia databases can be distinguished from other multimedia repositories foremost by the purpose of their content (i.e. to purposely elicit specific emotional reactions), but also by the models employed to describe knowledge about emotions and by specific experimental protocols with which emotion metadata is carefully acquired and verified. Affective multimedia databases are particularly useful in research of emotion and attention [3], but they can also be employed in studies of stress-related mental disorders [4] and methods for their prevention or mitigation [5].

The Nencki Affective Picture System (NAPS) [6], together with its extensions [7][8], is currently the largest and the most recently developed database. It features significant data models improvements compared to previous databases such as semantic categories, multi-word stimuli annotations and also provides normative ratings in line with discrete and dimensional models of affect. Because of these advanced features the NAPS is suitable for exploratory analysis of stimuli clustering in affective multimedia databases.

The reported research in clustering of affective dimensions in pictures is motivated by the intention to develop a model formally describing affective multimedia emotion, semantics and their functional relationships. Because of the uncertain nature of the data such a model must support combined representation of logical and statistical knowledge. In this paradigm ontologies are the preferred method for explicitly representing the meaning of data, while the statistical approach helps to encode the regularities in the affective database content. Once developed the knowledge model could have many useful applications in human computer interfaces, affective computing and cognitive sciences such as in an intelligent decision support system for construction of affective multimedia databases, video recommendation or emotion estimation.

The remainder of this paper is organized as follows; the next section describes emotion models and the NAPS data structure. Section 3 explains data mining software developed specifically for this study and the *k*-means algorithm. Section 4 presents in detail the results of the clustering analysis. Finally, the last section concludes the paper and gives an outlook to future research which will be based on the clustering results reported here.

II. AFFECTIVE PICTURE DATASET

In cognitive and computer sciences two predominant theories are used to describe knowledge about affect: the discrete category and the dimensional model [9]. Values in the former are often referred to as emotion norms or basic emotions, while the latter model is also called the

pleasure arousal dominance emotional state model (PAD) [10] or the circumplex model of affect [11]. Both theories can effectively describe emotion in digital systems but are not mutually exclusive. All affective multimedia databases have been characterized according to at least one of these models [9], and for some – including the NAPS – normative ratings collected in line with both models are available [6]. The dimensional theories of emotion propose that affective meaning can be well characterized by a small number of dimensions. The dimensions are chosen on their ability to statistically characterize subjective emotional ratings with the least number of dimensions possible [12].

Importantly, an explicit transformation operator between the two models does not exist. An emotion value can be transposed from one model to the other only with statistical inference [13]. Therefore, databases providing both discrete and dimensional annotations are important tools for research and utilization of relationship between emotions in digital systems.

The dimensional model is built usually around three emotion dimensions: valence (*Val*), arousal (*Ar*) and dominance (*Dom*). All three are described with continuous variables $val \in [1,9] \in Val$, $ar \in [1,9] \in Ar$, $dom \in [1,9] \in Dom$. Usually only the first two are used because dominance is the least informative measure of the elicited affect [9]. Subsequently, for all practical purposes, a single multimedia document can be represented as a coordinate in a two-dimensional emotion space $\Omega_{Emo} = Val \times Ar$. Therefore, documents in affective multimedia databases can be displayed as points defined with means of their respective valence and arousal annotated values as in Figure 1.

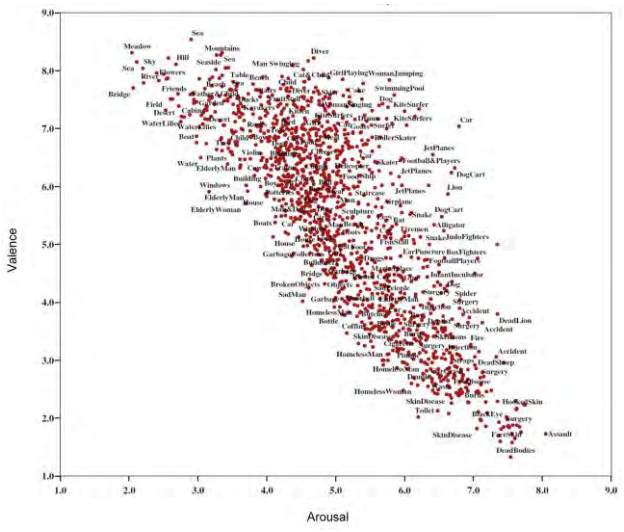


Figure 1. The NAPS database pictures displayed in Ω_{Emo} with annotated keywords. Adapted from [6].

The emotion values of pictures are acquired with rigidly controlled experiments in which participants express, most commonly in a form of a questionnaire, their subjective beliefs in elicited emotions. For example, during development of IAPS subjective ratings from around one hundred participants have been acquired for each picture and statistically analyzed to acquire aggregated mean and standard deviation of emotional dimension values [3]. Other multimedia databases follow the same pattern.

Our previous research in affective multimedia databases has demonstrated that stimuli semantics and emotion are coupled [13]. A given meaning, or semantic content, of pictures is aligned with specific emotion values. In other words, perceived semantics will elicit certain sets of emotions. Analysis of clustering in dimensional emotions is a step toward a deeper understanding of statistical couplings in affective multimedia databases.

III. CLUSTERING METHOD AND TOOLS

For the experiment a custom desktop computer application was developed in Java programming language. The application has a simple graphical user interface implementing only features important for clustering analyses of valence-arousal emotion space in affective multimedia databases. The main features are: 1) affective picture data import, 2) *k*-means clustering and determining sum of squared errors for any value of *k*, 3), analysis of maximum total error, maximum variance and weight factor, 4) plotting of picture clusters and different data graphs, 5) data output in a numerical format.

The tool can process three NAPS datasets, for men, women and both sexes, describing valence and arousal values of pictures and, for that matter, any other compatible database file in CSV format (e.g. physical parameters of the stimuli such as luminance or complexity). The visualization layer relies on jFreeChart Java component for rendering of clusters and graphs [14]. The display can be configured by user or captured as a picture file. However, with regards to functionality the most important part are components for data processing and clustering. These components were imported from the Waikato Environment for Knowledge Analysis (WEKA) which is a widely used open-source software suite for knowledge discovery and data mining [15]. The WEKA suite includes Java-based implementations of many algorithms for supervised, unsupervised and semi-supervised learning, data analysis, classification, clustering and data modeling [15]. Apart from visualization the obtained data can also be saved in a formatted text file for analysis. The building software components were intentionally chosen so as to increase the application's reliability, modularity and commonality with other data analysis tools.

The *k*-means uses the squared Euclidean distance to allocate objects to clusters [16]. The implicit assumption is that the data should have similar scale which is compatible with clustering in the NAPS since all picture have emotion values in the identical range and scale. The *k*-means algorithm tries to find a minimum for the Sum of Squares Error (SSE) [16]. In other words, SSE is the sum of the squared differences between each observation and its cluster's mean. It can be used as a measure of variation

within a cluster. For the entire set of objects, the SSE for k clusters is calculated by

$$SSE = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - c_j \right\| \quad (1)$$

where x_i is a data point (i.e. NAPS picture) in cluster C_j and c_j is a representative point – usually centroid – for the cluster C_j.

Hence, the goal of k-means is to find a solution with the lowest SSE. To find the optimal k, or more accurately the optimal range of values of parameter k, it is necessary to find the distribution of SSE for all relevant values of k.

IV. RESULTS

The SEE graph for $k = 1 – 94$ clusters is shown in Figure 2. All 1356 pictures in the NAPS database were used for the analysis based on their valence and arousal emotion dimensions [6]. NAPS BE annotations [7] and dominance dimension were not used.

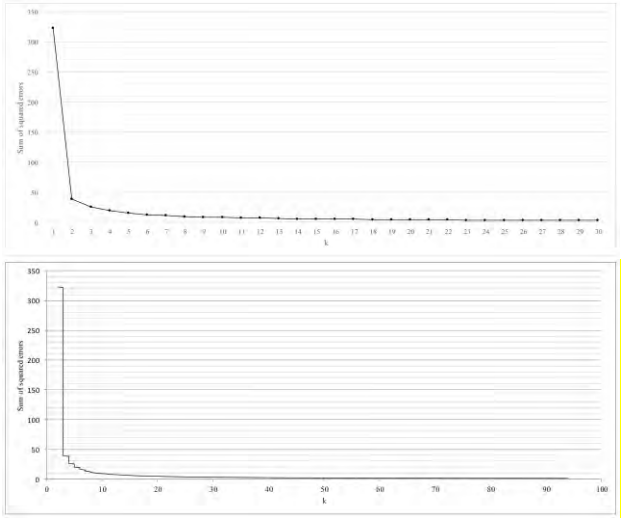


Figure 2. The sum of squares error (SSE) for $k = 1 – 30$ (above) and $k = 1 – 94$ clusters (below).

As can be expected the SEE decreases as k gets larger. The SSE starts with a relatively high value 322.66 for $k = 1$ (no clustering) and falls sharply with addition of clusters as 38.72, 25.34, 19.03, 15.19, 12.47, 10.74, 9.49, 8.55, 7.86, 7.24, 6.80, 6.25, 5.62, 5.25, 5.08 for $k = 2 – 16$. There is a noticeable change in the slope of the curve between $k = 2$ and $k = 10$. This trend in the curve gradient is extended to $k \sim 16$. After the inflection it can be assumed that the SSE asymptotically approaches ~2 with $k \to \infty$.

Another important information in the analysis is clusters' size. In Figure 3 the average size and variance of clusters for $k = 2 – 30$ is shown. As can be seen the average size of clusters, as well as their variance, linearly decreases as k gets larger. For heuristically chosen optimal range in the number of clusters $k = 2 – 16$ the average cluster sizes are 677.5, 451.67, 338.75, 271, 225.83, 193.57, 169.38, 150.56, 135.5, 123.18, 112.92, 104.23, 96.79, 90.33 and 84.69, respectively. Since each NAPS picture is tagged with one keyword expression, the same number of keywords will also be available for semantics analysis of clusters. The variances in the same range are 152.5, 70.04, 81.92, 56.20, 52.99, 54.02, 55.23, 40.02, 42.82, 33.59, 38.14, 39.53, 36.73, 27.91 and 32.32.

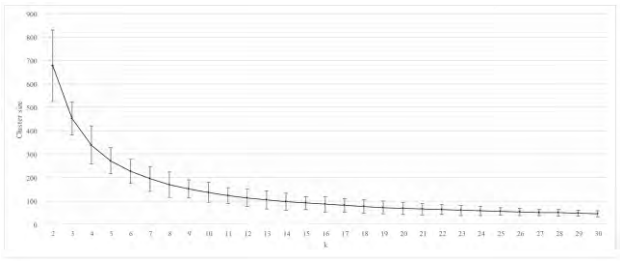


Figure 3. Average size and variance of clusters for $k = 2 – 30$.

Histograms of dimensional emotion values distribution (Figure 4) show that the densest data intervals (i.e. the highest bin counts) are in bins 5 and 6 with 264 and 335 pictures for valence, and 492 and 445 pictures for arousal, respectively.

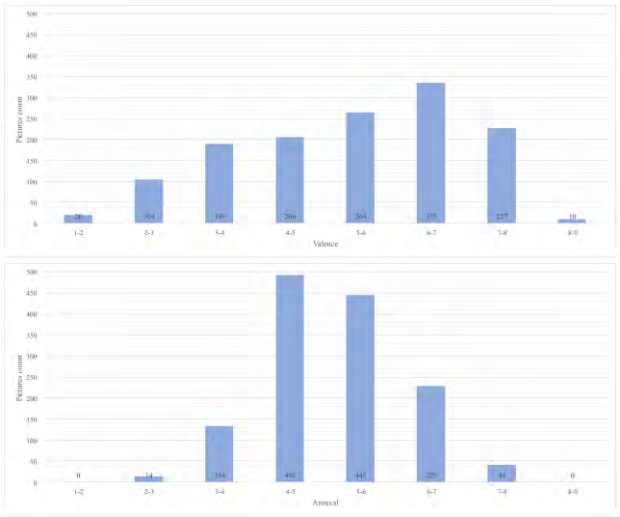


Figure 4. Frequency distribution of NAPS database pictures for valence (above) and arousal (below) emotion dimension.

The allocation of pictures over possible emotion values is important for density-based clustering but also might also affect the performance of k-means method. Interestingly, there are no pictures with very low and very high arousal, i.e. in bins 1 and 9. Also there is a very small proportion of pictures in arousal bins 2 and 8. This implies that a hypothetical classifier utilizing this dataset for statistical inference will not be able to make predictions for these bins or that the confidence in predictions will be weak. Pictures count in valence dimension is more even than in arousal, but also the majority of pictures are grouped around neutral valence values, i.e. between 4 and 6.

Distribution of NAPS pictures clusters in valence-arousal dimensional emotion space, for representative values of k, are shown in Figure 5 and the related most frequently occurring keywords are listed in Table 1.

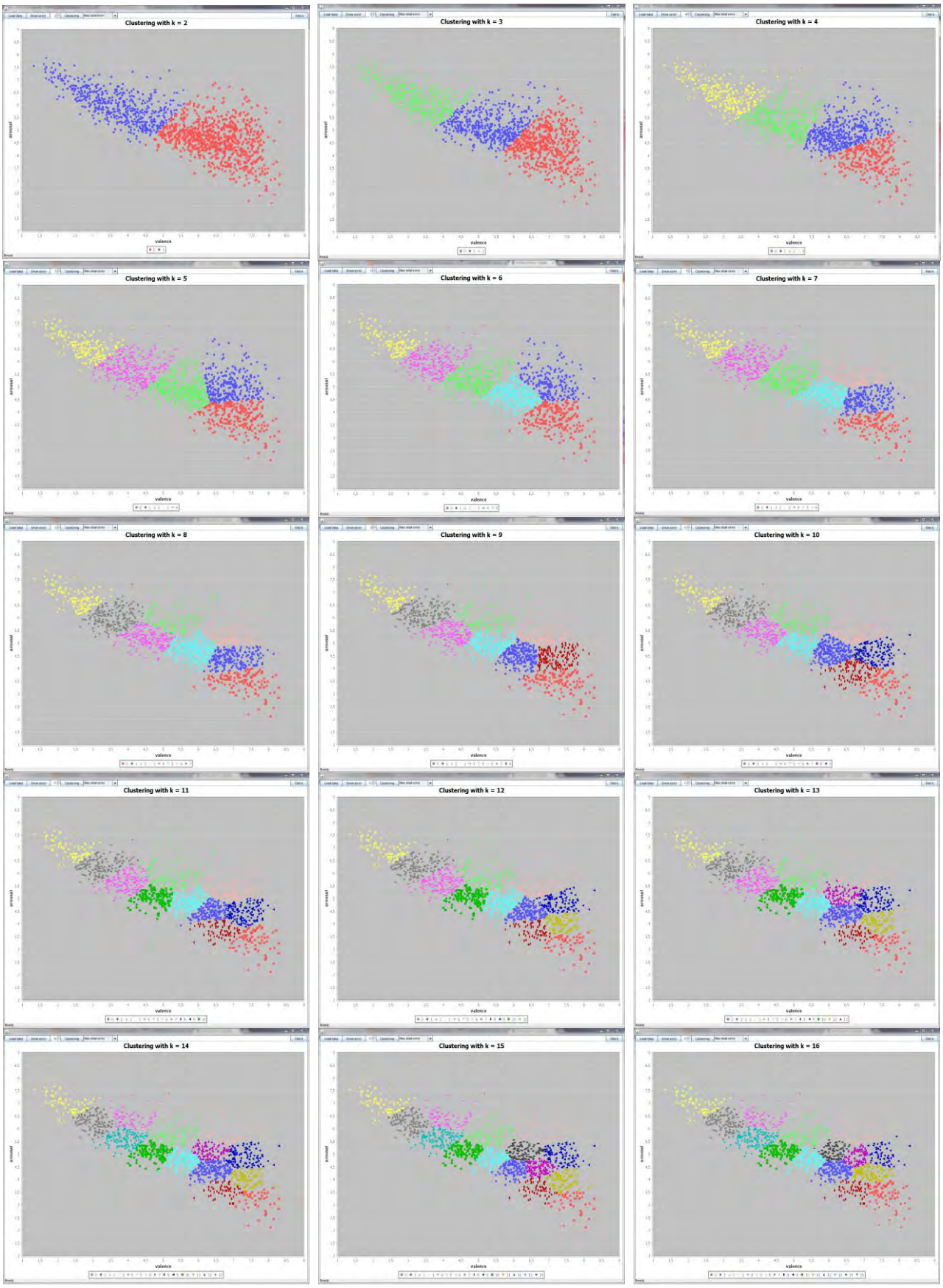


Figure 5. Clusters in the NAPS database's valence-arousal dimensional emotion space attained with *k*-means algorithm for $k = 2 – 16$. Clusters for $k = 2$ are presented in the upper-left and $k = 16$ in the lower-right corner of the figure.

The ten most frequently occurring NAPS keywords describing pictures in clusters c_i, $i = 1 - k$ and $k = 2 - 10$ are listed in Table 1. In total the table contains 510 keywords.

TABLE I. THE 10 MOST FREQUENTLY OCCURING KEYWORDS IN CLUSTERS

k	Cluster	Keywords
2	1	woman, man, elderly, girl, boy, children, fish, cat, dog, child
	2	dead, mutilated, man, homeless, car, surgery, child, elderly, snake, sick
3	1	woman, man, girl, elderly, boy, children, child, dog, cat, fish
	2	man, woman, elderly, snake, car, bicycle, baby, devastated, meal, cat
	3	dead, mutilated, surgery, homeless, car, man, child, accident, child, sick
4	1	girl, water, boy, woman, desert, fish, children, flowers, cat, plants
	2	woman, man, elderly, girl, boy, meal, dog, children, bicycle, car
	3	snake, man, woman, elderly, car, devastated, meat, child, mutilated, butcher
	4	dead, mutilated, homeless, surgery, man, accident, sad, child, sick, elderly
5	1	water, girl, desert, children, woman, flowers, cat, fish, plants, turtle
	2	woman, man, elderly, boy, girl, dog, children, child, people, meal
	3	man, woman, elderly, car, girl, bicycle, boy, meal, snake, cat
	4	dead, mutilated, surgery, homeless, accident, car, sick, sad, man, elderly
	5	mutilated, child, snake, man, woman, car, dead, butcher, homeless, meat
6	1	water, desert, girl, flowers, fish, woman, plants, cat, turtle, children
	2	woman, man, elderly, boy, girl, dog, children, meal, mother, people
	3	snake, man, elderly, woman, car, meat, cat, sea, baby, crashed
	4	dead, mutilated, surgery, accident, homeless, car, starved, child, sad, man
	5	mutilated, dead, homeless, child, man, car, child, surgery, sick, sad
	6	man, woman, elderly, girl, child, house, car, boy, bicycle, cat
7	1	desert, water, fish, flowers, plants, river, turtle, garden, girl, sea
	2	woman, children, girl, boy, man, child, dog, elderly, cat, baby
	3	snake, man, woman, elderly, meat, cat, sea, fish, child, baby
	4	dead, mutilated, surgery, accident, homeless, car, starved, child, sad, man
	5	mutilated, dead, homeless, man, car, surgery, sick, child, sad, meat
	6	man, woman, elderly, girl, boy, house, car, bicycle, couple, meal
	7	woman, elderly, man, girl, meal, car, snake, dog, mother, boys
8	1	desert, fish, water, flowers, river, plants, garden, turtle, girl, sea
	2	woman, girl, boy, child, children, man, elderly, dog, cat, baby
	3	snake, meat, lion, wolf, bat, small, spider, cat, baby, wounded
	4	dead, mutilated, surgery, accident, homeless, car, starved, child, sad, man
	5	man, woman, elderly, child, crashed, devastated, broken, car, sea, butcher
	6	man, woman, elderly, car, girl, bicycle, fish, boy, couple, house
	7	woman, elderly, man, girl, meal, boy, dog, mother, people, children
	8	mutilated, dead, homeless, sick, child, man, surgery, sad, meat, car
9	1	desert, water, fish, plants, flowers, river, garden, turtle, girl, sea
	2	woman, man, girl, elderly, house, child, meal, child, cow, cat
	3	snake, meat, lion, spider, wolf, bat, small, cat, stew, food
	4	dead, mutilated, surgery, accident, sad, man, homeless, starved, sick, child
	5	child, woman, devastated, man, butcher, crashed, broken, car, sick, mutilated
	6	man, elderly, woman, car, devastated, bicycle, fish, boy, garbage, cat
	7	woman, elderly, man, boy, meal, dog, mother, girl, car, dog
	8	mutilated, homeless, dead, surgery, man, sad, child, meat, car, sick
	9	woman, boy, girl, children, man, child, elderly, cat, dog, people
10	1	Desert, fish, flowers, river, plants, water, mountains, duck, turtle, girl
	2	Man, woman, elderly, girl, boy, bicycle, house, meal, child, couple
	3	Snake, meat, spider, wolf, bat, small, lion, cat ,baby, boy
	4	Dead, mutilated, surgery, accident, sad, man, homeless, starved, sick, child
	5	Child, woman, crashed, objects, child, sick, devastated, mutilated, butcher, broken
	6	Man, elderly, woman, car, devastated, garbage, sea, baby, shoes, bicycle
	7	Woman, elderly, man, boy, car, dog, mother, girl, boys, people
	8	Mutilated, homeless, dead, surgery, man, sad, child, car, elderly, meat
	9	Water, woman, cat, boy, girl, cow, child, children, turtle, butterfly
	10	Woman, man, girl, children, boy, elderly, dog, child, cat, people

As can be noticed in Table 1, as k gets larger semantics of clusters converges. This trend is particularly visible in clusters with negative valence and high arousal as they contain pictures with keyword such as “dead”, “mutilated”, “surgery”, “accident”, “homeless”, “car”, “starved” and “sad”. Negative semantics associated with disease, sickness, death, accidences and associated meanings are the most homogenized across all clusters listed in Tabled 1 and displayed in Figure 5. Clearly these semantic categories elicit the most similar negative emotional reactions in subjects. However, on the opposite side of valence-arousal scales, pleasant and mildly activating pictures converged in some individual clusters. Examples of such keywords are “flowers”, “plants”, “river”, “garden”, “girl”, “boys”, “mother” etc. With increase in k clusters with low arousal and semantics containing people, scenery and nature become noticeable as well.

For determining the optimal number of cluster we used the Elbow method [17]. In simple terms this heuristic approach seeks to find k at which the SSE decreases abruptly. However, since the goal of this research was to find the smallest set of clusters with which semantics and emotions in NAPS database could be brought into a meaningful statistical relationship, the Elbow method was enough for making preliminary conclusions.

Based just on dynamics of values and gradient of the SEE curve the optimal k is $k_{OPT} \in [2,16]$, with naively expected value (median) of $k_{OPT} = 9$. As can be seen in Figure 2, the SEE drops off most rapidly for $k = 2$ but

depending on the Elbow criteria the upper limit for k_{OPT} varies. However, it is also necessary to simultaneously consider picture keywords. Cluster semantics begins to homogenize with $k \sim 7$ which entails $k_{OPT} \geq 7$. This is in accordance with our estimate that $k_{OPT} = 9$. Nonetheless, the exact value of k_{OPT} should be determined with further research into coupling of semantics and emotion in the NAPS. We expect the semantic distance between keywords in a cluster to be important indicator of its homogeneity. Intuitively, semantically more uniform clusters should be better differentiated, subsequently leading to more accurate prediction of emotion from semantics, and vice versa.

Because of the paper format limitations, the complete set of results could not be listed but can be obtained per request by contacting the first author.

V. CONLUSION

A close examination of dimensional emotions clustering in the NAPS database revealed a possibility that valence and arousal emotion dimensions could be effectively used for classification of pictures. Statistical relationship between semantic and dimensional emotion annotations may be exploited to define methods for estimation of picture semantics with emotions, and the opposite, of emotion values using a provided set of keywords.

Even though we used the heuristically proven Elbow method for establishing the optimum k, it was difficult to determine its precise value. However, the optimum number of clusters is certainly in an open interval $k_{OPT} = 2 - 16$ with median $k_{OPT} = 9$.

We believe that this research in k-means clustering of the NAPS database pictures only scratches the surface on possible avenues of research in knowledge discovery and data mining of affective multimedia databases. Since these databases are standardized, have been methodically developed and jointly, if not individually, store sufficiently large number of documents, they represent a substantial experimental corpus which is sufficient for investigation of emotions induced by multimedia and the perceived semantic content.

In the future we will focus on developing more expressive and robust models of relationship between emotion and semantics in pictures. As the work progresses we plan to report on the design, development and quality of a decision-support system for construction of affective multimedia databases. Development of video recommendation and emotion estimation systems, at least as prototypes, is in the scope of our long-term plans as well.

REFERENCES

[1] J. P. Jokinen, "Emotional user experience: Traits, events, and states," International Journal of Human-Computer Studies, Elsevier, vol. 30(76), pp. 67–77, April 2015.

[2] M. Horvat, M. Vuković, and Ž. Car, "Evaluation of keyword search in affective multimedia databases," Transactions on Computational Collective Intelligence, Springer, in press.

[3] P. Lang and M. M. Bradley, "The International Affective Picture System (IAPS) in the study of emotion and attention," Handbook of emotion elicitation and assessment, vol. 29, 2007.

[4] K. Ćosić, S. Popović, M. Horvat, D. Kukolja, B. Dropuljić, B. Kovač, and M. Jakovljević, "Computer-aided psychotherapy based on multimodal elicitation, estimation and regulation of emotion, " Psychiatr. Danub., vol. 25, pp. 340–346, 2013.

[5] K. Ćosić, S. Popović, M. Horvat, D. Kukolja, B. Dropuljić, I. Kostović, M. Judaš, M. Radoš, M. Radoš, L. Vasung, B. B. Spajić, S. Doričić, and D. Mesić, "Virtual reality adaptive stimulation in stress resistance training," In HFM-205 Symposium on "Mental Health and Well-Being across the Military Spectrum", January 2011.

[6] A. Marchewka, Ł. Żurawski, K. Jednoróg, and A. Grabowska, "The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database," Behavior research methods, Springer, vol. 46(2), pp. 596–610, 2014.

[7] M. Riegel, Ł. Żurawski, M. Wierzba, A. Moslehi, Ł. Klocek, M. Horvat, A. Grabowska, J. Michalowski, K. Jednoróg, and A. Marchewka, "Characterization of the Nencki Affective Picture System by discrete emotional categories (NAPS BE)," Behavior research methods, Springer, in press.

[8] M. Wierzba, M. Riegel, A. Pucz, Z. Leśniewska, W. Ł. Dragan, M. Gola, K. Jednoróg, and A. Marchewka, "Erotic subset for the Nencki Affective Picture System (NAPS ERO): cross-sexual comparison study," Frontiers in Psychology, vol. 6(1336), 2015.

[9] C. Peter and A. Herbon, "Emotion representation and physiology assignments in digital systems," Interacting with Computers, Oxford Journals, vol. 18(2), pp. 139–170, 2006.

[10] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament," Current Psychology, vol. 14(4), pp. 261–292, 1996.

[11] J. A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, vol. 39, pp. 1161–1178, 1980.

[12] M. M. Bradley and P. J. Lang, "Measuring emotion: The Self-Assessment Manikin and the semantic differential," Journal of Behavior Therapy & Experimental Psychiatry, Elsevier, vol. 25, pp. 49–59, 1994.

[13] M. Horvat, S. Popović, and K. Ćosić, "Towards semantic and affective coupling in emotionally annotated databases," In MIPRO, 2012 Proceedings of the 35th International Convention, IEEE, pp. 1003–1008, May 2012.

[14] JFreeChart, http://www.jfree.org/jfreechart/, retrieved on 19 February 2016.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update. ACM SIGKDD explorations newsletter," vol. 11(1), pp. 10–18, 2009.

[16] C. Ding and X. He, "K-means clustering via principal component analysis," In Proceedings of the twenty-first international conference on Machine learning, ACM, pp. 29–36, July 2004.

[17] J. P. Baudry, A. E. Raftery, G. Celeux, K. Lo, and R. Gottardo, "Combining mixture components for clustering," Journal of Computational and Graphical Statistics: A Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America, vol. 9(2), pp. 332–353, 2010.

Challenges in Adopting Big Data Strategies and Plans in Organizations

A.Budin Posavec* and S. Krajnović **
* Ericsson Nikola Tesla, Zagreb, Croatia
** Ericsson AB, Stockholm, Sweden
andrea.budin@ericsson.com, sinisa.krajnovic@ericsson.com

Abstract - In today's world, organizations of all types – be it in the industrial, governmental or non-governmental segments - are beginning to realize that embracing the full power and potential of big data and advanced analytics can bring significant benefits to their businesses and to the society in general. Technologies involved in big data analytics are already quite advanced and allow reaping the benefits of analytics results. However, many organizations are facing different issues in developing and adopting big data plans and strategies. Those issues can be internal and related to organizational challenges and the value of information, but they can also be related to regulatory and legal requirements. The focus of the paper will be on the above mentioned organizational and regulatory/legal challenges, and not the technology related ones.

I. INTRODUCTION

The world is today awash with data – be it data generated by humans or data generated by various kinds of connected devices. Data is proliferating at an astonishing rate, and much of it is or was collected in order to improve decisions about some aspect of business, government or society. In the last decade the concept of big data was embraced – data characterized by three Vs: variety, volume and velocity. When talking about big data we are referring to data that is either semi-structured or unstructured, that is present in very big volumes and is created or changing at a fast pace.

Organizations continue experiencing exponential growth in the amounts of data they capture and retain within their data centers. Many data-driven companies (such as Google, Amazon, etc) currently manage multiple petabytes of data. If that data isn't turned into better decision making through advanced analytics, the data is both wasted and probably creating suboptimal performance [1]. Despite the advanced big data technologies and methodologies (NoSQL, Hadoop, MapReduce, programming language R, etc.), many organizations are struggling with exploiting the full potential and getting real value from big data because of different organizational inhibitors and the lack of data governance mechanisms, adequate analytical skills and data science capabilities. Furthermore, organizations sometimes exploit big data without regard for issues of data legality, value, risk and cost.

This paper will first give a brief overview of the various uses of big data analytics in business organizations, in the healthcare sector and in creating public value, followed by an overview of the stages of analytics adoption in companies along with the obstacles they are facing. A perspective on data governance mechanisms will be given with respect to the value of information assets and related costs and risks. Finally, a view on the legal aspects of managing big data will be given.

II. AREAS OF CREATING VALUE WITH BIG DATA ANALYTICS

Without going into details and for the sake of illustrating the vast and complex area of application of big data analytics and the implications it may bring, but also of illustrating the complexity of required, we bring some of the examples of big data analytics benefits that have already been recognized.

Business organizations can approach various types of business decisions by applying analytics in the following business areas [1]:

- Marketing: decisions on pricing, locations of stores and branches, targeting promotions, web site customizations, digital media advertising placement;
- Finance: decisions on financial performance drivers, balanced scorecards, forecasts;
- Supply chain management: decisions on inventory, locations of distribution centers and warehouses, transport routing;
- Human resources: decisions on compensations, education, hiring employees;
- Research and development: decisions on product features, product effectiveness, product design-

The healthcare sector worldwide is facing different challenges – from increasing costs due to ageing populations and the increasing prevalence of chronic diseases in the developed world, to healthcare access in the developing countries. Big data analytics can bring significant benefits to the healthcare sector in areas including, but not limited to [2], [3]:

- Disease management: predicting hospitalization risk for individual patients that can result in reduced spending on patients with chronic diseases:
- Outcomes evaluation: measuring cost and quality of treatments that can result in more cost-efficient care;

- Guideline compliance: monitoring patient treatments and comparing them with medical guidance, which can result in reduced spending for e.g. unnecessary hospital stays;
- Productivity comparison: benchmarking hospital productivity which can result in optimized hospital budgets;
- Real-world evidence: measuring efficacy of specific drugs or procedures that can result in better discounts or rejected reimbursement for selected drugs;
- Timely detection of epidemics: tracking social media interactions that can results of earlier detection of upcoming disease (e.g. flu) epidemics;
- Determining optimal treatments: matching patient profiles with research data in order to determine the best treatment for an individual patient (e.g. IBM Watson in the Memorial Sloan Kettering Cancer Center).

Big data analytics are already today used in advanced communities for creating public value, examples of which are [4]: crowdsourcing, web interfaces for citizen engagement, open innovation platforms, online city halls, open government blogs, social media and calls to public actions.

III. ANALYTICS ADOPTION IN BUSINESS ORGANIZATIONS

Despite the new technologies that are collecting more data than ever before, many organizations are wondering whether they are getting full value from the massive amounts of data that are available. The *MIT Sloan Management Review* study (in partnership with the IBM Institute for Business Value) that covered more than 30 industries and 100 countries worldwide [5], resulted in grouping organizations into three analytical capability levels according to the stages of their adoption of analytics: aspirational, experienced and transformed.

Organizations with aspirational analytics capabilities are mostly focusing on cost efficiency or automation of their business processes. Those organizations more or less lack the necessary resources – people, processes and tools – to collect, understand or act upon analytical insights.

Organizations somewhat experienced in analytics are looking to go further than cost efficiency and are developing better ways to collect data, incorporate analytics and act based on analytics results to optimize their organization.

Transformed organizations with respect to analytics adoption substantially use analytics in various functions. Analytics is used as a competitive differentiator, whereas those organizations are already adept at organizing people, processes and tools to optimize and differentiate. Those organizations are focused on driving profitability and acquiring new customer segments, and have shown to be significantly more likely to outperform their competitors.

Aspects related to analytics adoption in organizations, such as motives, business challenges, data management and use of analytics for aspirational, experienced and transformed analytics adoption levels are depicted in Table I. (adapted from [5]).

TABLE I. ANALYTICS CAPABILITY LEVELS

Capability / Aspect	Aspirational	Experienced	Transformed
Motive	• Use analytics to justify actions	• Use analytics to guide actions	• Use analytics to prescribe actions
Business challenges	• Competitive differentiation through innovation • Cost efficiency • Revenue growth	• Competitive differentiation through innovation • Cost efficiency • Revenue growth	• Competitive differentiation through innovation • Profitability • Acquiring and retaining customers
Data management	• Limited ability to capture, gather, analyze and share information and insights	• Moderate ability to capture, gather and analyze data • Limited ability to share information and insights	• Moderate ability to capture, gather and analyze data • Effective sharing of information and insights
Use of analytics	• Rare use of analytics in decision making • Limited use of analytics results to guide strategies and actions	• Some use of analytics in decision making • Growing use of analytics results to guide strategies and actions	• Significant use of analytics in decision making • Growing use of analytics results to guide strategies and actions

When it comes to key obstacles in organizations when adopting analytics in their business, the study ([5]) shows that obtaining the data itself is not the biggest challenge. The analytics adoption barriers are usually more of a managerial and cultural nature than related to data and technology, and have been shown to include the following – in order of appearance from biggest towards smaller ones:

- Lack of understanding of how to use analytics to improve business
- Lack of management bandwidth due to competing priorities
- Lack of analytical skills
- Ability to obtain data
- Culture that does not encourage sharing information
- Unclear ownership of data or ineffective data governance
- Lack of executive sponsorship

Whereas organizations today state that the top analytics techniques most used today are trends analyses, forecasting, standardized reporting (business intelligence) and data visualization, they have also stated that they expect the following analytics techniques to create the most value in the coming years: data visualization, simulations and scenario development, analytics applied within business processes, regression analysis, discrete choice modelling and mathematical optimization [5].

IV. PERSPECTIVES ON DATA VALUE, RISK AND COST

Organizations are experiencing almost exponential yearly growth (from 40% up to 150% according to [6], [7]) in the data they capture and retain in their datacenters. Organizations can have their own datacenters or use the services of specialized datacenter service providers. Both options incur significant costs, whereas organizations are in their investment decisions often driven by the assumption that the value of stored data must exceed the data retention costs. However, at most of the organizations, generating significant business value from increased amounts of data is today still an aspiration [7].

Without a clear understanding of the real value of big data and how that value can change in time, organizations can create inappropriate data retention strategies. Insufficient investments in data storage can result in various technical, economic and reputation risks (e.g. storing data about clinical tests of a new expensive drug in an unreliable cheap storage), while huge investments into expensive data storages while the value of that data is still insignificant can prove to be unjustified.

Organizations are therefore facing a challenge to develop data governance mechanisms – policies and structures – that will strike a balance between risk and value in the face of growing data quantities and innovation delivering better, faster and cheaper storage technologies.

A. Valuing data/information assets

When certifying annual financial reports, auditors are expected to confirm that the organization owns valuable data assets which are adequately protected. However, today's financial reporting standards do not enable valuing data assets in balance sheets (except for mergers and acquisitions) [6], which means those assets are treated as having zero value even in the most data-rich organizations. Consequentially, many organizations lack methodologies for accurately measuring the value of data and information or tracking changes of that value in time.

In the face of high storage costs and rapid data growth, the concept of information life cycle management has emerged to help management understand information needs and to structure storage spending in a way that meets those needs [8].

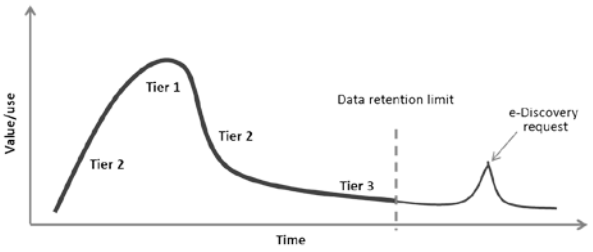


Figure 1. Information life cycle curve

A practical application of the information life cycle curve (Figure 1) requires understanding of the length of time over which information value rises and falls. This value is often in the eye of the beholder and the expected life of an information asset varies widely. For example, stock price information can exist for a second before its value changes, while electronic medical records for a patient are valuable during the patient's whole life.

Although it is a challenge for organizations to value their data assets, some other insights are possible. One of them is to view how much value can be lost (in the form of profits, revenues, data recovery costs, etc) when critical data is lost or inaccessible. This value is referred to as value-at-risk [6]. By knowing how much value is at risk from adverse events such as hardware failures or accidental data destruction, strategic decision can be made on investments in different types of storage based on storage cost at value-at-risk.

Tier 1 storage, which has the highest reliability, lowest latency and highest cost, is essential for mission critical data that must be available 24/7. Tier 2, built on storage area networks for structured data and network access storage for unstructured data, is less costly, but is generally sufficient for storing data where latency and reliability are less critical. Tier 3 storage, the least expensive (although it is most expensive to access), usually optical or magnetic tape systems, can be used for data that is rarely accessed. Figure 1 depicts the different storage tiers for data in different information life cycle phases. Data storage processes must enable moving the data from different storage tiers. Regulations that mandate data archiving and disposal policies must be taken into account along with perceived data value and storage costs.

In many cases, the anticipation of value-at-risk can be sufficient to justify data retention. Data retention is justified if the costs of storing the data are less than the opportunity costs of not having the data available for decision making or analysis. When the opportunity costs fall below the retention costs, this might signal the need to remove the data or to consider less expensive storage. There is, however, another concern that must be taken into account – so-called e-discovery request wherein organizations are obliged to deliver data (such as e.g. e-mail correspondences) as a part of a legal dispute. Such e-discovery requests may occur beyond the point where data is archived because its value does not justify retention, however, failure to produce that data could trigger legal damages.

B. Big data storage costs

Although storage costs are declining and are likely to decline further in terms of hardware costs, the data center operating costs include a variety of other operating costs that comprise the total cost of ownership of data - costs such as energy, maintenance and software that increase in direct proportion to the growth of big data (Figure 2). Research has shown that the total cost of ownership can be five to seven times higher that purely hardware acquisition costs [6].

Some organizations are already close to the point where storage costs are so significant that they compromise strategic IT investments, while IT budgets are already under increased pressure. The costs of managing big data and the evolution of analytics software are, on the other hand, likely to imply needs for higher spending on IT infrastructure and information processing,

so flat or declining IT budgets could jeopardize strategic IT projects in general. The rising need for cost control within big data environments involves the use of storage tiers (as mentioned in the previous section) that span through the information life cycle. Data can migrate from lower to higher storage tiers as its value rises, while it can migrate from higher to lower storage tiers as its value drops.

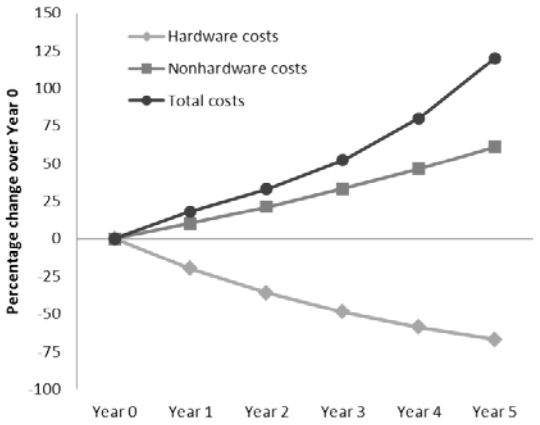


Figure 2. Big data cost simulation **[6]**

The key to managing the costs of big data is therefore, physically moving it to a higher or lower tiered storage each time there is a significant rise or decline of its value. The difficulty with determining those value changes often results in applying only simple data value metrics such as the date of last use or the frequency of use. Similar or more complex metrics require setting up data governance practices in organizations.

V. DATA GOVERNANCE PRACTICES

Data governance practices, i.e. organizational policies and procedures that describe how data is to be managed throughout its useful life cycle comprise three categories covering practices including the following [6]:

1. Structural practices
   - user involvement in policy setting and evaluation
   - steering groups to asses data value and costs
   - data ownership/stewardship rights and responsibilities
2. Operational practices
   - data retention policies
   - backup and recovery practices and parameters
   - establishing and monitoring data access rights
   - service levels for protection of information
   - monitoring risk factors for value-at-risk
   - criteria for migration between storage tiers
   - planning and provisioning storage capacities
   - e-discovery and archiving procedures
3. Relational practices
   - education of users and non-IT personell on storage utilization and cost
   - communication regarding policy effectiveness and user needs

Different organizational and technical factors can, however, impact the endeavor to implement data governance practices that can help organizations to manage the value, cost and risk of big data. The most notable data governance enablers and inhibitors resulting from research [6] are listed in Table II.

TABLE II DATA GOVERNANCE ENABLERS AND INHIBITORS [6]

Data governance enablers	Data governance inhibitors
Organizational issues	
Highly focused business strategy Aligned IT and business strategy Centralized IT and organization	Complex product and service mix Strategic misalignment Decentralized IT and organization
Industry issues	
Regulations Predictable data growth rates	Varying regulations by region (US, EU, etc.) Lack of industry-wide data standards
Technological issues	
Culture of recognizing strategic importance of IT IT standardization	Data hoarding culture Legacy IT systems

The enablers of data governance can be industry-specific and vary by industry, whereas data governance inhibitors can be present equally across all industries.

VI. REGULATORY AND LEGAL ASPECTS OF BIG DATA

Organizations exploit big data in different ways, sometimes pressing the limits of legality and data privacy. Potential negative outcomes enabled by big data can therefore extend into different social, economic and political realms. Appropriate legal frameworks for managing and controlling various big data aspects are needed which will help minimize potential negative outcomes.

A. Big data contexts

Some of today's big data sources are instances that have been created and collected already for a couple of decades – e.g. e-mail records, corporate documents, etc. More and more big data sources, however, have begun to exist recently with the omnipresent social networks and data gathering technologies (such as data from various connected sensors). In some cases, big data is available in relatively coherent data sets, while it is often coming from a variety of sources [9]. In order to illustrate the need for an appropriate big data legal/regulatory framework, some of the big data sources that require tight control of data ownership and data rights are given:

- patients' medical records
- government data about citizens;
- social media and social networks data;
- data gathered from use of loyalty cards;
- consumer profile databases;
- sensor data (from sensors communicating via RFID; Bluetooth, Wi-Fi, etc.);
- data from smart meters;
- aerial surveillance data (e.g. satellite imagery, unmanned aerial vehicles (drones), etc.)

B. A regulatory framework for big data

As the volumes of data grow and the Internet of Things get hold along with social networks, universal surveillance is no longer just a sporadic paranoid concern, but has already become a practicable proposition [9]. Free societies depend on individuals' rights in relation to their data and the interests of big data beneficiaries subject to tight controls. This is why some kind of a legal/regulatory framework is essential. A common legal/regulatory framework [10] for big data can be conceptualized by a six-layered model depicted in Figure 3. Within that framework, levels 3, 4 and 5 are specific to big data, while levels 1, 2 and 6 relate to IT systems in general

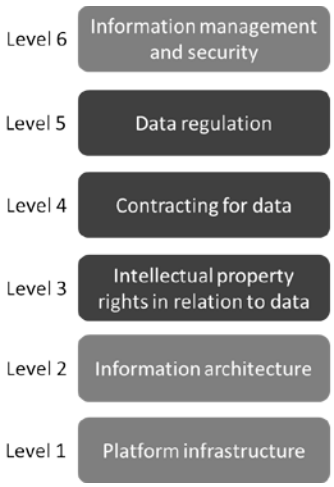


Figure 3. Big data – a legal framework **[10]**

The platform infrastructure level refers to the physical infrastructure for storing and managing big data, i.e. storage, servers, routers, local networks, internet connectivity, etc., and the software running on the physical infrastructure, i.e. operating systems, middleware, connectivity software, etc. The legal/regulatory aspects at this level relate to the traditional software copyrights issues and the relationships between copyright and database rights in relation to database software and accessing/extracting the data from that software. At this level, attention also needs to be paid on how the relevant technical standards have been adopted.

The information architecture level, as the intermediate level between the platform infrastructure and the data itself, is from the legal and regulatory perspective viewed through the appropriate copyright protection of the documentation specifying he actual information architecture. Furthermore, the formal database structure is protectable in the EU under Chapter II, Article 3 of the Database Directive [11].

The intellectual property rights in relation to data itself can be grouped into three categories: copyright, database rights and confidentiality and trade secrets. Patents can in some parts of the world apply to software and business processes, but generally no not apply in relation to the data itself.

Copyright can protect the form or expression of information, but cannot protect the underlying information or data itself. Copyright applies to software, certain databases and documentation. In the EU it arises automatically, so a formal copyright registration is not required.

Database intellectual property rights arise from the premise that the maker of the database has made a substantial investment into obtaining, verifying and presenting data [10]. The owner of the database right is generally the maker of the database and the right lasts for fifteen years from initial creation (in the EU).

While the copyright and database rights protect expression and form rather than the actual substance of the information, rules and directives for confidentiality and trade secrets aim to protect the substance of data and information. The European Commission published in 2013 a draft directive to harmonize trade secret protection through common trade secret standards [12]. The preliminary agreement on the directive text has been reached in December 2015.

The contract rights in relation to data are entirely separate from intellectual property rights, wherein a contract in relation to data confers rights and imposes obligations recognized by law. The key focus areas for contracting in relation to big data (or data in general) include the following [10]:

- scope of rights that are being licensed (internal use, further dissemination, territories where the rights may be used, combination/use with other data, treatment of derived data, etc.);
- purposes for which the data can be used (checking whether the intended data analysis is permitted, mechanisms for adding or changing purposes, etc.);
- ownership of underlying rights and rights to derived data;
- warranties of compliance with laws and regulations (data protection, sector-specific regulations, etc.);
- risk allocation (supplier and customer indemnity and liability);
- duration and termination of supply of data;
- use of data and derived data after contract expiration.

The data regulation level of the legal/regulatory framework for big data is of an increasing importance with the omnipresent growth of big data. This level comprises three important areas: data protection, competition laws and sector-specific regulation.

Data protection means conferring rights and imposing obligations on processing personal data as data relating to an identified or identifiable individual. It is becoming clear that an explicit informed consent of an individual is required to use and process his/her personal data, which in the EU is the subject of the Data Protection Regulation [13]. Data protection deals with data protection issues such as:

- fairness (concerned with the manner in which personal data is obtained);
- individuals' or organizations' consent to use their data;
- limitation of data use purposes;

- data anonymization;
- data privacy impact assessments;
- information governance.

Data regulation is in the last years analyzed also through the lens of competition laws at national or regional levels through viewing business patterns, licensing and contracting for data in a number of sectors, especially in the financial sector [10]. Data regulation is also gaining increased importance in many vertical industry sectors with the digitalization of data. Examples of such industry sectors are the already mentioned financial sector, the insurance industry (e.g. data on client solvency), the air travel industry (e.g. data on an airline customer's itinerary) and healthcare (e.g. data about anonymized clinical patient data).

The information management and security level is mostly related to compliance with common standards – such as the ISO 27000 series of Information Security Management Systems, standards in USA in how service companies should report on their information security (e.g. SSAE 16, ISAE 3402), as well as industry specific standards such as the Data Security Standards (DSS) in the payment card industry.

VII. CONLUSION

Despite the many technological advancements and the aspirations for big data analytics in their businesses, many organizations are struggling with challenges in adopting and implementing big data strategies and plans. Challenges may be of an organizational, cultural or competence nature, while in some organizations they may be related to the big data value and costs. Challenges can also arise from unfamiliar or insufficiently adopted regulations related to big data. Organizations are recognizing the need for data governance practices, policies and structures that should span across the whole organization in order to benefit as much as possible from big data analytics.

REFERENCES

[1] T. H. Davenport, Keeping Up with the Quants: Your Guide to Understanding and Using Analytics, Boston: Harvard Business Review Press, 2013.

[2] S. Biesdorf, D. Court and P. Willmott, "Big Data: What's Your Plan?," *McKinsey Quarterly,* vol. 2013, no. March, 2013.

[3] J. Kolb and J. Kolb, The Big Data Revolution, Chicago: Applied Data Labs Inc., 2013.

[4] A. McGahan, "Unlocking the Big Promise of Big Data," *Rotman Management,* vol. 2013, no. Fall, 2013.

[5] S. LaValle, E. Lesser, R. Schockley, M. S. Hopkins and N. Kruschwitz, "Big Data, Analytics and the Path from Insights to Value," *MIT Sloan Management Review,* vol. 52, no. 2, 2011.

[6] P. Tallon, "Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost," *IEEE Computer,* vol. 46, no. 6, 2013.

[7] C. Beath, J. Becerra-Fernandez and J. Short, "Finding Value in the Information Explosion," *MIT Sloan Management Review,* no. Summer, 2012.

[8] P. Tallon and R. Scannell, "Information Life Cycle Management," *Communications of the ACM,* vol. 50, no. 1, 2007.

[9] M. Wigan and R. Clarke, "Big Data's Big Unintended Consequences," *IEEE Computer,* vol. 46, no. 6, 2013.

[10] R. Kemp, "Legal Aspects of Managing Big Data," Kemp IT Law, London, 2014.

[11] European Commission, "Protection of Databases," 1996. [Online]. Available: http://ec.europa.eu/internal_market/copyright/prot-databases/index_en.htm.

[12] European Commission, "Trade Secrets," 2014. [Online]. Available: http://ec.europa.eu/growth/industry/intellectual-property/trade-secrets/index_en.htm.

[13] European Commission, "Protection of Personal Data," 1995. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV:l14012..

A survey of intelligent system techniques for Indian stock market forecasting

Subhash Panwar, Ved Prakash Upadhyay, S K Bishnoi
Govt. Engineering College, Bikaner, Rajasthan, India
Panwar.subhash@gmail.com, iamvedprakash@outlook.com, bishnoi.sk.ece@gmail.com

Abstract - Forecasting the movements of stock market is strenuous and noteworthy task for both the researchers and investors. The movements of stocks market is affected by local and global economic factors related to the market and political movements. In this field of research a substantial amount of knowledge regarding finance, statistics and Artificial Intelligence is needed to achieve good results. To perceive the movements of stock markets we need to interpret significant amount of information from the non-linear, volatile and non-parametric raw data. For reducing the complexity of stock market forecasting we need to extract some features from the raw data. To make the task of stock market forecasting easier for researchers and traders we made a study about Indian stock market and also we have presented a complete summary report. In this summary report, we have been included 50 research articles related to Indian stock market with some highly cited articles related to other international markets.

I. INTRODUCTION

The volatile behavior of stock market has been debated broadly for so many decades. Researchers from both economics and engineering fields have been examined the market by several financial and soft computing models for making predictions to future trends. Also they have been built various models to predict the volatility of stock market. They focus on cultivating different approaches to successfully predict future values of stock prices and market indices. The main goal of the researchers is to build models for prediction of stock market using least amount of the data and achieving higher amount of accuracy.

The task of forecasting is tenacious in nature because of so many intricacies involved in it. To predict stock markets accurately researchers need to choose correct input variables, modeling techniques and also need accurate performance measures for the model.

Nowadays researchers are highly fascinated to use AI and business models for the prediction of stock market volatility. A stock market index is a statistical indicator which is used for the measurement of changes in the market value of stocks. Stock market index give us information about the overall movements of different stocks of the market.

In countries like US and other developed countries stock market forecasting is became a highly researched field but in developing countries like India it is still a novice field and need to be researched more. In India the researchers are still not much aware of the potential of the field.

Therefore, the purpose of our research is to review and classify all the techniques derived in literatures. We have been reviewed 50 highly cited research papers including Indian markets and other most reputed International markets. In this paper we have been summarized the approaches used in various stock market forecasting systems.

We have produced a complete report on data sample & size, input variable, preprocessing and classifiers techniques used in stock market forecasting models. Also we made comparison for both simple and hybrid models.

Section II show various input variables used by researchers, section III elaborate complete preprocessing approach applied on raw data. In section IV the feature indicators are classified with financial and macro-economic fundamentals. Section V and VI show the comparative study of classification models and performance measures respectively used in various forecasting system. Finally we conclude our study in section VIII.

II. INPUT VARIABLES

Various input variables like opening, low, high and closing values of stocks or index of stock market are used to predict the short term movement of market. From the survey we have found that number of input variable changes according to the requirements of the researchers and the available data.

Many Indian researchers in [13], [17], [27], [47], [29], [32], [34] and international researchers in [10], [11], [44] used only closing value as single input variable to forecasting the movement of market.

Some researchers work for long term prediction and use macro-economic variables such as exports and imports, money supply, interest rates, inflation rates, foreign exchange rates, unemployment figures, and specific company financial profile [3],[14] (dividend yields, earnings yield, cash flow yield, book to market ratio, price-earnings ratio, lagged returns, size, etc).

In some cases researchers have used economical and financial factors as input variable for the model [5], [22], [26], [42] and [3].

III. DATA PREPROCESSING

Data preprocessing is one of the most important step in the data mining process. Also in the process of forecasting a time series like stock market, we need to preprocess our raw data, because the available raw data is not good enough for the modeling. Raw data from data

sources is highly inconsistent in nature and it contains noise such as redundancy and unavailability of data. The quality of data used for modeling is going to affect accuracy of prediction model. To improve the accuracy and complexity of the model we use data preprocessing.

In data preprocessing, data is normalized so that each input component is linearly scaled in a specified range like [-1.0, 1.0] or [0, 1]. Researcher who scaled their samples in range of [0, 1] are shown in articles [1], [4], [33] and [48]. In articles [18], [32], [23], [44] and [27] researchers have scaled sample in range of [-1, 1]. Also some researcher use to scale their data in a different ranges like [-0.9, 0.9] and [-0.5, 0.5] which shown in article [40] and [41] respectively.

To reduce the dimensionality of the data generally the principal component analysis (PCA) is used by authors of articles [8], [19], [21] and [26]. The higher dimensionality by redundancy in source data is generally summarized from many independent variables to a small set of derived variables. These derived variables are known as principal components. The number of principal components is always less than or equal to the number of original variables. Principal components are derived from the variance matrix. First principal component is the linear combination of the matrix elements which have largest possible variance and second element is the combination of matrix elements which have second largest variance and so on.

Time series data is non-stationary and high dimensional in nature. Time series data is said to be stationary if there is no systematic change in mean (no trend), in variance, and, if so, periodic variations have to be removed. To test stationarity of data unit root test is used like Augmented Dickey Fuller [16], [22], and Philip Perron [29].

.

IV. FEATURE EXTRACTION

To increase the accuracy of the model we need to extract some good features from the data set. If source data is large in size and there is the possibility of redundancy in data then we need to reduce the data set into a set of important information which is known as feature indicator. It is expected that feature indicator are extracted from raw data of stock market time series must contain all relevant pertinent information.

In literature review various feature indicators are used which are generally categorized in two terms as fundamental and technical, depending on stock market analysis.

Fundamental analysis is based on macroeconomic data, such as exports and imports, money supply, interest rates, inflation rates, foreign exchange rates, unemployment figures, and specific company financial profile (dividend yields, earnings yield, cash flow yield, book to market ratio, price-earnings ratio, lagged returns, size, etc.).

Technical analysis are generally ignores completely the efficient market hypothesis and it is based on the rationale that history will repeat itself and that the correlation between price and volume reveals market behavior. Prediction takes place by exploiting implications hidden in past trading activities, and by analyzing patterns and trends shown in price and volume charts.

The specific time series of stock market known as a forecasting horizon is used to select the feature indicators. The features indicators can be selected on the basis of the important factors which affects the forecasting horizon. If the forecasting time horizon covers one year or more, fundamental analysis is preferred; otherwise, if the horizon is shorter than one year, technical analysis is preferred.

Since forecasting systems are capable of choosing the relevant features automatically, the exact number of indicators is not of much importance, only those indicators which contain any relevant information will be selected by the system.

In literature we found that most of researchers classify the feature indicators in three types, namely, volume based, price based and overlay indicators. Total trading volume in stock markets is analyzed in volume based indicators and same the price value of stock is analyzed in price based indicators. As in technical chart, the feature indicators are often squiggly lines found above, below and on-top-of the price information. Such indicators are called overlay based indicators that are used the same scale as prices and typically plotted on top of the price bars. Here in Table 1, we have categorized all feature indicators which are used in literature.

Table 1: Feature Indicators

V. CLASSIFICATION AND MODELING

The most important part of the stock market forecasting system is to choose a good classifier for better performance in both intraday and long term prediction. Researchers have been experimented on various types of forecasting classifiers. Various forecasting models like ARIMA (Auto Regressive Moving Average) [9], [13], [29], EGARCH (Exponential Generalized Auto Regressive Conditional Heteroskedastic), TARCH (Threshold ARCH) [16], [17], Hidden Markov Model [19], ARFIMA-FIGARCH (Auto Regressive Fractionally Integrated Moving Average-Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity) [34] are using statistical techniques for the forecasting. Recently, researchers have been shifted their interest towards machine learning algorithms for predictions as they provide promising results.

On other hand, there are also few affecting parameters that researchers must care about the modeling process. Survey shows that most prominent parameters that affect share prices are their immediate opening and closing values [2]. Also the length of forecasting horizon and investment strategies also influences trading simulations. The forecasting horizon is the period of time in which the indexes are realized [3].

We have been made few observations that the ability of neural networks to discover nonlinear relationships in input data makes them more preferable

Indicators	Explanation	Type
On-Balance Volume(OBV)	OBV is a momentum indicator that uses **volume flow** to predict changes in stock price	Volume based indicators
Money Flow Index(MFI)	Compares the traded value of the up-days to the traded value of down-days and puts it in a percentage value.	
Volume Price Trend Indicator	It consist cumulative volume line that adds or subtracts a multiple of the percentage change in share price trend and current volume, depending upon their upward or downward movements.	
Chaikin Money Flow	CMF Oscillator is derived from MACD. Used in technical analysis as measurement of buying and selling pressure with a purpose of generating **trading signals**.	
Relative Strength Index (RSI)	A technical momentum indicator that compares the **magnitude** of recent gains to recent losses in an attempt to determine overbought and oversold conditions of an asset.	Price based indicators
Moving Average Convergence Divergence (MACD)	The difference between a fast and slow exponential moving average **(EMA)** of closing prices. (Fast means a short-period average, and slow means a long period one)	
Rate Of Change(ROC)	The **percentage difference** between the current price and the price n-time periods ago.	
Stochastic Oscillator (SO)	It compares a security's closing price to its price range over a given time period. The **oscillator's sensitivity** to market movements can be reduced by adjusting the time period or by taking a moving average of the result	
William's %R	This is a momentum indicator measuring overbought and oversold levels, similar to a **stochastic oscillator**	
Momentum	To identify trend lines used **an oscillator** which is known as momentum.	
Chaikin Oscillator	Combines price and volume to show how money may be flowing into or out of a stock. Based on **Accumulation/Distribution Line**	
Moving Average	To emphasize the direction of a trend and smooth out price and volume fluctuation that can confuse interpretation.	Overlay based indicators
Bollinger Bands	A chart overlay that shows the upper and lower limits of price movements based on the **Standard Deviation** of prices.	

over other models for modeling nonlinear dynamic systems such as the stock market [4]. Also Artificial Neural Network (ANN) classifier has the ability to identify outliers and erroneous data [3]. It implements the Empirical risk minimization principle, outperforms other traditional statistical models [11], [15]. It is observed that the risk-adjusted performance of the NN based trading model is generally better than Buy and Hold strategy [41].

ANN is widely accepted for stock market prediction due to its capacity of learning non linear pattern in more accurate way [48]. Also various versions of Neural Network classifiers in stock market forecasting are used as denoted in [1], [15], [22], [27], [39] and [41].

Survey shows that the over fitting risk was higher in ANN. Also it suffers from local minimum traps and difficulty in determining the hidden layer size and learning rate [11]. When we compare it to Support Vector Machine (SVM) model [20], we found that SVM provides good generalization performance, absence of local minima, and sparse representation of solutions due to its characteristics. SVM classifier is based on the structural risk minimization (SRM). This induction principle is different from the commonly used empirical risk minimization (ERM) principle which only minimizes the training error.

Established on the unique principle, SVM usually achieves higher generalization performance than traditional ANN that implement the ERM principle in solving many machine learning problems. Another key characteristic of SVM is that training SVM is equivalent to solving a linearly constrained quadratic programming problem so that the solution of SVM is always unique and globally optimal, unlike other networks training which requires nonlinear optimization with the danger of getting stuck into local minima.

The solution to the problem that SVM is only dependent on a subset of training data points which are referred to as support vectors. Using only support vectors, the same solution can be obtained as using all the training data points. Literature shows that SVM outperformed random forest, neural network and other traditional models. The superior performance of SVMs over the other models is due to the reason that SVMs implement the structural risk minimization principle which minimizes an upper bound of the generalization error rather than minimizes the

training error [10] [23]. This eventually leads to better generalization than the neural network and random forest which implements the empirical risk minimization principle [50]. Various stock market forecasting models as in [31], [18], [25], [50], [37], [38], [40], [43] and [44] used the SVM classifier to utilize the structural risk minimization principle.

Just like SVM the regularized RBF neural network minimize the regularized risk function, rather than the empirical risk function as used in the BP neural network. So they are robust to over fitting, eventually resulting in better generalization performance than the BP neural network. [10]

One disadvantage of SVM is that the training time scales somewhere between quadratic and cubic with respect to the number of training samples. So a large amount of computation time will be involved when SVM is applied for solving large-size problems [10].

In literature we have found that various experiments have also done with mixed classifier and achieved better results compare to a single classifier models. In such hybrid models as in article [7] researchers has been used decision tree-KPCA-ANFIS hybrid system which uses decision tree for feature selection outperforms simple neural network and naïve Bayesian models. Also in [8] the decision tree rough set based prediction systems outperform compare to stand-alone rough set based prediction system without any feature selection and artificial neural network based prediction system without any feature selection.

A hybrid ARIMA-GARCH model which involves a MA filter based decomposition as a pre-processing step outperforms compare to other models such as ARIMA, GARCH and trend-ARIMA, wavelet-ARIMA [9].

A hybrid neuro fuzzy adaptive control system has been developed and compared with 13 other soft computing based approaches to show its superior performance [14].

In article [18] the performance of proposed models were improved significantly when they learnt through trend deterministic data. The accuracy of 86.69 %, 89.33 %, 89.98 % and 90.19 % was achieved by ANN, SVM, random forest and naive-Bayes respectively.

The Cuckoo Search (CS) [20] is based on the Swarm Intelligence optimization technique and is very simple to tune the parameters of SVM. The obtained experimental results indicate that the CS-SVM method can achieve higher accuracy rate than regular SVM method. The proposed experimental results demonstrate that CS-SVM provides higher accuracy with lower MSE and MAPE among ANN and SVM. A hybrid ARIMA-neural network model [47] outperforms linear ARIMA and nonlinear ANN. The Genetic Algorithm (GA) with optimized decision tree-SVM hybrid system [31] outperforms both the artificial neural network and the naïve bayes based prediction systems and also the hybrid system [37] with genetic and SVM models outperforms the stand alone SVM model.

VI. PERFORMANCE MEASURES

Performance measure or quality measure plays an indispensable role in the field of machine learning and data sciences. Performance measures are used for evaluating strength of the classification models and also they are used to make criterion for designing heuristics to develop classification models. Performance measures can be classified in two types known as statistical and non-statistical measure.

Statistical measures further classified as parametric and non-parametric tests. Parametric tests are those test that make some kind of assumptions about its parameters (properties) of the distribution from the one's data are driven. While non-parametric is one that makes no any kind of assumptions. The non-statistical performance measure includes hit ratio which is widely used for time series problems.

Table 2: Performance Measures [7]

MSE	Mean squared error	Mean (et^2)
RMSE	Root mean squared error	√ (MSE)
MAE	Mean absolute error	mean(|et|)
Md. AE	Media absolute error	median(|et|)
MAPE	Mean absolute percentage error	mean(|pt |)
Md. APE	Median absolute percentage error	median(|pt|)
SMAPE	Symmetric mean absolute percentage error	mean(2|Yt - Ft |/(Yt +Ft))
SMd. APE	Symmetric median absolute percentage error	median(2|Yt -Ft |/(Yt +Ft))

If the problem is a regression problem then we should use root mean squared error (RMSE), Mean absolute error (MAE), Steady state error (SSE) etc. While in problems related to classifications we use confusion matrix or error matrix to measure the performance. Table 2 shows the performance measure which are generally used in literature.

VII. Conclusion

Our study has surveyed various articles which used several approaches to solve the stock market forecasting system. In our survey, we have found that Artificial intelligence techniques perform much better then statistical techniques and others. Moreover we found that such soft computing techniques as support vector machine is most beloved classification model and also it provides better accuracy as compare to other models. Literatures have evidence that hybrid models give most promising results compare to single classification model.

REFERENCES

[1]. A. Victor Devadoss and T. Antony Alphonnse Ligori, "Stock Prediction Using Artificial Neural Networks", International Journal of Data Mining Techniques and Applications, Volume: 02, PP: 283-291, 2013.

[2]. Ajith Abraham, Ninan Sajith Philip and P. Saratchandran, "Modeling Chaotic Behavior of Stock Indices Using Intelligent Paradigms", Neural, Parallel & Scientific Computations - Special issue: Advances in intelligent systems and applications, Volume 11 Issue 1 & 2, PP: 143 – 160, 2003.

[3]. An-Sing Chen, Mark T. Leung, Hazem Daouk, "Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index", Computers

& Operations Research, Volume 30 Issue 6, PP: 901 – 923, 2003.
[4]. Asif Ullah Khan,T.K. Bandhopadhyaya, Sudhir Sharma, "Genetic Algorithm Based Backpropagation Neural Network Performs better than Backpropagation Neural Network in Stock Rates Prediction", IJCSNS International Journal of Computer Science and Network Security, Volume.8 No.7, 2008
[5]. Avijan Dutta, Gautam Bandopadhyay, Suchismita Sengupta, "Prediction of Stock Performance in the Indian Stock Market Using Logistic Regression", International Journal of Business and Information, Volume 7, Number 1, 2012.
[6]. Gourishankar S Hiremath and Kamaiah Bandi, "Some Further Evidence on the Behaviour of Stock Returns in India", International Journal of Economics and Finance Volume 2, No. 2; May 2010.
[7]. Binoy B. Nair, M. Minuvarthini, Sujithra B. and V.P. Mohandas, "Stock market prediction using a hybrid neuro-fuzzy system", International Conference on Advances in Recent Technologies in Communication and Computing, 2010.
[8]. Binoy.B.Nair, V.P. Mohandas and N.R.Sakthivel, "A Decision Tree- Rough Set Hybrid System for Stock Market Trend Prediction", International Journal of Computer Applications, Volume 6-No.9, 2010.
[9]. C. Narendra Babu and B. Eswara Reddy, "Selected Indian Stock Predictions using a Hybrid ARIMA-GARCH Model", International Conference on Advances in Electronics, Computers and Communications (ICAECC),2014, DOI: 10.1109/ICAECC.2014.7002382, INSPEC Accession Number: 14841743.
[10]. Cao and Tay, "Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting", IEEE Transactions on neural networks, Volume. 14, No. 6, 2003.
[11]. Yeh,Huang and Lee, "A multiple-Kernel support vector regression approach for stock market price forecasting", Volume 38, Issue 3, PP 2177-2186, 2011 .
[12]. Crina Gro¸san, Ajith Abraham, Sang Yong Han and Vitorino Ramos, "Stock market prediction using Multi Expression Programming".
[13]. Debadrita Banerjee, "Forecasting of Indian Stock Market using Time-series ARIMA Model", 2nd International Conference on Business and Information Management (ICBIM), Print ISBN: 978-1-4799-3263-4,INSPEC Accession Number: 14790920, 2014.
[14]. George S. Atsalakis and Kimon Valavanis, "Forecasting stock market short-term trends using a neuro-fuzzy based methodology", Expert Systems with Applications , Volume 36 Issue 7,PP 10696-10707, 2009.
[15]. Guoqiang Zhang, B. Eddy Patuwo, Michael Y. H, "Forecasting with artificial neural networks: The state of the art", International Journal of Forecasting , Volume. 14, issue 1, PP: 35-62, 1998.
[16]. Harvinder Kaur, "Time Varying Volatility in the Indian Stock Market", Vikalpa , Volume 29, No 4, 2004.
[17]. Hojatallah Goudarzi , C.S. Ramanarayanan , "Modeling Asymmetric Volatility in the Indian Stock Market" , International Journal of Business and Management ,Volume 6, No. 3, 2011
[18]. Jigar Patel, Sahil Shah, Priyank Thakkar , K Kotecha, "Predicting stock and stock price index movement using trend Deterministic and machine learning techniques", Expert Systems with Applications, Volume 42,PP 259–268,2015.
[19]. Jyoti Badge, "Forecasting of Indian Stock Market by Effective Macro- Economic Factors and Stochastic Model", Journal of Statistical and Econometric Methods, Volume 1, no.2, PP: 39-51, 2012.
[20]. Ms.K.Nirmala Devi, Dr.V.Murali Bhaskaran, G.Prem Kumar, "Cuckoo Optimized SVM for Stock Market Prediction", IEEE Sponsored 2nd International Conference on Innovations in Information, Embedded and Communication systems (ICJJECS), Print ISBN: 978-1-4799-6817-6, INSPEC Accession Number: 15366666, 2015.
[21]. K.V.Sujatha and S. Meenakshi Sundaram, "A Combined PCA-MLP Model for Predicting Stock Index", ACM, Article No-17, ISBN: 978-1-4503-0194-7, 2010.
[22]. Kunwar Singh Vaisla and Ashutosh Bhatt, "An Analysis of the Performance of Artificial Neural Network Technique for Stock Market Forecasting", International Journal on Computer Science and Engineering (IJCSE), Volume 02, No. 06, PP: 2104-2109, 2010.
[23]. Kyoung-jae Kim, "Financial time series forecasting using support vector machines", Neurocomputing 55, PP: 307–319, 2009.
[24]. Lipika Dey, Anuj Mahajan and SK. Mirajul Haque, "Document Clustering for Event Identification and Trend Analysis in Market News", Seventh International Conference on Advances in Pattern Recognition, Pages 103-106, ISBN: 978-0-7695-3520-3, 2009.
[25]. Mr. Manish Kumar and Dr. M. Thenmozhi , "Support Vector Machines Approach to Predict the S&P CNX NIFTY index returns", http://ssrn.com/abstract=962833
[26]. M.P. Rajakumar and V.Shanthi, "Computationally intellectual structure for forecasting share price", International journal of soft computing, Volume: 8, PP: 218-222, 2013.
[27]. M.Thenmozhi, "forecasting stock index returns using neural networks", Delhi Business Review, Volume 7, No. 2, 2006.
[28]. M.V.Subha, S.Thirupparkadal Nambi, "Classification of Stock Index movement using k-Nearest Neighbours (k-NN) algorithm", E-ISSN: 2224-3402, Volume 9, Issue 9, 2012.
[29]. Manish, kumar, "A time-varying parameter vector autoregression model for forecasting emerging market exchange rates", International Journal of Economic Sciences and Applied Research, Volume. 3, No. 2, PP. 21, 2010.
[30]. Md. Rafiul Hassan, Baikunth Nath, Michael Kirley. "A fusion model of HMM, ANN and GA for stock market forecasting", Expert Systems with Applications, Volume 33 Issue 1, PP: 171-180,2007.
[31]. Binoy B. Nair , V.P Mohandas and N.R. Sakthivel, "A Genetic Algorithm Optimized Decision Tree- SVM based Stock Market Trend Prediction System", International Journal on Computer Science and Engineering(IJCSE) Volume 02, No. 09 ,PP: 2981-2988, 2010.
[32]. Neeraj Mohan, Goutam Dutta, Pankaj Jha, Arnab Kumar Laha, "Artificial Neural Network Models for forecasting stock price index in Bombay stock exchange" , Journal of Emerging Market Finance, volume 5, no. 3, PP: 283-295, 2006.
[33]. Nitin Merh, Vinod P. Saxena and Kamal Raj Pardasani, "A Comparison between hybrid approaches of ANN and ARIMA for Indian stock trend forecasting", Business Intelligence Journal,Volume: 3,Issue: 2, PP 23, 2010.
[34]. P. Bagavathi Sivakumar and Dr. V. P. Mohandas , "Modeling and Predicting Stock Returns using the ARFIMA-FIGARCH A case study on Indian Stock data", IEEE 978-1-4244-5612, 2009.
[35]. Chang, P.-C., & Liu, C.-H., "A TSK type fuzzy rule based system for stock price prediction", Expert Systems with Applications: An International Journal, Volume 34 Issue 1, PP 135-144, 2008
[36]. Luigi Troiano and Pravesh Kriplani , "Predicting Trend in the Next-Day Market by Hierarchical Hidden Markov Model" ,978-1-4244-7817-0, INSPEC Accession Number:11664101, PP: 199 – 204, 2010.
[37]. Rohit Choudhry, and Kumkum Garg, "A Hybrid Machine earning System for Stock Market Forecasting", World

Academy of Science, Engineering and Technology, Volume 41,PP :315-318, 2008.
[38]. Rudra Kalyan Nayak, Debahuti Mishra, Amiya Kumar Rath, "A Naïve SVM-KNN based stock market trend reversal analysis for indian benchmark indices" , Applied Soft Computing , Volume 35, Issue C, PP: 670-680, 2015.
[39]. Satyajit Dhar, Tuhin Mukherjee, Arnab Kumar Ghoshal, "Performance Evaluation of Neural Network Approach in Financial Prediction: Evidence from Indian Market", Proceedings of the International Conference on Communication and Computational Intelligence , PP:597 – 602, INSPEC Accession Number: 11887488,2010
[40]. Shom Prasad Das and Sudarsan Padhy," Support Vector Machines for Prediction of Futures Prices in Indian Stock Market", International Journal of Computer Applications, Volume 41– No.3, 2012.
[41]. Subrata Kumar Mitra, "Optimal Combination of Trading Rules Using Neural Networks", International Business Research, Volume 2, No. 1,2009,
[42]. T.D. Chaudhari and Indrani Ghosh, "Using Clustering Method to Understand Indian Stock Market Volatility", Communications on Applied Electronics (CAE) – ISSN: 2394-4714, 2015.
[43]. Wei Huanga, Yoshiteru Nakamoria, Shou-Yang Wang, "Forecasting stock market movement direction with support vector machine", Computers & Operations Research, Volume 32, Issue 10, PP: 2513–2522, 2005
[44]. Yakup Kara, Melek Acar Boyacioglu, Ömer Kaan Baykan ,"Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange", Expert Systems with Applications, Volume 38, no. 5, PP. 5311-5319, 2011
[45]. Chen,Dong and Zhao, "Stock Index Modeling using EDA based Local Linear Wavelet Neural Network", Print ISBN:0-7803-9422-4, INSPEC Accession Number: 9072728, PP: 1646 - 1650, 2005.
[46]. Yuehui Chen, Ajith Abraham, Ju Yang and Bo Yang, Hybrid Methods for Stock Index Modeling, Volume: 3614, PP: 1067-1070, 2005.
[47]. Manish Kumar and M. Thenmozhi, "Stock Index Return Forecasting and Trading Strategy Using Hybrid ARIMA-Neural Network Model", International Journal of Financial Management , Volume 1 Issue 1, 2012.
[48]. Richa Handa, H.S. Hota , S.R. Tandan, "Stock Market Prediction with Various Technical Indicators Using Neural Network Techniques", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 3, Issue 6, 2015.
[49]. George S. Atsalakis , Kimon P. Valavanis , "Surveying stock market forecasting techniques – Part II: Soft computing methods", Expert Systems with Applications , Volume 36, Issue 3, Part 2, Pages 5932–5941, 2009.
[50]. Manish Kumar, M. Thenmozhi, "Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest", ssrn.876544, 2006.

The effect of class distribution on classification algorithms in credit risk assessment

K. Andric * and D. Kalpic **
* Privredna Banka Zagreb, Radnička 50, 10000 Zagreb, Hrvatska
** Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva, Unska 3, 10000 Zagreb, Hrvatska
E-mails: kristina.andric@fer.hr; damir.kalpic@fer.hr

Abstract - This paper makes a practical contribution to classification in credit risk assessment by providing empirical evidence on which class distribution in the training sample should be used to maximize performance of different classification algorithms. Data used for credit risk assessment are often imbalanced due to the number of defaulting clients, who represent the minority class, being much less numerous than the non-defaulting majority class. Classification algorithms are usually biased towards the majority class and can show deceivingly high overall prediction accuracy, while at the same time exhibiting poor performance in prediction of the minority class. Although altering class distribution can be an effective method for alleviating the adverse impact of class imbalance, limited research efforts have been devoted to empirical evaluation of the role of class distribution in credit risk assessment, especially when dealing with real life data samples. To address this issue, an empirical study on the effect of different proportion of training examples belonging to each class on classifiers' performance is presented. In addition to logistic regression, neural networks and gradient boosting methods were evaluated using several real life and publicly available datasets from the UCI Machine Learning Repository.

I. INTRODUCTION

Credit risk can be defined as the potential that a borrower or counterparty would fail to meet its his or her, in further text for simplicity "his", obligations in accordance with agreed terms [1] and it is one of the principal risks for financial institutions [2]. Credit risk assessment is one of the key factors for long-term success of financial institutions [1]. The ability to distinguish between borrowers who would be able to meet their obligations and those who might not, in other words those who would default on their obligations, is crucial to ensure sustainability of the credit business. Therefore, increasing prediction accuracy, even by a small amount, would bring substantial savings to credit institutions.

Credit risk assessment is a binary classification problem. Borrowers are classified into one of two classes based on historical data on clients bearing similar features. Data used for credit risk assessment are often imbalanced, since the number of defaulting clients (bad clients), representing the minority class, is much less numerous than the non-defaulting majority class (good clients). In credit risk modelling this is usually the case when trying to assess default risk for sovereign, bank or large corporate credit borrowers. These portfolios are often referred to as low default portfolios. Class imbalance has been identified as one of the key challenges in data mining and has received considerable attention in the recent years [3]. It presents a challenge for classification algorithms, as they are usually biased towards the majority class and can show ambiguously high overall prediction accuracy, while at the same time exhibiting poor performance in prediction of the minority class. Level of imbalance can be described with Imbalance ratio (IR), a ratio of the number of majority class cases and minority class cases. To improve prediction accuracy where class imbalance is present, several methods have been proposed, namely altering class distribution in the training data sample. In particular, simple random undersampling has proven to be quite effective [4]. However, relatively little attention has been paid to evaluating the role of class distribution on credit risk assessment when dealing with imbalanced data samples, especially when dealing with real life data samples.

This paper represents an extension of previous research that already demonstrated that altering class distribution may increase prediction accuracy. It makes a practical contribution in credit risk assessment by providing empirical evidence on how the training sample should be designed in terms of class distribution in order to maximize performance of different classification algorithms. An empirical study on the effect of different proportion of training examples belonging to each class on various classification algorithms' performance is therefore presented. Logistic regression, neural networks and gradient boosting methods were assessed using several real life and publicly available datasets from the UCI Machine Learning Repository. Prediction accuracy was evaluated by using appropriate performance metrics, depending on the level of imbalance.

The remainder of the paper is organized as follows: Section 2 provides an overview of the background work, followed by Section 3 that gives an overview of the classification methods used. Experimental framework and datasets analysed are described in Section 4, whereas

Section 5 provides a detailed analysis of the effect of class distribution on the classification performance. The final chapter presents conclusions drawn from the study and possible future research direction.

II. BACKGROUND WORK

Credit risk assessment has been explored using a wide range of classification methods. Earlier research focused on traditional statistical methods, such as linear discriminant analysis (LDA), logistic regression etc. [5], [6],[7]. Nonparametric methods, which unlike traditional statistical methods do not make assumptions about the relation between predictors and the dependent variable [8], have also shown good performance in credit risk assessment [9], [10], [11]. Various studies evaluated performance of those methods, such as k nearest neighbours [12], neural networks [13], [14], [15], genetic algorithms [6], [16], [17] and support vector machine (SVM) method [18], [19], [20], [21]. In [11] the authors compared seventeen different classification methods on eight real-life data samples. The conclusion was that complex methods, such as radial basis function, support vector machines and neural networks perform rather well in terms of AUC (Area Under Curve), but so do traditional, simpler methods such as logistic regression and linear discriminant analysis. Different studies often reach different conclusions as to which classification method yields the best performance, showing that performance depends on the domain, data characteristics and assessment criterion [5], [12], [22].

Although a lot of research has been devoted to credit risk assessment and comparing different classification algorithms, limited research efforts have been devoted to dealing with class imbalance and the role of class distribution on prediction accuracy. Class imbalance is said to be present when the number of defaulting clients, the class of interest, is much smaller than the number of non defaulting clients belonging to the majority class. It has received considerable attention in data mining in recent years and was identified as one of the top 10 problems in data mining research [3]. In [23] the authors studied performance of eight different classification algorithms using five imbalanced real life data sets. In addition to using imbalanced data samples, class imbalance was progressively increased to identify the extent to which each of the classification algorithms was affected by the level of class imbalance. Their findings indicate that random forest and gradient boosting methods, relatively new methods used in credit risk assessment, perform very well in terms of AUC, especially when strong class imbalance is present. Logistic regression, one of the most commonly used methods in credit risk assessment, showed performance not significantly different from that of the gradient boosting and random forest algorithms. On the other hand, performance of decision tree algorithm C4.5, quadratic discriminant analysis and LS-SVM was significantly worse.

Various methods have been proposed to address the issue of class imbalance. Methods can broadly be categorized into these groups [24]:

- data pre-processing techniques
- adjustments of the classification algorithm or
- cost-sensitive learning

The aim of data pre-processing techniques is to alter the original imbalanced data sample in order to produce a more balanced class distribution. Various data pre-processing techniques have been proposed in general data mining literature to find an effective way of dealing with class imbalance. The key advantage of data pre-processing techniques is that they can be applied independently, in other words they do not depend on the classification algorithm. The most commonly used are random oversampling, which implies increasing the number of instances belonging to the minority class and undersampling, or decreasing the number of instances belonging to the majority class. Some studies have found that balancing data sets by random oversampling yields better prediction accuracy than by using random undersampling [25]. A number of studies have compared various oversampling and undersampling methods and reached different conclusions on prediction accuracy derived by oversampling versus undersampling, indicating that the optimal approach might be more dependent on the dataset characteristics rather than the pre-processing technique [26].

The optimal class distribution, ie. distribution of majority and minority class instances in the training data sample has also been studied. Although not focused on the credit risk domain, in [27] the authors analysed classification accuracy of the C4.5 classifier, using 25 different data sets. They found that the best class distribution for learning, to achieve the best performance in terms of AUC, tends to be near the balanced class distribution. That means that the imbalance ratio should be roughly equal to one, which is not easy to achieve in real life credit risk assessment problems where the number of minority class examples is usually very small to begin with. In [28] authors presented an alternative approach, a wrapper technique to define the best percentage to perform both undersampling and oversampling.

III. CLASSIFICATION ALGORITHMS

For the purpose of this study, three classification algorithms were selected: logistic regressions, neural networks and gradient boosting. Logistic regression is a well established method and one of the most frequently used methods for credit risk assessment [25]. Neural network is a nonparametric method with a good track record of prediction accuracy in credit risk assessment [13], [14], [15]. Gradient boosting, one of the newest methods to be used in credit risk assessment, already showed promising results when dealing with imbalanced datasets [22] and will be evaluated further in this study. A short overview of these classification algorithms is provided below.

A. Logistic regression

Logistic regression is one of the most commonly used methods in credit risk assessment. The dependent variable Y is equal to 1 if the client is not able to make necessary repayments, or in other words if he is classified as a bad client. Alternatively, if the client is classified as a good client, Y equals 0. If we denote a collection of n independent predictor variables by a vector x = ($x_1, x_{2,\dots} x_n$), and the conditional probability that the outcome is present by P(Y = 1|x), the logistic regression model can be described by the following equation:

$$(1) \quad \text{logit}\,(P(Y=1|x)) = \ln\left(\frac{P(Y=1|x)}{1-P(Y=1|x)}\right) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

b_0 is the intercept parameter and $b_1 \dots b_n$ are the variable coefficients. For the purpose of this study stepwise variable selection method was used.

B. Neural networks (Multi-layer perceptron)

Neural networks are mathematical representations modelled after the human brain [29]. Unlike logistic regression, no assumptions on the relation between independent (predictor) and the dependent variable are made. Therefore, virtually any type of dependency of the independent variables and dependent variable can be described. The most commonly used type of neural networks is the multilayer perceptron (MLP), which is composed of an input layer, a hidden layer, and an output layer. Each layer can contain several neurons that process their inputs into output values, which are then used by the neurons in the next layer. To each connection between neurons are assigned initial weights, which are subsequently adjusted during training. If we denote a collection of n independent predictor variables by a vector x = ($x_1, x_{2,\dots} x_n$) and w_{ij} as the weight connecting input j to neuron i, the output of the hidden neuron is determined via an activation function f;

$$(2) \quad h_i = f\,(b_i + \sum_{j=1}^{n} w_{ij}x_j)$$

b_i represents the bias term. Assuming there are m hidden neurons in the hidden layer and v_i is the weight connecting hidden neuron h to output neuron, the output of the output layer is determined using a similar function:

$$(3) \quad o = g\,(b_i + \sum_{j=1}^{m} v_i h_i)$$

In this study the standard backpropagation algorithm was used to determine the weights, where the weights of the neural network were initialized to random values and adjusted during training to minimize the error term.

Number of units in the hidden layer was set to 3, with no direct connections between input and output nodes allowed. In other words, inputs are connected to the hidden layer and hidden layer is connected to the output.

C. Gradient boosting

Gradient boosting creates a single predictive model consisting of a series of simple decision trees. It is an ensemble algorithm that aims to improve accuracy by minimizing the error term [30]. After the initial learning algorithm is set up, each subsequent algorithm is fitted to the so-called ‘‘pseudo residuals’’ of the previous learning algorithm in order to minimize the prediction error.

The model can be described as:

$$(4) \quad F(x) = G_0 + \sum_i b_i L_i(x)$$

G_0 is the first value for the series, and $L_1, \dots, L_n$ are the learners fitted to the pseudo-residuals. B_i are coefficients for the respective tree nodes computed by the algorithm. Similarly to decision trees, gradient boosting does not make assumptions about the distribution of the data. Unlike the decision tree, it is less prone to overfitting. Since each successive sample is weighted according to the classification accuracy of previous trees, it is sometimes called stochastic gradient boosting [31].

For the purpose of this study, the number of terms in the boosting series was set as 50 and the maximum branch size used in the splitting rule was set to 2, resulting in binary trees.

IV. EXPERIMENTAL SETUP

A. Data sets

In this study several real life cases obtained from a Croatian commercial bank were used, as well as publicly available datasets from the UCI Machine Learning Repository [32], to enable fair comparison with other studies.

The characteristics of the data sets are outlined in Table I.

TABLE I. DATA SETS

Data set	No of cases	No of features	No of good clients	No of bad clients	Imbalance ratio
German Credit data	1.000	20	700	300	2,3
Japanese credit screening	690	13	361	329	1,09
DS1	3.053	206	646	17	38
DS2	25.226	218	25.016	210	119

Real life datasets are DS1 and DS2, whereas Japanese credit screening and German credit data were obtained from the UCI Machine Learning Repository [32]. 70% of the cases were used in the training data sample, whereas 30% of the cases were used as a validation sample for all the combinations of the classification algorithm and imbalance ratio.

B. Performance metrics

Performance metrics in imbalanced datasets is the key factor for proper evaluation of the classification accuracy. The confusion matrix is often used to demonstrate classification algorithm's performance by showing the cases that were classified correctly and incorrectly in both classes.

TABLE II. CONFUSION MATRIX

		Prediction	
		Positive	***Negative***
Actual	***Positive***	TP (true positive)	FN (false negative)
	Negative	FP (false positive)	TN (true negative)

As a result, the accuracy rate is defined as follows:

$$(5)\quad AR = \frac{TP + TN}{TP + TN + FN + FP}$$

Even though Accuracy rate is a very commonly used performance indicator, it is not appropriate when dealing with imbalanced datasets. For example, a classifier could achieve an accuracy of 95% in a dataset where the percentage of minority class examples is 5% if all cases are classified as majority class.

When dealing with imbalanced datasets, performance indicators should be carefully selected to take into account the class distribution. Different measures can be defined to measure classification performance for both minority and majority class:

- True positive rate or sensitivity (recall), which measures the percentage of correctly classified cases of the minority class:

$$(6)\quad TPrate = \frac{TP}{TP+FN}$$

- True negative rate or specificity, which measures the percentage of correctly classified cases of the majority class:

$$(7)\quad TNrate = \frac{TN}{FP+TN}$$

- False negative rate, which measures the percentage of incorrectly classified cases of the minority class:

$$(8)\quad FNrate = \frac{FN}{TP+FN}$$

- False positive rate which measures the percentage of incorrectly classified cases of the majority class:

$$(9)\quad FPrate = \frac{FP}{FP+TN}$$

Even though prediction accuracy for the minority class is of greater interest, accuracy of the majority class should not be overlooked. Therefore, in this study measures that describe prediction accuracy for both classes are taken into account. A widely used indicator that takes into account both, the Receiver Operating Characteristic (ROC) and the appertaining Area under the ROC Curve (AUC) are used. ROC is a two dimensional graph that visualizes the dependency of the TPrate (on the Y-axis), percentage of correctly classified cases of the minority class, and FPrate (on the X-axis), percentage of incorrectly classified cases of the majority class. AUC is a single measure of a classifier's performance and is very useful when comparing average performance of different classification algorithms. The perfect classifier would incorporate point (0, 1) where all the cases would be classified correctly and the random classifier the point (0.5, 0.5). A classification method is said to have good performance when it outperforms a random model and the ROC curve should be as far to the top left-hand corner as possible.

Another example of a metric used, that provides a combined measure of the prediction accuracy for both classes is the geometric mean of the true rates measure, defined as follows:

$$(10)\quad GM = \sqrt{TPrate * TNrate}$$

C. Class distribution and input selection

In the first trial run we analyzed performance of the classification algorithms without changing the underlying class distribution. Afterwards, the initial datasets were altered to give various imbalance ratios, ranging from 90/10, where 90% of the cases in the training sample are majority class examples, to 20/80, where 20% of the cases in the training sample are majority class example. This was achieved by randomly undersampling majority class observations (good clients) or randomly undersampling both majority class observations (good clients) and minority class observations (bad clients) for data samples where ratio of bad clients was higher than the target ratio. In other words, for this empirical study the minority class ratio, the bad clients ratio, in each of the training data sets was changed artificially, by a factor, starting from 10% and up to 80% to create larger differences in the underlying class distribution. As a result, eight altered data sets were created for each of the original data sets. The resulting ratio of minority class in the new data sets was 10%, 20%, 30%, 40%, 50%, 60%, 70% and 80%. Class distribution was changed only in training datasets, whereas the validation datasets were not altered. To assess performance of each combination of class distribution and classification algorithm, 4-fold cross validation was performed.

Predictive variables were selected by analysing R-square (squared correlation) and Chi-square, whereas the maximum number of predictive variables was set to seven, to limit overfitting.

V. RESULTS

Tables III, IV and V summarize the overall results with AUC as the performance measure, which reflect performance on the validation datasets. They show the

AUC values for the eight data sets at different minority class ratios, or imbalance ratios, for each of the three classification algorithms. The class distribution resulting in the highest AUC for each data set is underlined. Table III demonstrates that logistic regression is a fairly robust method, relatively insensitive to the imbalance ratio. Nevertheless, the best result is achieved when imbalance ratio is greater than 1, especially with real life datasets. Neural networks show similar results, achieving best results when imbalance ratio ranges from 80%/20% to 70%/30%. For gradient boosting class distribution of 80%/20% yields best performance in terms of AUC for UCI datasets, whereas class distributions of 40%/60% and 50%/50% yield best performance in terms of AUC for real life datasets.

Overall it can be observed that changes in class distribution do not cause great disturbances in AUC for all the classification algorithms.

TABLE III. AUC FOR LOGISTIC REGRESSION

Data sets	Imbalance ratio							
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8
German Credit data	0.72	0.74	0.74	0.74	0.74	0.74	0.74	0.73
Japan credit screening	0.89	0.90	0.89	0.89	0.89	0.89	0.89	0.88
DS1	0.79	0.79	0.79	0.79	0.79	0.78	0.76	0.77
DS2	0.76	0.75	0.74	0.74	0.73	0.74	0.73	0.71

TABLE IV. AUC FOR NEURAL NETWORK

Data sets	Imbalance ratio							
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8
German Credit data	0.70	0.75	0.74	0.73	0.72	0.72	0.70	0.68
Japan credit screening	0.87	0.89	0.89	0.88	0.89	0.89	0.88	0.88
DS1	0.77	0.76	0.78	0.75	0.74	0.73	0.74	0.73
DS2	0.76	0.77	0.76	0.75	0.73	0.75	0.74	0.72

TABLE V. AUC FOR GRADIENT BOOSTING

Data sets	Imbalance ratio							
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8
German Credit data	0.69	0.75	0.74	0.74	0.73	0.75	0.74	0.72
Japan credit screening	0.90	0.91	0.90	0.90	0.90	0.89	0.89	0.89
DS1	0.80	0.81	0.78	0.77	0.80	0.81	0.80	0.79
DS2	0.74	0.76	0.74	0.77	0.84	0.84	0.82	0.81

The following tables, Tables VI, VII and VIII, provide a similar analysis summary, but using geometric mean (GM) as a performance metric instead of AUC. Table VI clearly shows that for logistic regression best results are achieved when class distribution ranges from 50%/50% to 30%/70%, or in other words when imbalance ratio is less than one. An exemption is one of the real life data samples, DS2, where the best result is achieved when class distribution equals 80%/20%. Neural network and gradient boosting exhibit similar results - best results are achieved when class distribution ranges from 50%/50% to 40%/60%, when imbalance ratio is less than one. For one of the real life data samples, DS2, the best result is achieved when class distribution equals 80%/20% .

It can also be noticed that all of the algorithms progressively become more accurate as class imbalance decreases. In addition, all methods show limited power in correctly classifying observations where a large class imbalance is present and only a small number of bad observations exist, while balancing class distribution might improve classification accuracy. It can also be noted that in some cases the gradient boosting algorithm did not classify a single case as a bad client, resulting in the TPRate being equal to 0% and consequently GM being equal to 0%.

TABLE VI. GM FOR LOGISTIC REGRESSION

Data sets	Imbalance ratio							
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8
German Credit data	0.13	0.28	0.53	0.67	0.70	0.68	0.60	0.47
Japan credit screening	0.70	0.79	0.83	0.84	0.85	0.85	0.85	0.84
DS1	0.14	0.49	0.65	0.70	0.73	0.69	0.59	0.50
DS2	0.56	0.81	0.63	0.68	0.66	0.64	0.58	0.48

TABLE VII. GM FOR NEURAL NETWORK

Data sets	Imbalance ratio							
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8
German Credit data	0.26	0.49	0.60	0.67	0.68	0.67	0.65	0.53
Japan credit screening	0.72	0.77	0.81	0.83	0.85	0.85	0.84	0.84
DS1	0.29	0.51	0.65	0.67	0.68	0.66	0.62	0.57
DS2	0.53	0.81	0.63	0.70	0.75	0.63	0.60	0.50

TABLE VIII. GM FOR GRADIENT BOOSTING

Data sets	Imbalance ratio							
	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8
German Credit data	0.07	0.37	0.56	0.61	0.68	0.69	0.63	0.50
Japan credit screening	0.68	0.79	0.84	0.86	0.86	0.85	0.85	0.85
DS1	0.11	0.32	0.00	0.00	0.74	0.72	0.65	0.53
DS2	0.56	0.81	0.63	0.68	0.66	0.64	0.58	0.48

It should also be noted that using undersampling techniques with DS2 generally increases prediction accuracy compared to the original distribution where the ratio of minority class examples is less than 1%, both in terms of AUC and GM. For example, without changing the original distribution GM amounts to 0% when using gradient boosting, as none of the bad clients are classified accurately, whereas GM of logistic regression and neural network is less than 0.2, which is significantly lower than when undersampling is used.

VI. CONCLUSION

In this study we analysed the effect of different class distributions on various classification algorithms' performance. Classification algorithms analysed include logistic regression, neural network and gradient boosting. Real life data samples were used in addition to publicly available datasets from the UCI Machine Learning Repository. Class distribution was altered to give various imbalance ratios, ranging from 90/10, where 90% of the cases in the training sample were majority class examples, to 20/80, where 20% of the cases in the training sample were majority class examples. In other words, the ratio of minority class examples, the bad clients, was artificially increased by a factor of 10%, ranging from 10% up to 80%, by randomly undersampling majority class observations (good clients) or randomly undersampling both majority class observations (good clients) and minority class observations (bad clients). Performance of the algorithms was assessed by using the area under the receiver operating characteristic curve (AUC) as well as geometric mean (GM), measures appropriate for imbalanced data samples.

The results show that changes in class distribution do not have a significant impact on classification accuracy when considering AUC as a performance measure. Specifically, logistic regression has proven its insensitivity to the balancing of the data sample, which can be attributed to its statistical efficiency. When considering GM as a performance measure, greater disparities can be observed, namely the fact that a reduction in classification accuracy can be observed, as class imbalance increases. All methods have shown limited power in correctly classifying observations when larger class imbalance was present, indicating that balancing class distribution might be helpful while trying to improve classification accuracy.

Leveraging on these findings, further work could be conducted, namely analyzing performance of other classification algorithms, such as support vector machines or random forests, yet to be fully researched in the context of credit risk assessment. This study focused on using undersampling techniques to alter class distribution. However, artificially altering class distribution might be addressed by using oversampling techniques, where the number of minority class instances, the bad clients count, is increased. It might also be interesting to analyse additional real life datasets with a larger number of minority class examples, given that the number of defaulting clients in the data samples used in this study was very small in absolute numbers.

VII. REFERENCES

[1] "Bank for International Settlements," "Principles for the Management of Credit Risk - final document," 2000. [Online]. Available: http://www.bis.org/publ/bcbs75.htm.

[2] B. Schmid, Credit Risk Pricing Models: Theory and Practice. Springer Finance, 2011.

[3] Q. Yang, "10 challenging problems in data mining research," vol. 5, no. 4, pp. 597–604, 2006.

[4] J. Van Hulse and T. Khoshgoftaar, "Knowledge discovery from imbalanced and noisy data," Data Knowl. Eng., vol. 68, no. 12, pp. 1513–1542, 2009.

[5] L. C. Thomas, D. B. Edelman, and J. N. Crook, Credit Scoring and Its Application. SIAM, 2002.

[6] V. S. Desai, D. . Convay, and J. N. Crook, "Credit scoring models in the credit union environment using neural networks and genetic algorithms," IMA J. Math. Appl. Bus. Ind., vol. 8, pp. 323–346, 1997.

[7] E. I. Altman and A. Saunders, "Credit risk measurement: Developments over the last 20 years," J. Bank. Financ., vol. 21, pp. 1721–1742, 1998.

[8] W. Chen, C. Ma, and L. Ma, "Mining the customer credit using hybrid support vector machine technique," Expert Syst. Appl., vol. 36, no. 4, pp. 7611–7616, May 2009.

[9] M. Doumpos and C. Zopounidis, Multicriteria Decision Aid Classification Methods. Dordrecht: Kluwer Academic Publishers, 2002.

[10] A. Capotorti and E. Barbanera, "Credit scoring analysis using a fuzzy probabilistic rough set model," Comput. Stat. Data Anal., vol. 56, no. 4, pp. 981–994, Apr. 2012.

[11] B. Baesens, T. Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring," J. Oper. Res. Soc., vol. 54, no. 6, pp. 627–635, 2003.

[12] Y. Peng, G. Wang, G. Kou, and Y. Shi, "An empirical study of classification algorithm evaluation for financial risk prediction," Appl. Soft Comput., vol. 11, no. 2, pp. 2906–2915, Mar. 2011.

[13] D. L. Tong and R. Mintram, "Genetic Algorithm-Neural Network (GANN): a study of neural network activation functions and depth of genetic algorithm search applied to feature selection," Int. J. Mach. Learn. Cybern., vol. 1, pp. 75–87, 2010.

[14] R. Malhotra and D. K. Malhotra, "Evaluating consumer loans using neural networks," Int. J. Manag. Sci., vol. 31, no. 2, pp. 83–96, 2003.

[15] P. Ravi Kumar and V. Ravi, "Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review," Eur. J. Oper. Res., vol. 180, no. 1, pp. 1–28, Jul. 2007.

[16] B.-W. Chi and C.-C. Hsu, "A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance

of credit scoring model," Expert Syst. Appl., vol. 39, no. 3, pp. 2650–2661, Feb. 2012.

[17] M. B. Yobas and P. Ross, "Credit scoring using neural and evolutionary techniques," pp. 111–125, 2000.

[18] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen, "Comprehensible credit scoring models using rule extraction from support vector machines," Eur. J. Oper. Res., vol. 183, no. 3, pp. 1466–1476, Dec. 2007.

[19] W. Chen, G. Xiang, Y. Liu, and K. Wang, "Credit risk Evaluation by hybrid data mining technique," Syst. Eng. Procedia, vol. 3, pp. 194–200, Jan. 2012.

[20] N. Hsieh, "Hybrid mining approach in the design of credit scoring models," Expert Syst. Appl., vol. 28, no. 4, pp. 655–665, May 2005.

[21] T. Bellotti and J. Crook, "Support vector machines for credit scoring and discovery of significant features," Expert Syst. Appl., vol. 36, no. 2, pp. 3302–3308, Mar. 2009.

[22] C.-F. Tsai and M.-L. Chen, "Credit rating by hybrid machine learning techniques," Appl. Soft Comput., vol. 10, no. 2, pp. 374–380, Mar. 2010.

[23] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets," Expert Syst. Appl., vol. 39, no. 3, pp. 3446–3453, Feb. 2012.

[24] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," Inf. Sci. (Ny)., vol. 250, pp. 113–141, Nov. 2013.

[25] S. F. Crone and S. Finlay, "Instance sampling in credit scoring: An empirical study of sample size and balancing," Int. J. Forecast., vol. 28, no. 1, pp. 224–238, 2012.

[26] G. Batista, R. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM Sigkdd Explor. Newsl., vol. 6, no. 1, pp. 20–29, 2004.

[27] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," J. Artif. Intell. Res., vol. 19, pp. 315–354, 2003.

[28] N. V. Chawla, D. A. Cieslak, L. O. Hall, and A. Joshi, "Automatically countering imbalance and its empirical relationship to cost," Data Min. Knowl. Discov., vol. 17, no. 2, pp. 225–252, 2008.

[29] C. M. Bishop, "Neural networks for pattern recognition," J. Am. Stat. Assoc., vol. 92, p. 482, 1995.

[30] J. H. Friedman, "Stochastic gradient boosting," Computational Statistics and Data Analysis, vol. 38, no. 4. pp. 367–378, 2002.

[31] SAS, "Getting Started with SAS," 2012.

[32] "UCI Machine Learning Repository." [Online]. Available: http://mlr.cs.umass.edu/ml/datasets.html. [Accessed: 03-Feb-2016].

Software Solution for Optimal Planning of Sales Persons Work based on Depth-First Search and Breadth-First Search Algorithms

E. Žunić*, A. Djedović*, B. Žunić**
* Info Studio Ltd, Sarajevo, Bosnia and Herzegovina
** Goethe-Institut, Sarajevo, Bosnia and Herzegovina
emir.zunic@infostudio.ba

Abstract - This paper presents and describes the practical usage of Depth-First Search and Breadth-First Search algorithms in the planning and optimization of sales persons work. Experiments for optimal implementation of these two algorithms for planning purposes are made through a specially developed MATLAB simulator. The application consists of two parts: Web application and Mobile application. Web application is developed using the Application Development Framework Technology (ADF), while for the development of the mobile version of application, Oracle ADF Mobile is used. Both applications are interactive with Google Maps, and based on the selected input parameters (such as the origin and destination of sales persons, the way of planning, any possible obstacles, etc.) they visually show the results, indicators and analysis. The commercial version of the software, eSalesmanPlan, which uses the aforementioned algorithms, is used in several companies in Bosnia and Herzegovina. All indicators point to significant savings both in human as well as financial resources and it will be also presented.

Keywords - Software, Simulator, Depth-First Search, Breadth-First Search, Sales Persons Planning

I. INTRODUCTION

The objective of this paper is to provide an introduction about graphs and the commonly used algorithms used for traversing the graph and their concrete practical usage. Breadth-First Search (BFS) and Depth-First Search (DFS) are the two popular algorithms asked in most of the programming interviews. Graphs are one of the most interesting data structures in computer science. Graphs and the trees are somehow similar by their structure. In fact, tree is derived from the graph data structure. However there are two important differences between trees and graphs:

- Unlike trees, a node can have many parents in graphs,
- The link between the nodes may have values or weights.

Graphs are good in modeling real world problems like representing cities which are connected by roads and finding the paths between cities (such as salespersons' problem which will be described in this paper), modeling air traffic controller system, etc. These kinds of problems are hard to represent using simple tree structures. A graph can be a directed/undirected and weighted/un-weighted graph. Every graph has two components, Nodes and Edges. This is the starting point for any software implementation of these algorithms. Nodes are implemented by class, structures or as Link-List nodes. Edges represent the connection between nodes. There are two ways to represent edges: Adjacency Matrix and Adjacency List.

The BFS and DFS are the two algorithms used for traversing and searching a node in a graph. They can also be used to find out whether a node is reachable from a given node or not. The aim of DFS algorithm is to traverse the graph in such a way that it tries to go far from the root node. Stack is used in the implementation of the DFS. The aim of BFS algorithm is to traverse the graph as close as possible to the root node. Queue is used in the implementation of the BFS. This is a very different approach for traversing the graph nodes.

As we can see, it is necessary to develop a programming environment which will use these algorithms to optimize the planning of commercial work. For this purpose, numerous tests of these algorithms have been performed in order to determine which algorithm in some situations shows better results. That is how a stimulator was created whose graphic results will be presented in the fourth chapter.

In the application of these algorithms, a big step was made using a concrete example of their application in solving one of the major problems in companies dealing with transport of goods with lots of commercialists working there. The problem is reflected in difficulties during the optimal planning of their everyday work. Therefore, we made the software which creates a presumption of optimal plan of commercialists' work on the basis of demonstrated knowledge and skills relating these two algorithms, as well as information about real-time locations. This software is being improved by adding new obstacles that could be found on optional path.

II. DESCRIPTION OF ALGORITHMS

Breadth-First Search is an algorithm for traversing or searching tree or graph data structures. It starts at the tree

root (or some arbitrary node of a graph, sometimes referred to as a "search key" [1]) and explores the adjacent nodes first, before moving to the next adjacent level.

BFS was invented in the late 1950s by E. F. Moore, who used it to find the shortest path out of a maze [2], and discovered independently by C. Y. Lee as a wire routing algorithm [3][4] (published 1961). BFS is one of the most popular graph algorithms, and forms the basis of a number of more advanced topics within graph theory. The main goal of BFS is to take a graph, denoted by $G = (V, E)$ where V represents the vertices in the graph and E represents the edges, and find the shortest path between the source vertex (analogous to the root node in the tree data structure) and all other vertices that are reachable from that source vertex. In this case, the shortest path is simply denoted as the number of edges travelled from the source vertex to the destination one. In doing so, BFS produces what is referred to as a breadth-first tree, where the source vertex is set as the root node and all reachable nodes are set as children of that root.

As indicated by its name, BFS works in a breadth-first manner, meaning that all reachable vertices at a specified distance, k, are discovered before all other ones reachable at $k + 1$. Algorithmically, this is done by "coloring" the vertices as they are visited. An unvisited vertex is denoted as a "white" one, vertices that have had all edges extended from them explored are marked as "black", and vertices that have been visited, but whose edges have not all been explored, are denoted as "gray." Therefore, a BFS algorithm allows for each reachable vertex in the graph to be explored, and the path between the source vertex and each reachable vertex to be tracked.

Depth-First Search is an algorithm for traversing or searching tree or graph data structures. One starts at the root (selecting some arbitrary node as the root in the case of a graph) and explores as far as possible along each branch before backtracking. A version of DFS was investigated in the 19th century by French mathematician Charles Pierre Trémaux [5] as a strategy for solving mazes [6][7].

DFS is similar to BFS. While the goals of each algorithm are the same, the ways we reach them are very different. As implied by its name, DFS searches as "deep" into a graph as possible before pulling back and exploring other paths. For example, if the source vertex has two adjacent edges, DFS will pick one and explore it fully before backtracking and moving on to the second.

DFS is also aided algorithmically by "coloring" each visited vertex, in the same manner as BFS. However, a vertex is not marked as being "black" until its adjacency list has been completely discovered and explored, which in turn can only happen when each vertex in that adjacency list has also had its adjacency list fully explored. Whereas BFS explores each current adjacency list at the same level, DFS picks one and explores it fully before moving on to the next one.

DFS has many attractive characteristics, such as the parenthetical structure property and the white-path theorem, each of which is useful when other algorithms attempt to extend the DFS algorithm.

III. DFS AND BFS ALGORITHMS INSTRUCTIONS

DFS and BFS are common methods of graph traversal, which is the process of visiting every vertex of a graph [8]. Stacks and queues are two additional concepts used in the DFS and BFS algorithms.

A stack is a type of data storage in which only the last element added to the stack can be retrieved. It is like a stack of plates where only the top plate can be taken from the stack. The three stacks operations are:

- *Push* – put an element on the stack,
- *Peek* – look at the top element on the stack, but do not remove it,
- *Pop* – take the top element off the stack.

A queue is a type of data storage in which the elements are accessed in the order they were added. It is like a cafeteria line where the person at the front of the line is next. The two queues operations are:

- *Enqueue* – add an element to the end of the queue,
- *Dequeue* – remove an element from the start of the queue.

Considering a given node as the parent and connected nodes as children, DFS will visit the child vertices before visiting siblings using this algorithm:

```
Mark the starting node of the graph as visited and push it onto the stack
While the stack is not empty
    Peek at top node on the stack
    If there is an unvisited child of that node
        Mark the child as visited and push the child node onto the stack
    Else
        Pop the node off the stack
```

BFS will visit the sibling vertices before the child vertices using this algorithm:

```
Mark the starting node of the graph as visited and enqueue it into the
queue
While the queue is not empty
    Dequeue the next node from the queue to become the current node
    While there is an unvisited child of the current node
        Mark the child as visited and enqueue the child node into the
        queue
```

Examples of the DFS and BFS algorithms are given next.

A. Examples of DFS Algorithm

An example of the graph which will be discussed in details is shown in Figure 1.

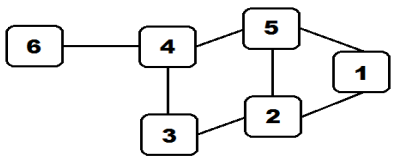


Figure 1. Example of the graph which will be analyzed

In the case of this graph, DFS algorithm will process the presented graph in the order shown in Figure 2.

Action	Stack	Unvisited Nodes	Visited Nodes
Start with node 1	1	2, 3, 4, 5, 6	1
Peek at the stack Node 1 has unvisited child nodes 2 and 5	1	2, 3, 4, 5, 6	1
Mark node 2 visited	1, 2	3, 4, 5, 6	1, 2
Peek at the stack Node 2 has unvisited child nodes 3 and 5	1, 2	3, 4, 5, 6	1, 2
Mark node 3 visited	1, 2, 3	4, 5, 6	1, 2, 3
Peek at the stack Node 3 has unvisited child node 4	1, 2, 3	4, 5, 6	1, 2, 3
Mark node 4 visited	1, 2, 3, 4	5, 6	1, 2, 3, 4
Peek at the stack Node 4 has unvisited child node 5	1, 2, 3, 4	5, 6	1, 2, 3, 4
Mark node 5 visited	1, 2, 3, 4, 5	6	1, 2, 3, 4, 5
Peek at the stack Node 5 has no unvisited children	1, 2, 3, 4, 5	6	1, 2, 3, 4, 5
Pop node 5 off stack	1, 2, 3, 4	6	1, 2, 3, 4, 5
Peek at the stack Node 4 has unvisited child node 6	1, 2, 3, 4	6	1, 2, 3, 4, 5
Mark node 6 visited	1, 2, 3, 4, 6		1, 2, 3, 4, 5, 6

Figure 2. DFS algorithm steps

There are no more unvisited nodes so the nodes will be popped from the stack and the algorithm will terminate.

B. *Examples of BFS Algorithm*

In the case of analyzed graph, BFS algorithm will process the presented graph in the order shown in Figure 3.

Action	Current Node	Queue	Unvisited Nodes	Visited Nodes
Start with node 1		1	2, 3, 4, 5, 6	1
Dequeue node 1	1		2, 3, 4, 5, 6	1
Node 1 has unvisited children nodes 2 and 5	1		2, 3, 4, 5, 6	1
Mark 2 as visited and enqueue into queue	1	2	3, 4, 5, 6	1, 2
Mark 5 as visited and enqueue into queue	1	5, 2	3, 4, 6	1, 2, 5
Node 1 has no more unvisited children, dequeue a new current node 2	2	5	3, 4, 6	1, 2, 5
Mark 3 as visited and enqueue into queue	2	3, 5	4, 6	1, 2, 5, 3
Node 2 has no more unvisited children, dequeue a new current node 5	5	3	4, 6	1, 2, 5, 3
Mark 4 as visited and enqueue into queue	5	4, 3	6	1, 2, 5, 3, 4
Node 5 has no more unvisited children, dequeue a new current node 3	3	4	6	1, 2, 5, 3, 4
Node 3 has no more unvisited children, dequeue a new current node 4	4		6	1, 2, 5, 3, 4
Mark 6 as visited and enqueue into queue	4	6		1, 2, 5, 3, 4, 6

Figure 3. BFS algorithm steps

There are no more unvisited nodes so the nodes will be dequeued from the queue and the algorithm will terminate.

IV. RESULTS OVERVIEW

Development and implementation of the entire system consisted of several steps and stages. The first experimental part consisted of programming of analyzed algorithms as well as creating simulators which will visually present their application. MATLAB was used for these purposes. The input parameter for each algorithm is a matrix showing the very shape of the graph, as well as the definition of initial and final graph node. For example, the matrix with seven nodes will be shown as Figure 4 in simulator.

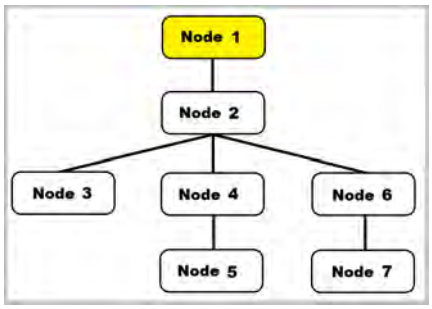


Figure 4. Simulator before algorithm search

DFS results will be shown here, since the procedure is identical for BFS too (they are complement). If in the following example the initial node is Node 1, and the ultimate one Node 7, then the steps will be as follows shown in Figure 5.

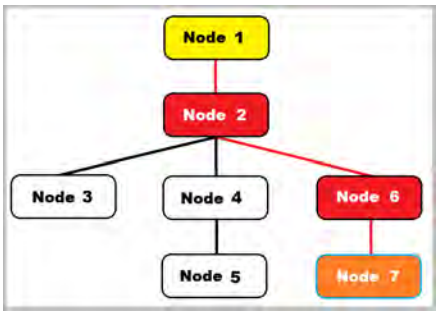


Figure 5. Simulator after algorithm search

Since this example is optimal and rarely used in practice, the more realistic problem containing certain obstacles from the initial to the ultimate node, as it is shown in the Figure 6, is being analyzed.

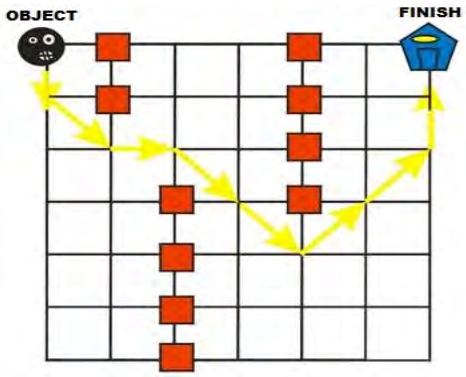


Figure 6. Example graph with obstacles

The object can move in paths north/west, west/east, and diagonally NW/NE/SW/SE in each node, or a section of the grid. The yellow line represents the path an object can achieve. The question is: How can an object reach the goal using the path bypassing the obstacles with? To solve this problem, we use DFS algorithm to produce the path of the object with given limitations. This example can be easily and in an identical way used during the movement and orientation of robots in the labyrinths. Solving this problem is of crucial importance, since the autonomous robots require this capability in planning their route.

Avoiding the obstacle is very important, especially if there is a possibility of robot damaging objects with strikes or vice versa. Programming of objects to move from the initial to the final location is usually more complicated than programming of rectilinear path, because there are obstacles between the initial and final point. There are several ways of avoiding obstacles used in generating the path, from the very simple ones of "changing the direction when reaching the obstacle" to using an A* algorithm which uses the information from the previous conditions and possible future ones for generating the path. DFS starts on the starting position of the object (the root node) and searches the target by expansion of all the following nodes (successors). The successors can only be those the object is allowed to go to, actually only those nodes that will not lead to the collision or removing the object from the working area. The Figure 7 shows all the directions the object can move in (8 directions).



Figure 7. Directions the object can move

If DFS comes to the point where it is impossible to find successors for the expanding node, it will move to the latest discovered one that has not been expanded yet. DFS will be easier to understand if we look at the tree and use the LIFO (last in-first out) structure. The structure contains a list of new nodes. The newest one will be on the top of the list. The new expanding node is taken from the top of the list and all its successors have been added to the list. When the goal is reached, the searching of the graph can go back through the tree and determine the shortest path or the path of the robot in the case of labyrinth. The simulator has also been developed for such cases of finding the path from the initial to the final node if obstacles are included.

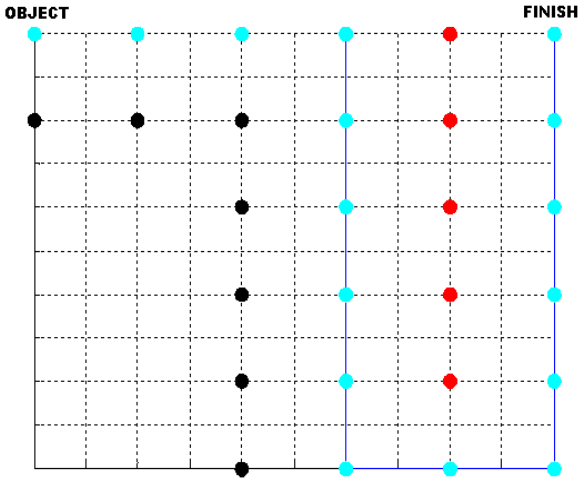


Figure 8. Simulator result with obstacles

The obstacles are presented by red dots, the searching limit by black ones, while the path from the initial to the final node is presented by cyan dots and the blue line.

The next thing we wanted to do is setting the obstacle in the way that algorithm reaches it, without any possibility of further progress and expansion, actually to be obliged to go back to the previous node and continue the searching of the space.

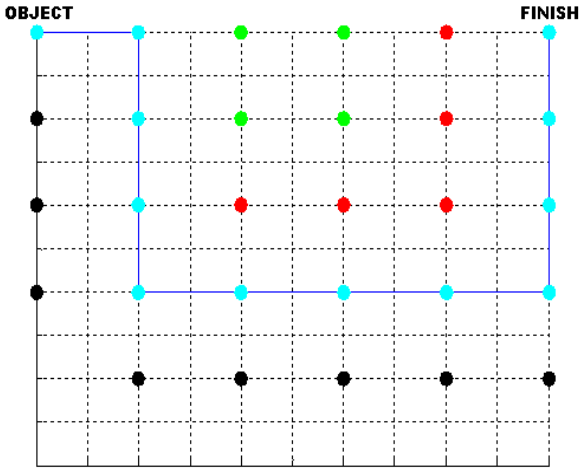


Figure 9. Simulator result with obstacles

The previous illustration shows the searching procedure and the path determination from the initial to the final node. We can also notice the green dots aside from the red, black and cyan ones, as well as the blue line. They explain that these nodes were notices during the search, but their further explanation was not possible because of the obstacle, so it was better to go back to some of the previous nodes in order to find the path to the target node. Initial simulator is made for the BFS algorithm also.

According to the searching graph results and finding the path from the initial to the final node and depending on the obstacles, a mechanism was made which, on the basis of comparative results of these algorithms determines which of them is better for an adequate situation. Criteria used for this purpose were the optimum path (fewer nodes) and the searching graph speed.

Criterion	BFS	Uniform Cost	DFS	Depth-Limited	Interactive deepening
Time	b^d	b^d	b^m	b^l	b^d
Space	b^d	b^d	b.m	b.l	b.d
Optimal	YES	YES	NO	NO	YES
Complete	YES	YES	NO	YES if l^3 d	YES

b: branching factor — m: maximum depth of the search tree

d: depth of the solution — l: depth limit (in Depth Limited Search)

Figure 10. Comparative results between DFS and BFS algorithms

According to the numerous tests, BFS algorithm proved to be superior to the DFS. But in some situations DFS is better than the BFS. The results of the analyses are shown in the Figure 10.

Then, we started to create the application for the practical usage on the basis of all the results. Web application is developed using the Application Development Framework Technology (ADF) [9], while for the development of the mobile version of application, Oracle ADF Mobile is used [10]. Both technologies being used are based on Oracle ADF components, as well as Java EE. Both applications are interactive with Google Maps, and based on the selected input parameters (such as the origin and destination of sales persons, the way of planning, any possible obstacles, etc.) they visually show the results, indicators and analysis. After the users logs into the application, they can see the screen where they, with the help of interactive map, chooses initial and final point (starting point and the destination) and possible additional obstacles that might exist in the database. It is important to understand that there are stored predefined values (coordinates) of eventual obstacles (rivers, mountains, etc.). The third step important for user to do is choosing additional parameters, such as the way of planning (DFS, BFS, or the system will automatically determine which algorithm is more favorable) and perhaps storing of additionally added obstacles in the database and expand it for future usage. That is the way of improving and expanding knowledge base for better result of this software.

Once the user has finished the initializing, the mechanism of creating the graph starts over the input data. It is of great importance to know that initial and final point are placed on the edges of the graph (as it has been shown on the very simulator), and network is being created between the mentioned edges with extremely high density. The nodes of the graph are created in these horizontal and vertical sections. Simplified results (Banja Luka - Sarajevo) are shown in the Figure 11. Using this software, the system administrator has the ability to prepare the work plan for any worker (commercialist) and to send it via mail and/or push notifications to the one the plan refers to. With his user application, commercialist has an access in such prepared plans he can behave on.

The commercial version of the software, eSalesmanPlan, which uses the aforementioned algorithms is used in several companies in Bosnia and Herzegovina. All indicators point to significant savings both in human as well as financial resources and it will be also presented. Thus, in one company using this software during the 2015 financial expenses were decreased by 7%, compared to the period software was not used in. Saving in human resources being very visible is important to mention. According to some analyses, out of 100 commercialists, two to three persons could stop doing this amount of work by the implementation of this system. This s showed that the detailed analyses of the problem and usage of these algorithms can practically lead to favorable and successful results.

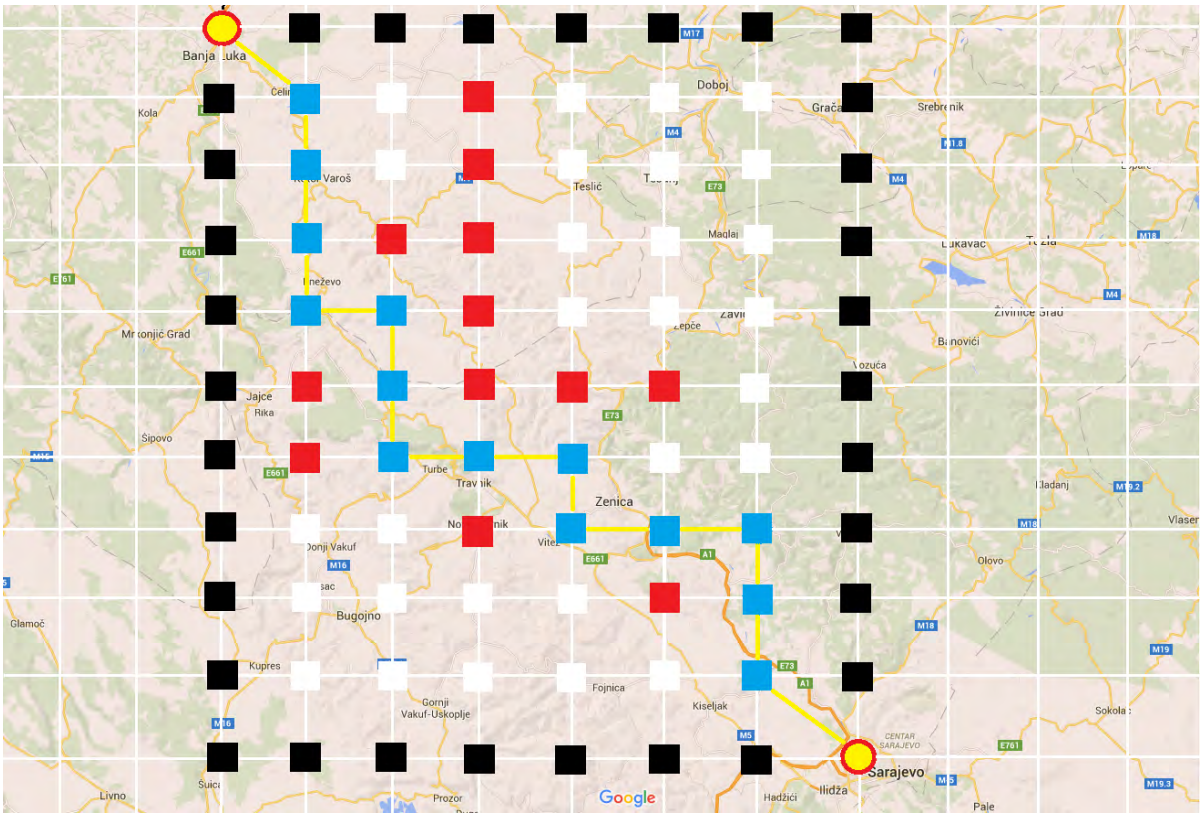


Figure 11. Web application, eSalesmanPlan, simplified result (Banja Luka – Sarajevo)

V. CONLUSION

Depth-First Search and Breadth-First Search are search algorithms used for graphs and trees. When you have an ordered tree or graph, it's quite easy to search the data structure to find the node that you want. But, when given an unordered tree or graph, the BFS and DFS search algorithms can come in handy to find what you're looking for. The decision to choose one over the other should be based on the type of data that one is dealing with.

BFS and DFS both are techniques that use searching for some solution in a state space given. Both of these construct spanning trees with certain properties are useful in other graph algorithms. Besides, both of them are very useful for directed graphs. Depth-First Search is "any search algorithm that considers outgoing edges of a vertex before any neighbors of the vertex that is, outgoing edges of the vertex's predecessor in the search. Extremes are searched first".

During the very implementation of the software, the realization of the simulator was the first step to find out the comparative results for the two algorithms. Comparing BFS and DFS, the big advantage of DFS is that it has much lower memory requirements than BFS, because it's not necessary to store all of the child pointers at each level. Depending on the data and what you are looking for, either DFS or BFS could be advantageous. The advantages of either vary depending on the data and what you're looking for.

On the basis of observed findings, we started the realization of the application that could have a practical application. In accordance to everyday problems of many companies dealing with this kind of work, it was found out that this software could be created for optimizing commercialists' work. It was done, so the applications (web and mobile ones) justified all the expectations and returned the invested funds within the first year of their application.

Beside the algorithms already mentioned, testing and using Dijkstra's algorithm for the problem is the next step in software development, directed graph and all weights must be non-negative It applies the greedy method pattern to the single-source shortest-path problem. For that reason it will be the next challenge in our next work.

REFERENCES

[1] S. Skiena, "The Algorithm Design Manual," Springer. p. 480, 2008.

[2] C. E. Leiserson, Schardl, Tao B, "A Work-Efficient Parallel Breadth-First Search Algorithm (or How to Cope with the Nondeterminism of Reducers)," ACM Symp. on Parallelism in Algorithms and Architectures, 2010.

[3] C. Y. Lee, "An Algorithm for Path Connections and Its Applications," IRE Transactions on Electronic Computers, 1961.

[4] S. Russel, P. Norvig , "Artificial Intelligence: A Modern Approach (2nd ed.)," Prentice Hall, 2003.

[5] C. P. Trémaux, "École Polytechnique of Paris (X:1876)," French engineer of the telegraph in Public conference, December 2, 2010 – by professor Jean Pelletier-Thibert in Académie de Macon (Burgundy – France) – Abstract published in the Annals academic, March 2011

[6] S. Even, "Graph Algorithms (2nd ed.)," Cambridge University Press, 2011.

[7] R. Sedgewic, "Algorithms in C++: Graph Algorithms (3rd ed.)," Pearson Education, 2002.

[8] A. Aziz, A. Prakash, "4. Algorithms on Graph," Algorithms for Interviews, 2010.

[9] S. O'Brien, S. Shmeltzer, "Oracle Application Development Framework Overview," An Oracle White Paper, June 2011.

[10] F. Desbiens, "Back-end to the Future: Using your Existing Oracle ADF Applications as a Pillar of your Mobile Strategy," An Oracle White Paper, December 2013.

Iterated Local Search algorithm for planning the sequence of arrivals and departures at airport runways

Eliot Bytyçi[1], Kadri Sylejmani[2], Agni Dika[2]
[1]Faculty of Mathematical and Natural Sciences, [2]Faculty of Electrical and Computer Engineering
University of Prishtina
Prishtina, Kosovo
eliot.bytyci@uni-pr.edu, kadri.sylejmani@uni-pr.edu, agni.dika.@uni-pr.edu

***Abstract*—In this paper, in order to tackle one of the dozens of the challenging optimization problems in aviation industry, we introduce an algorithm for planning the sequence of arrivals and departures of aircraft in a runway. The algorithm is based in Iterated Local Search (ILS) algorithm and uses two effective operators, namely swap and shift, for exploring the neighborhood of current running solution. The experiments that are conducted with real instances show that this approach achieves better results than the First Come First Served approach, while computing the results in relatively short period of time, which is in order of milliseconds.**

Keywords - air transportation, iterated local search, aircraft sequencing

I. INTRODUCTION

Air traffic control is the service provided to the aircraft by licenced air traffic controllers, with the primary aim to avoid collision between aircraft on the air and ground, and other moving vehicles on the ground. Air traffic controllers need to direct the aircraft to and from a specific airport. An aircraft prior starting asks permission from the air traffic controller to start and to taxi – move to the active runway, and then depart. After the permission, aircraft will start and will be ready for the journey. From this moment on, air traffic controllers will coordinate as smooth as possible the journey of the aircraft with minimum delays possible. The aircraft will taxi to the active runway that is used for departure depending on weather conditions, and will be ready for departure. But, in the same time, there could be one or more aircraft waiting to land in the same runway and therefore the air traffic controller will have to make decisions: first of all to accept the request of the aircraft to start, then to taxi and at the end to depart. If an aircraft waits for departure, a number of things will be affected: delay of the passengers, burn of fuel, CO2 emission and possible emergency situations that could emerge. Not every aircraft would have the same impact: some produce less CO2, some burn less fuel, some are aircraft that depart only for training purposes but on the other hand there are aircraft that are ambulances and need priority. A tool, that facilitates and takes into the account the above mentioned facts, is very welcome for the aviation community. It will create a sequence for the runway according to the importance of the aircraft.

Runway according to the International Organization for Civil Aviation (ICAO), is: a defined rectangular area on a land aerodrome prepared for the landing and take-off of the aircraft [1]. Most of the big airports today have more than one runway that they use for landing and take-off but even so, sometimes due to wind and weather factors they are limited to using one. And one runway in a busy environment means: more delay, increased percentage of possible emergency situation, higher environmental pollution and increased workload for both pilots and controllers. Therefore, there exists a need for a software tool that would allow the best possible usage of the runway and decrease the above mentioned factors.

With the number of air traffic forecasted to increase until 2035 by 50%, according to EuroControl "Challenges of growth" forecast released in 2013 [6], airports need to have more tools to help in dealing with the increased traffic. Since late 90's, aviation has started to actively use tools such as arrival manager (AMAN) and departure manager (DMAN) to assist with aircraft management especially during more challenging times, such as bad weather and runway closure [2], which produce more delicate situations and creation of problems. Furthermore, airports need to integrate both systems in order to facilitate the flow of air traffic more naturally. One such a system, with integrated AMAN and DMAN, would provide more benefits for the airports such as: predictability and stability of sequence, tactical scheduling, reductions on taxi time, cuts in runway delays and reduction in fuel and CO2 emission [4]. Such systems are being utilized every day and more, latest being the Istanbul Ataturk airport, which deployed those systems by the end of 2014 [3].

Even the airports that do not have any software in place for regulating departure and arrival sequence, they use a kind

of algorithm to serve the ones that come first, so a First Come First Serve (FCFS) algorithm. It means that according to the schedule of the aircraft, the aircrafts will be served. Nevertheless, that can change due to several factors such as delays of the aircraft (for which they should send a new schedule time) or weather conditions. Still, FCFS can serve also as basis for comparison with the new proposed implementation of the algorithms from the research community.

II. RELATED WORK

Aircraft landing problem was tackled by [11] based on the simplex algorithm creating a partial ordering list of landing aircraft and was used for both: an airport with single and multiple runways.

In [10] besides using three different algorithms for solving sequencing problems, they have taken into the account also a number of real life constraints important to the specific scenario. That has been done with the aim to minimise the workload of the pilot and controller.

In [5], authors proposed a Rolling Horizon algorithm for the optimization problem with the aim to reduce delays and also the workload of the air traffic controllers. They have – formulated the problem using Mixed Integer Programming Model and by using real world air traffic instances (12 instances) from Milano Linate Airport, and then compared results with First Come First Served (FCFS) sequence.

Another improved version of the Rolling Horizon algorithm was proposed by [9], in which the sequence of the aircrafts is partitioned in the chunks and each chunk is solved individually, where the respective chunk results are combined to produce the final solution. This approach proved to be more efficient then the approach where chunks were considered as whole.

III. MATHEMATICAL FORMULATION

In order to achieve results by using the Iterated Local Search (ILS), we have to fulfil some hard and soft constraints that are related to the nature of the specific problem - air traffic control. The hard constraints that ought to be fulfilled will help in achieving the soft constraint, which is the goal of the problem. Basis for mathematical formulation of the problem, which represents the core of the algorithm, is based on [5]. The soft constraint is to optimize the system and that by minimizing the cost (w), which is actually the waiting time that an aircraft will depart or land, based on their original schedule.

$$min \sum_{i=1}^{n} w_i (z^i - t_i^i) \quad (1)$$

Equation (1) represents the soft constraint and is represented by the sum of differences between actual arrival/departure time (z) and scheduled time (t), multiplied by the cost (w) of such a delay. The best possible scenario, according to equation (1), is when all aircraft arrive or depart on the same scheduled (requested) time. Due to the facts mentioned in previous sections, that is not always the case and especially not with the airport that have a high flow of traffic. Thus, the failure to comply to depart or arrive the aircraft at the specific requested time, has costly consequences. Cost is calculated according to many factors: weight of the aircraft, number of passengers, fuel burned, CO2 emissions etc., and ICAO has an actual list of the "weight" of the aircraft according above mentioned criteria.

Now, to achieve equation (1) several conditions have to be met:

1) First of all, every aircraft should have exactly one position in the sequence of the arriving and departing aircrafts. It is the first and the outmost condition to be taken into the account, when the algorithm starts functioning and the aircraft sequence is changed. This is formulated mathematically by the equation (2)

$$\sum_{s \in S, p \in P} x_s^p = 1 \quad (2)$$

where S and P, are sets of possible solution of the sequence and position of the aircraft.

2) Actual time of arrival / departure of the aircraft, should respect a lower bound, that is, it cannot be earlier than the scheduled time. This means that, an aircraft can arrive/depart at the exact scheduled time or later, depending on other conditions that it has or simple delays due to other reasons. It should be mentioned that in real time situations, due to shortcuts provided by air traffic control, some of the aircraft may come earlier than scheduled, but in our case we assume that all aircraft are flying the pre-planned route. This condition is represented by the following mathematical equation, equation (3), where aircraft can be assigned a time as schedule (t) or greater (d – as a possible delay):

$$a_p \leq t_p^p + d_p \quad (3)$$

3) If an aircraft changes the position, due to sequencing, its scheduled time should be taken into the account; not the time of the aircraft that had previously that position.

An aircraft can't arrive or depart in the same time as the previous arrival / departure. According to ICAO, there are well established separation minimums between arrivals vs arrivals, departures vs departures, arrivals vs departures and departures vs arrivals. These separation minima are taken into the account in our algorithm according to ICAO separation rules for the weight of the aircraft.

IV. ITERATED LOCAL SEARCH ALGORITHM

Iterated local search (ILS) algorithm [7] is based on the so called Hill Climbing algorithm. It tries to search for local optimum but in more intelligent manner by using the hill climbing technique. Every time a new better local optimum is searched and that is a repeated process, with the heuristic being in the fact that local optimum can sometimes be near the found optimum and in a way outperforming the

searching of new locations randomly from the beginning [7].

ILS is straightforward: it initially takes a solution as initial solution and the best solution at the same time. Then it tries to find better solutions by randomly changing the position of the elements in the sequence of initial set.

The pseudo code for ILS is:

Input: – (Maximum iterations, Initial iterations, Neighbour distance, Alternation frequency, Restart frequency)
Output: S_{best} *- best solution*
$S_c \leftarrow$ *random initial solution from sequence of aircraft*
HomeBase $\leftarrow S_c$
Repeat until maximum number of iterations not exceeded
T ←Near future random time from set of predetermined times
Repeat until time T not exceeded
$R \leftarrow$ *Random operator* (S_c) *– random application of swap and shift*
If Quality of R better than Quality of S_c
$S_c \leftarrow R$
If quality of S_c *better than quality of* S_{best}
$S_{best} \leftarrow S_c$
HomeBase ← New random HomeBase
$S_c \leftarrow$ *Perturb HomeBase*
Return S_{best}

A. Neighborhood exploration operators

In our algorithm, based on ILS, we have used two basic operators: shift and swap. Both of the operators were chosen randomly. In the shift operator, figure 1, the position of the aircraft is shifted for several positions in order to gain a new sequence, more cost effective.

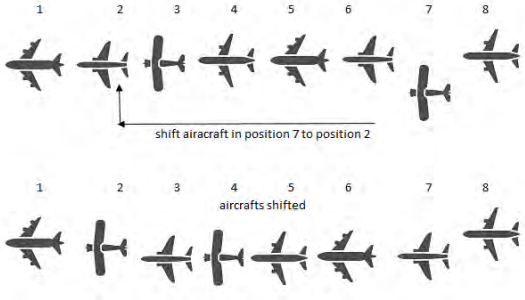


Figure 1. Aircraft shift operator

This affects all possible aircrafts that fall in line behind the first shifted aircraft. The swap operator on the other hand, affects only two aircrafts which position is changed, as seen in the figure 2. Shifting and swapping is done randomly but abiding the conditions described above.

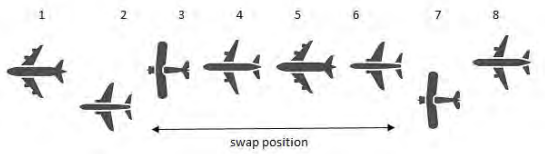


Figure 2. Aircraft swap operator

B. Perturbation mechanism

Perturbation, which is the part that differs ILS from the Hill Climb algorithm, was implemented using both of the operators: shift and swap and was used several times. The aim of the perturbation is to try other new optimums and not always fall in the same local optimum [7]. In our case, we have chosen a random variable, depending on which the perturbations consisted of shift function (several usages) or swap function (several usages) or both functions (used in half).

V. EXPERIMENTS

The approach in [5] compares its results against the first come first served approach (FCFS), but the results of FCFS approach presented in the paper do not correspond to the actual results of some the individual instances, namely Instances 6 to Instance 11. Initially when we created our algorithm, the idea was to compare it with the results of the algorithm proposed in [5] by taking the same instances used but we encountered difficulties due to the fact that some of FCFS results from algorithm [5] were not same as ours, especially the results from the middle instances (6-11). Therefore, we are unable to make proper a direct comparison of our results against the results in approach [5], hence we have also evaluated our results with against the actual results of FCFS sequence approach.

For testing purposes and gain of better results, we have designed the experiment, which was based on several initial tests, allowing to gain insight in best possible scenarios and parameters to be used. In the end, for final testing stage, we have chosen several parameters:

A. Machine, environment and programming language

All the experiments were performed in dual core computer, with i5 processor and processing power of 2.60GHz. The operating system of the machine was a Windows 8.1 version and the processor was 64 bit. The environment used was Netbeans 7.3 IDE, which run the algorithm programmed in Java language.

B. Instances

Instances used were taken from [8] and are the same instances used in [5], which are real traffic instances from Milano Linate Airport in Italy. In total are 12 instances, each having 60 aircraft, with the call sign, aircraft type (which can determine the importance of the aircraft), its category (which can determine the needed separation for the aircrafts), arrival or departure and its scheduled time of arrival / departure. The idea was to use instances generated by live traffic and

possibly used before, due to possible comparison between the results.

C. Design of the experiment

1) Initial iterations: which in fact determine the iterations that algorithm will perform. From initial tests, was concluded that raising the number of iterations more than 1000, will in fact result in no better results. The same was also concluded for the smallest number of iterations, 100, meaning that below that the results were approximate to FCFS algorithm. Therefore, we have chosen iterations from 100 to 800 for our experiments.

2) Neighbour parameter: which determines the distance between positions of the aircrafts in the sequence from one another. We have started with the nearest neighbours, 3, by analysing the sequence, from which we have concluded that mostly there are 3 aircraft that have the same scheduled time and therefore it makes sense to change their position. But, also other options were considered, like distance 5, then 3, 5 and 8 used jointly and in the end 2,3,5,8 and10. Above distance 10, based on preliminary results, made no sense to change the positions since the scheduled time was differing quite a lot.

3) Frequency: determines choosing inside of an optimum which determines the distance between: how many times to take the best result gained as a starting point and how many times to try and find another random starting point. We have chosen five frequency values from 10 up to 50, with increments of 10.

4) Restarting parameter: meaning to find another optimum and escaping from the local one. Values used were 25 up to 100, with increments of 25.

D. Results

Based on the parameters described in the previous section, we have performed the experiments. Initially we have concluded 10 experiments for each of the initial iterations described above and we have chosen their best results as shown in figure 3, where we can see the relations between best results and number of instances. So in this case, the best results were obtained from 400-800 iterations with 600 iterations being the best.

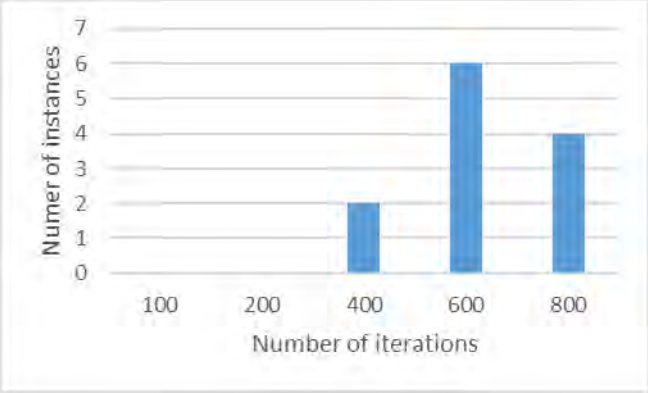


Figure 3. Results per number of iterations

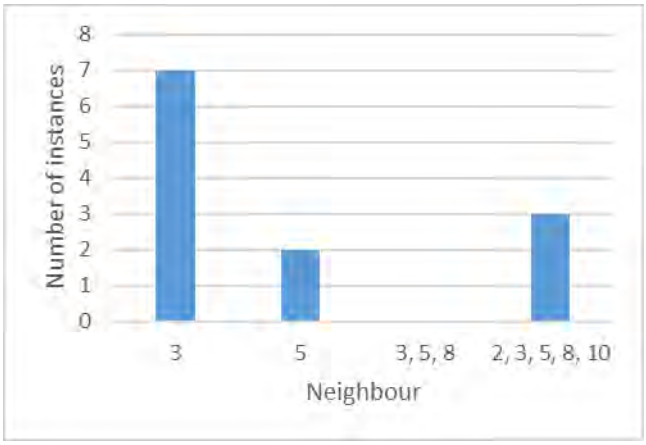


Figure 4. Results per neighbour

Then according to the best results of all iterations, we have chosen that result, to perform tests with different parameters. So in the figure 4, we see that the best parameter to be taken into considerations is 3, which is arguable because in most of the cases the number of the aircraft with the same time is around 3 or 4.

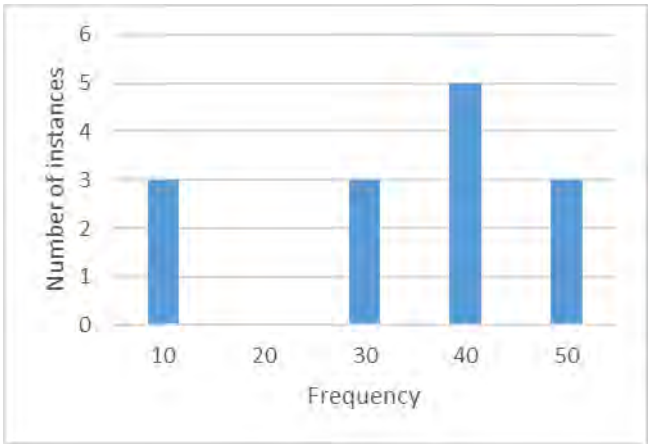


Figure 5. Results per frequency

Figure 5, gives a representation of the best possible results on the frequency of choosing best result or starting from another random initial solution. It is ambiguous, which is better, even though 40 is more stretched and therefore chosen for our experiment.

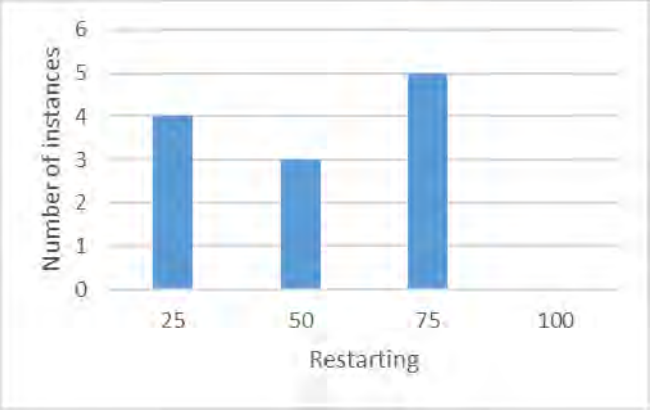


Figure 6. Results per restarting value

On figure 6, we see the best results for restarting value, which in our case is after 75 iterations.
Finally, when we have all the needed parameters and their outmost performance, we have conducted the final round of the tests and compared them with FCFS algorithm.

E. Comparing final results with FCFS results

Finally, our results were compared against the results of FCFS approach, and as we see in figure 7, we have improvements in the percentage and also as previously stretched, the results were computed in short amount of time, that is in order of milliseconds, which is acceptable for the air traffic controllers.

	FCFS	ILS	ms	Imp. %
Instance 1	401	329	35	17.96
Instance 2	457	387	39	15.32
Instance 3	393	328	55	16.54
Instance 4	480	414	39	13.75
Instance 5	440	383	20	12.95
Instance 6	516	302	48	41.47
Instance 7	581	375	51	35.46
Instance 8	346	271	25	21.68
Instance 9	336	296	45	11.90
Instance 10	370	324	49	12.43
Instance 11	394	348	75	11.68
Instance 12	269	222	21	17.47

Figure 7. Comparing results with FCFS

VI. CONCLUSION

Before conducting the experiments, on the live instances of the Linate Airport, we have conducted a series of tests in order to find out which parameters are best to be used but also their outmost performance. That being done, the final round of experiments have been conducted and evaluated against the first come first served algorithm. We have seen an average of 20% improvement, with some specific cases of improvement more than 40% improvement – most probably due to the big differences between aircraft timing.

As a future work, we intend to use clustering algorithms, in order to cluster the aircraft with certain time similarities and perform the actions only on those clusters. This would improve the performance, since the algorithm would not be trying for solutions with big time differences.

REFERENCES

[1] Aerodrome standards – ICAO, http://www.icao.int/safety/implementation/Library/Manual%20Aerodrome%20Stds.pdf, retrieved on 20.10.2015

[2] Eurocontrol arrival manager, http://www.eurocontrol.int/services/aarival-manager-aman, retrieved on 15.10.2015

[3] Eurocontrol airport corner, https://www.eurocontrol.int/airport_corner_public/LTBA, retrieved on 15.10.2015

[4] SESAR Joint Undertaking, http://www.sesarju.eu/sesar-solutions/traffic-synchronisation/dman-baseline-integrated-aman-dman, retrieved on 17.10.2015

[5] Furini, Fabio, Carlo Alfredo Persiani, and Paolo Toth. "Aircraft sequencing problems via a rolling horizon algorithm." Combinatorial Optimization. Springer Berlin Heidelberg, 2012. 273-284.

[6] Eurocontrol Challenges of Growth, http://www.eurocontrol.int/sites/default/files/content/documents/official-documents/reports/201306-challenges-of-growth-2013-task-4.pdf, retrieved on 13.10.2015

[7] Sean, Luke. "Essentials of metaheuristics." Lulu: Department of Computer Science, George Mason University, (2009).

[8] Bologna University Research Group - www.or.deis.unibo.it/research.html, retrieved on 14.10.2015

[9] Furini, Fabio, Martin Philip Kidd, Carlo Alfredo Persiani, and Paolo Toth. "Improved rolling horizon approaches to the aircraft sequencing problem."Journal of Scheduling (2015): 1-13.

[10] Ernst, Andreas T., Mohan Krishnamoorthy, and Robert H. Storer. "Heuristic and exact algorithms for scheduling aircraft landings." Networks 34, no. 3 (1999): 229-241.

[11] Atkin, Jason AD, Edmund K. Burke, John S. Greenwood, and Dale Reeson. "A metaheuristic approach to aircraft departure scheduling at London Heathrow airport." In Computer-aided Systems in Public Transport, pp. 235-252. Springer Berlin Heidelberg, 2008.

Energy Efficiency with Intelligent Light Management Systems

Ivan Britvić, Anja Nikitović
Visoka škola za informacijske tehnologije, Klaićeva 7, Zagreb, Croatia
ivan.britvic@vsite.hr, anja.nikitovic@vsite.hr

Improving our energy efficiency is one of most important technology goals set by EU Horizon 2020 agenda. Significant amount of energy is wasted in usage of sub-optimal lighting in home and especially in commercial sector.

This paper analyses recent developments in market of intelligent light systems as well as in Internet of Things technologies that could be utilized to achieve light savings. Solutions of leading light management system vendors are reviewed and their strong and weak points are listed.

Paper describes minimal requests for mathematical algorithms and offers practical guidelines and approaches for future development in light management area.

1. INTRODUCTION

1.1. Energy efficiency

"Energy efficiency delivering the same (or more) services for less energy and helps protect the environment. [1] When we use less energy, the less energy we need to generate at power plants, which reduces greenhouse gas emissions and improves the quality of our air." Energy efficiency helps the economy, by saving consumers and businesses millions of dollars in energy costs. Energy efficient solutions can reduce the energy bill for many homeowners and businesses by 20 to 30 percent. [2][10]

Energy efficiency is a fundamental component of our energy and environment policies. Cutting energy demand reduces greenhouse gas emissions in a more cost-effective way than any other energy or climate policy. The International Energy Agency's 2012 World Energy Outlook strongly featured energy efficiency as the key way forward. [3]

But energy efficiency has other environmental benefits than just climate protection. Local air quality, for instance, can be improved by reducing emissions through lower energy consumption

1.2. Lighting management systems

"A lighting control system is an intelligent network based lighting control solution that incorporates communication between various system inputs and outputs related to lighting control with the use of one or more central computing devices. Lighting control systems are widely used on both indoor and outdoor lighting of commercial, industrial, and residential spaces." [8] Lighting control systems serve to provide the right amount of light where and when it is needed.

Lighting control (management) systems are employed to maximize the energy savings from the lighting system, satisfy building codes, or comply with green building and energy conservation programs. Lighting control systems are often referred to under the term Smart Lighting. Also allow the management of operation of a lighting system by sending digital signals for switching on/off or reducing the quantity and color quality of the light emitted, depending on the actual requirements of the environment. Lighting conditions for areas may be modified using specific light control systems depending on the complexity of the environments and the type and number of light variations (scenes) considered necessary. [7][9]

Variations in light conditions can be obtained through programmed and automatic methods, or a combination of both. Light control systems allow the programming of a set of different scenes for luminaires to respond to the changed conditions. Automatic changes can be

activated by the light management systems using presence sensors, natural light sensors or manually with keyboards which, like a conventional switch, allow the user to retrieve the programmed scenes or to program others. “Amongst light control systems, those which allow variations in the color of the light in luminaires with RGB function, in particular LED luminaires, stand out.[1][3]

2. EXISTING LIGHT MANAGEMENT SYSTEMS

When analysing present market situation of lighting products we have to divide our research to two different segments:

- Industrial or more broadly referenced as commercial sector that includes various manufacturing and warehouse sites but also a public buildings like hospital and schools
- Private homes, including flats, small houses, garages, etc.

2.1. Industrial sector

Industrial light solutions in recent years are primary focused on implementation of LED technology equipment. Relatively high prices of bulbs and other hardware parts make this great revenue opportunity for lighting system vendors. New lines of products like General Electric LEDs [1], Cree industrial and warehouse lighting [2] or Philips manufacturing lighting [3] bring higher energy efficiency almost exclusively through lower power consumption. Software optimizations are not so interesting for large vendors probably because of lesser financial potential for their sales department. Also in their promotional materials are often listed paragraphs about intelligent buildings, smart cities and sensor and switch enabled lighting bodies yet specific products and solutions are not available. Instead these vendors, through own service departments, are eager to offer tailor made solutions with questionable price. Regarding above mentioned, penetration of existing lighting system in commercial spaces, excluding smart buildings, is basically negligible.

It is important to mention that there are some software solutions in commercial area, mostly based on two existing standards:

- KNX, network communications protocol for building automation, preferred in Europe [4]
- DALI (Digital Addressable Lighting Interface), standard made exclusively for lighting device management extensively used in North America [5]

Both standards aim to define data protocols, solving device addressing and discovery and enabling management from distance of compatible devices. There are number a products from vendors like Osram and Delmatic for DALI and formidable list of KNX vendors [6], including lot of small European vendors but also a global corporations like Philips and Panasonic.

The downturn of these two standards is that they expect wired connection between different elements though wireless and PoE (Power over Ethernet) communication is possible. Unfortunately majority of products are not made wireless so current systems are mostly implemented through wiring. For a new buildings this is a logical approach but wiring costs (in time and financial means) on an existing infrastructure are often reason for companies to avoid investments in smart lighting systems. However these protocols are open to possibility of wireless communication but the existing equipment usually doesn't support it.

2.2. Home market

On the other part of lighting market, home appliances are currently experiencing revolutionary changes. There are different home automation solutions and lighting management is integral part off all of them. Focus of these solutions is more on ambient light effects, like dimming, changing colors but energy saving component is also present. Example of typical schematic of smart home is shown on Picture 1:



Picture 1 – Smart home example [7]

Basic idea is that light (and other house appliances) should be remotely controlled through PC/mobile apps. Problem is, that although there are lot of products on market general connection, standard is not present. The products for smart home are using different basic communication platforms like:

- Xbee
- Wifi IEEE 802.11
- Z-wave
- Bluetooth 4.0 LE
- Power over Ethernet (PoE)
- Radio frequency (RF) on 433MHz and 868MHz (EU).

Crowding of home automation space with large numbers of products on different standards created market for new line of products, integration hubs. Task for hubs is to manage wide range of IoT devices like sensors, switches, cameras and give end user possibility to manually control elements but also to set a number of rules (e.g. switch off lights in living room when there is no occupancy). Usually user manages hub through mobile and web application. It has started with SmartThings acquired by Samsung in 2014 [8] but more products followed. In Table 1 we can see the comparison of most popular hubs on the market [9][10][11][12]:

Product	SmartThings Hub	Insteon Smart Hub Pro	Lutron Caseta Wireless Smart Bridge Pro	Wink hub
Compatible devices	Number of Xbee, z-wave and LAN devices from vendors like Aeon, Belkin, GE, Samsung [13]	Insteon devices, Nest support planned	Lutron, Serena, Nest	z-wave, Zigbee devices list including Philips Hue Bulbs, Honeywell, Wink

				products
Hub Management	Mobile app connected with cloud, need continuous Internet connection	Mobile app, there is version Apple Home kit compatible (voice control through Siri)	Mobile app, Apple Home kit (voice control through Siri)	Mobile app, Apple Home kit compatible (voice control through Siri)
Lighting Scenarios	Room lights switching off/dimming, rules, IFTT	Room and zone lights scenes and schedules, IFTTT	Rooms and zone rules, IFTTT	Basic rules, leans heavily on IFTT
Price (US market)	$99	$150	$120	$89

Table 1- Home automation hubs comparison

IFTTT ("If This Then That") is a web service that allows users to create basic rules, called "recipes", which are triggered based on changes to on Gmail, Facebook, Twitter accounts. User can define recipes on IFTTT website or mobile app. New rules are automatically synchronized with home automation hub, managing lighting or any other part of the smart home. [14]

New home automation hubs solve problems of interconnection of different devices and standards but have one essential flow: they are all closed solutions. Either no real software programming is possible, just basic configuration, or it is possible to use software API to program lights but only within vendor hardware ecosystem. Apple HomeKit standard also follows this pattern, it is possible to manage HomeKit devices and even create such applications but only for iOS devices (iPad, iPhone).

There is a rooted version of Wink hub [15] which could be used for development of intelligent light systems but this kind of hacking is legally questionable.

To conclude, lot of integration solutions had surfaced on market in last years but they are all targeting end users and not software developers.

3. GUIDELINES FOR SUCCESSFUL LIGHT MANAGEMENT SYSTEM

Despite the problem with existing solutions if we wanted to build intelligent light management system we still have some advantages that weren't present few years ago. There is a number of new sensor, light switches and other hardware equipment with attractive price and solid specifications. Software libraries that fill the communication gap with standards like z-wave, KNX, Bluetooth are present for number of programming languages like Java, C and Python. Eclipse Smartphone [16] is a new development initiative that promises platform for unification of programming interfaces for large number of devices. In future it could present a major standard in smart home area.

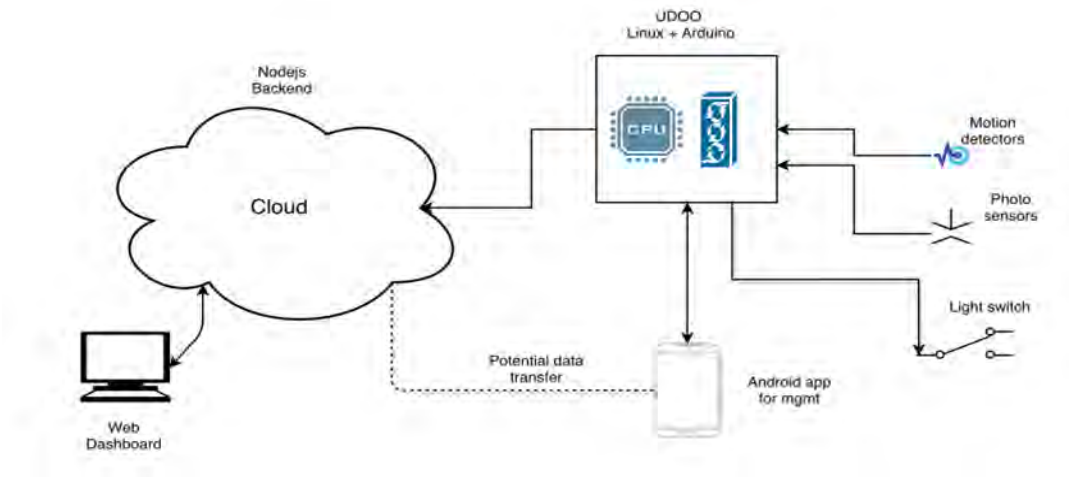
Last but not least, new IoT micro hardware platforms like Raspberry Pi, Linux credit size computer and Arduino, programmable microcontroller offer needed hardware infrastructure which is easy to place on production side and with almost negligible cost.

3.1. Experimental light management system

Having in mind all these new IoT and smart home developments on VSITE, College for Information Technologies we are currently experimenting with light management solution that could be potentially utilized in various commercial spaces. Our solution hopefully

can provide some guidelines for similar systems that could prevent unnecessary energy waste.

System architecture of the solution is shown on Picture 2:



Picture 2 – Experimental system architecture [45]

Hardware core of the system consists of UDOO board, born through Kickstarter project in 2013 [17], now very popular in IoT projects worldwide. Its main advantage is that it integrates the Raspberry Pi and Arduino unit on single board.

System is capable of detecting readings from remote sensors. This task is done in Arduino part of UDOO because of real time nature of platform and number of cheap hardware plugins and open source software libraries available for Arduino.

In Raspberry unit of UDOO real processing of sensor data is performed. Environment and configuration data are input in optimization algorithms which activate switch on/off, dimming actions.

3.2. Algorithms

Typical algorithms for light management systems include simple decision matrix which maps sensor readings and time of the day to specific switch actions.

But results from these kind of algorithms are not very different from rules used in a home automation hubs what imposes question is it worth of effort it if the almost same effect can be accomplished with packaged solutions.

But algorithm can be more sophisticated, bringing new advantages to the final system. Some of the interesting options for the optimization algorithm are:

- Self-learning (mainly of corrective actions from users like switch on when system switched off the lights)
- More advanced daylight harvesting and occupancy monitoring scenarios
- Recognizing false and incorrect sensor readings and alerting administrators on equipment failure
- Monitoring bulbs usage and calculating end of life for light bodies

Algorithms are not only limited to on-site boards, system should also be able to transfer data to the cloud database where data should be “crunched” to provide various reporting, dashboards, etc.

4. CONCLUSION

Recent home automation and IoT technology outbreaks are creating exciting possibility for development and implementation of highly intelligent light management systems. They could be primarily implemented in commercial sector, in building and warehouses, bringing higher energy efficiency and smaller carbon footprint, saving money and helping environment preservation efforts.

Existing complete home automation systems are still not ready for role of commercial lighting systems because of limited options, hardware support and closed design nature. But with combining single elements of these systems and other IoT hardware and software innovations today is possible build solutions that could match professional lighting systems. Strong driver for adoption of new systems should be their cost, it is expected to be be a fraction of cost associated to old professional solutions. However, this assumption should be tested in real production area where we could experience real total cost of ownership, test stability of system and measure energy saving effects.

References:

[1] http://www.iguzzini.com/Light_control_systems

[2] https://en.wikipedia.org/wiki/Lighting_control_system

[3] http://www.cpelectronics.co.uk/energy-saving-products/lighting-control-systems

[4] http://www.odyssee-mure.eu/publications/br/synthesis-energy-efficiency-trends-policies.pdf

[5] http://www.eceee.org/about-eceee/why_energy_efficiency

[6] http://www.legrand.us/lighting-controls.aspx

[7] http://www.cyberhomes.co.uk/what-we-do/automated-lighting-systems/

[8] http://www.energystar.gov/index.cfm?c=about.ab_learn_more

[9] https://www.smartenergygb.org/en/the-bigger-picture/about-the-rollout

[10] Energy efficiency in electrical systems Maheswaran, D.; Kailas, K.K.J.; Rangaraj, V.; Kumar, W.A.Power Electronics, Drives and Energy Systems (PEDES), 2012 IEEE International Conference

[11] http://www.gelighting.com/LightingWeb/na/solutions/industry/industrial/

[12] http://lighting.cree.com/applications/industrial-and-warehouse

[13]http://www.mea.lighting.philips.com/systems/system-areas/office-and-industry/manufacturing.html

[14] http://www.knx.org/knx-en/index.php

[15] http://www.dali-ag.org/

[16] http://www.knx.org/knx-en/manufacturers/list/

[17] Source: http://www.mccallsservices.com.au/img/home-automation.gif?a=162

[18] http://www.forbes.com/sites/aarontilley/2014/08/14/samsung-smartthings-acquisition-2/#712ac4957965

[19] https://support.smartthings.com/hc/en-us/articles/205956900-Meet-the-Samsung-SmartThings-Hub

[20] http://www.insteon.com/insteon-hub-homekit

[21] http://www.lutron.com/en-US/Products/Pages/SingleRoomControls/CasetaWireless/Overview.aspx

[22] http://www.wink.com/products/wink-hub/

[23] https://www.smartthings.com/compatible-products

[24] https://ifttt.com/categories/connect-your-home

[25] http://www.homeautomationhacking.com/c/openwinkhtm.htm

[26] http://www.eclipse.org/smarthome/documentation/index.html

[27] http://www.udoo.org/category/kickstarter

Adaptive and modular urban smart infrastructure

M.Klaric*, I.Kuzle ** and I.Livaja***
* Dalekovod d.d., Zagreb, Croatia
** University of Zagreb, Faculty of Electrotechnics and Computing, Zagreb, Croatia
*** Polytechnics of Sibenik, Sibenik, Croatia
email: mario.klaric@dalekovod.hr

***Abstract*—More and more countries and regions are trying to develop smart city projects in order to make transformation to cognitive and sensible cities and municipalities. Parallel to this process the aging of existing infrastructure also makes huge demands to public budgets regarding the need for large capital investments in reconstruction or maintenance. Latest developments in information and communication (ICT) industry created also large space for new kind of infrastructure very suitable for additional services as well as for update of existing ones. In this paper one solution of cost optimizing is described through reconstruction and expansion of public lighting systems connected with several static and dynamic ICT installations. These, existing infrastructures can be used as a basic platform where all additional IoT installations can fast and easily be done modularly and adaptively.**

Keywords—smart city, existing infrastructure, cost optimization

I. INTRODUCTION

Many communities in the world are facing new challenges when trying to develop various kinds of smart city [1] projects in order to make transformation to new kind or so called cognitive, smarter [2] and senseable [3] cities and municipalities. This demanding task gets even more complex when facing the aging of existing infrastructure which also makes huge demands to public budgets regarding the need for large CAPEX (capital expenditure) amounts in reconstruction and upgrade of public lighting, power system distribution and telecom installations.

On the other side, large OPEX (operating expenditure) costs and reliability costs usually demands urgent measures and decisions. Latest developments in ICT industry created large possibilities in the area of energy savings, traffic control, asset management and others. It also created space for new kind of infrastructure very suitable for additional services as well as for update of existing ones. Therefore various possibilities for the usage of ICT installations arise when analyzing its use in the existing municipal infrastructure.

Probably the most suitable existing installation that imposes itself for upgrade is public lighting system. These, mainly existing infrastructures can be used as a basic platform where all additional IoT (Internet of Things) installations can fast and easily be done very fast and without many problems. It is very convenient because in a lot of cases, existing public lighting systems are more than 20 years old. Furthermore, it is present in almost all parts of most important streets and squares in the public area, and not to neglect, it is already connected to power supply systems.

Also, all these installations can be curved out as a part of special Energy Management System through some kind of public private partnership (PPP), Energy Service Companies (ESCO) or Asset Management contracting.

Subject reconstructions in public lighting systems result in large energy savings. These additional power capacities can be used for the connection of additional power demanding devices. Probably the most suitable ones are electric vehicle (EV) chargers. These applications result as well in lower OPEX in general. Also, new kinds of ICT installations (EV chargers, WiFi Access Points (AP), Traffic control systems...) can create numerous opportunities for additional applications in marketing, public and social services. In this way they are very ESCO compatible and create additional free cash flows which can be a trigger for large boost of private investments in public infrastructure. In this paper, one such pilot project is described which integrates all these services through single spot cloud based solution. This project can be easily upgraded, scaled and multiplied as well as cut&pasted in many other industrial, touristic, urban or residential public areas.

II. SYSTEM DESCRIPTION

A. Fundamental framework

Fundamental system framework is structured as the integration of additional modules around existing basic infrastructure like energy grid, public lighting system, traffic control system etc. Main additional modules are shown schematically together with existing basic infrastructure in Figure 1.

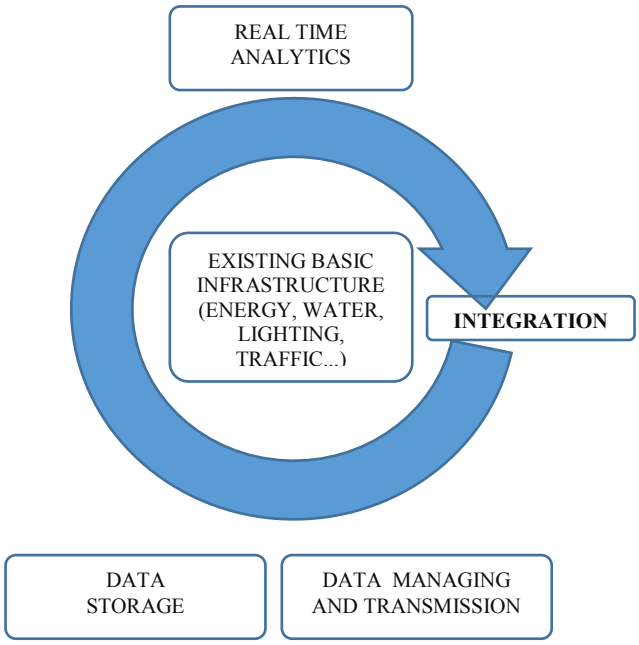


Fig. 1. Fundamental system framework

This fundamental system framework scheme can serve as a basic diagram for future strategy definition and roadmaps for each city, county or area. Every case ask for separate strategy building, solutions projecting and defining the steps to realize strategy and integration.

B. Basic architecture

Basic system architecture is defined as set of nodes clustered around gateway backbone sending crucial data and information to cloud based server. This architecture demands for strong telecommunication support (in most cases wireless telecommunications) but also must make a compromise in order not to be too expensive. For this reason it is important to identify the most suitable and cost effective technologies for such applications.

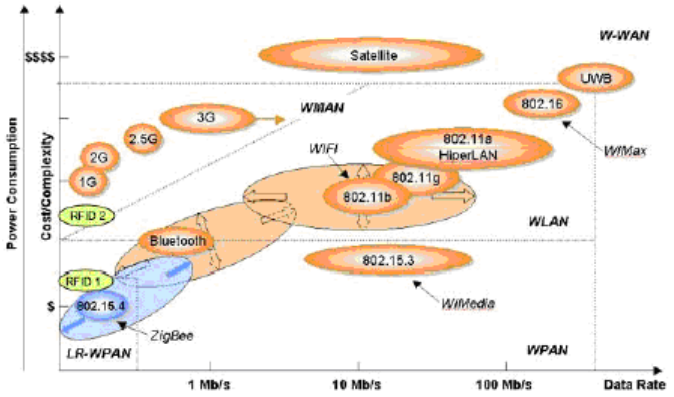


Fig. 2. Overview cost/complexity vs. data rates of wireless technologies

In figure 2 the overview of costs and data rates for various kinds of wireless technologies [4] are given. It can be seen from Table 1 that ZigBee and Bluetooth based technologies are very cost effective and useful for many applications [5].

TABLE I. COMPARISON BETWEEN ZIGBEE, BLUETOOTH AND WIFI

Standards	Bandwidth	Power Consumption	Protocol Stack Size	Stronghold
Wi-Fi	Up to 54Mbps	400+ mA TX	100+ KB	High Data Rate
Bluetooth	1 Mbps	40mA TX	~100+ KB	Interoperability, capable replacement
ZigBee	250 kbps	30mA TX	4”32KB	Long battery life, low cost

Since the number of nodes will most probably grow rapidly in the future, and since nodes will produce basic simple operations in most cases, for this primary level of communications a ZigBee, GPRS and Bluetooth based protocols are chosen. Other very suitable ones are Sigfox's, OnRamp Wireless, LoraWan [6] and others. The implementation of the 6LoWPAN protocol is one of the most innovative features, and it is recommended for low-power wireless and high security (128-bit AES encryption) applications. This communication will most probably be dominant for security demanding applications.

So, typically the basic architecture is made as wireless sensor network (WSN), built of nodes from a few to several hundreds or even thousands, where each node is connected to one or several sensors. In most applications, a WSN communicates with a specific Local Area Network or Wide Area Network through a gateway or similar data concentrator device. The Gateway serves as a link or a bridge between the WSN and the other network. In this way data to be stored and processed by devices with more resources, for example, in a remotely located server. Gateways can also have some kind of temporary storage that can be used for emergency conditions.

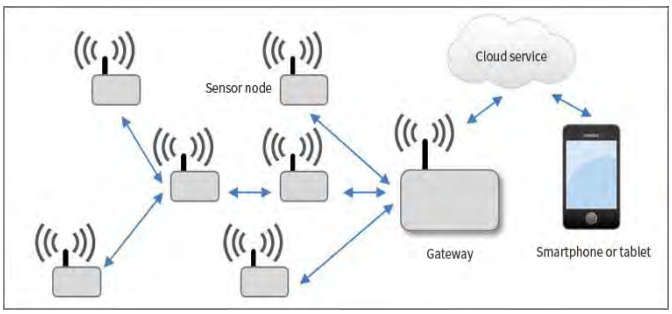


Fig. 3. Basic system architecture

In this way, like it is described in figure 1, the existing basic infrastructure like energy grid, water system or public lighting can be equipped with numerous kinds of sensor transmitting large streams of live data to data servers where they can be stored and analyzed producing useful information

from this big data structures. This new kind of infrastructure is very adaptive and modular (so called "adaptidular") towards existing infrastructure. Since in many cases existing infrastructure faces the challenges of aging and large CAPEX needs, this new sensor network architecture can serve in many cost optimization services.

C. System benefits

The most important benefits of this new kind of infrastructure is described in figure 4.



Fig. 4. Most important smart city system benefits

Also, in this way the main frame or smart city backbone is designed which can serve for all additional future applications and services.

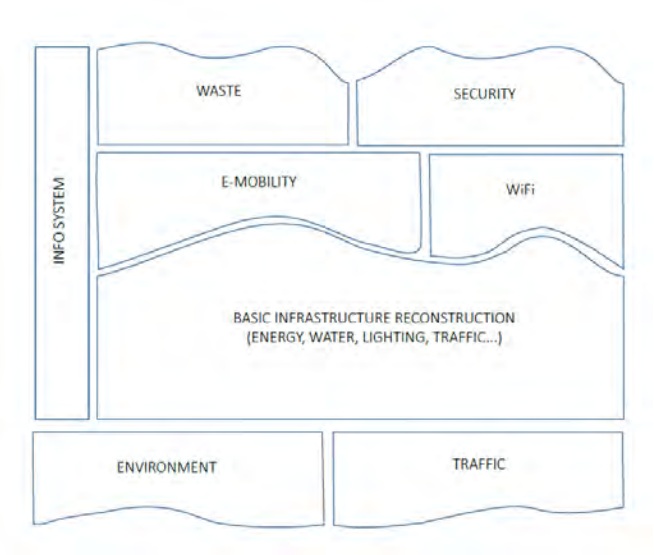


Fig. 5. Adaptive and modular (so called adaptidular) infrastructure characteristics

D. Additional dynamic ICT installations

All these installations and applications can be upgraded with so called dynamic ICT installations which are not installed firmly in one place but are placed on moving objects like public transportation vehicles, waste management trucks, courier services, e-bike sharing systems etc.

Also, through WiFi Access Points, additional smartphones, PDA's or computers can be connected to this network making it a suitable BYOD (Bring your own device) infrastructure.

III. Pilot Project

One pilot project was developed in order to describe the possibility of such projects in wider scale. As described earlier, it has all the characteristics of adaptive and modular infrastructure that can easily be installed, commissioned and later upgraded and integrated in large scale applications. This pilot project was developed by company Dalekovod and its subsidies, namely company Dalekovod ESCO as a part of energy services business development.

A. Current state of infrastructure

Infrastructure that is defined as a part of this scope is some 25 years old public lighting system that was used for lighting of outdoor surfaces in an industrial facility. Since there was a huge lack of information and documentation about the subject infrastructure, a quality energy audit was conducted in order to specify the current state and the necessary reconstruction measures. During the energy audit, measurements of electric power were made during two weeks in order to measure reference costs for future savings calculations. The measurements results are given in following figure.

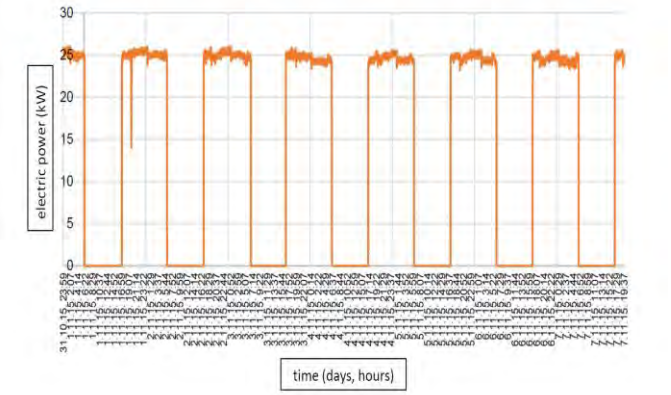


Fig. 6. Measurements of electric power in tested public lighting system

It can be seen that at this part of the infrastructure, peak power was 25 kW. Also, it should be stressed, that the existing system was designed by old lower lighting standards which were approximately 20% lower than the today's existing ones. Not to be neglected that some 20% of the existing lights were not functioning. During the energy cost summarization it was seen that during the past few years, due to the crisis and lower engagement, there was a total system shutdown of approximately 20% (the facility was not working designed 4100 hours yearly but around 3500 hours). In this way, a large reference cost curtailment showed which makes also smaller difference between the existing costs and future designed costs. Energy auditor suggested to reconstruct lighting system with

replacement of old halogen and sodium lights with modern LED lights and solutions.

B. *Project scope and design*

During the design phase, new LED lights were planned on the existing mounting poles. Since the lifetime of the poles and underground cables is estimated on 40 years, only some minor painting and corrosive protection works were suggested in the scope of civil works. The newly designed peak load was calculated at 8 kW, meaning 1/3 of currently installed power.

In this way, 2/3 of currently installed power was still not allocated for some additional services and devices to be connected to grid leaving much of the space for some new consumers like EV chargers, traffic controls and others.

Therefore, it was decided to define such additional devices to be connected at the same electric grid to from small microgrid [7]. These devices comprise of EV chargers, WiFi Access Points and parking control system. Also, it was defined to install smart meter and meteorological station. Basic Machine-to-machine (M2M) network is thus organized.

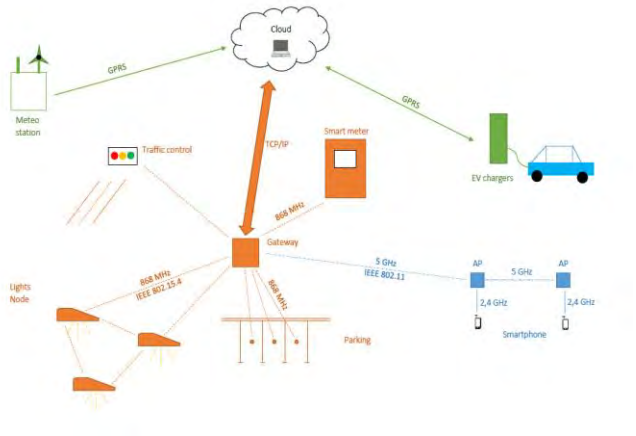


Fig. 7. Pilot project design

It can be seen form figure 7. that all these systems are connected in combined communication network where all the nodes in the network are sending data to the cloud based application making the intelligent automated infrastructure. In this pilot project it is practically negligible amount of sensor nodes so there is no need for some complex data storage system or big data application in cloud environment. However, in this way, one fundamental approach is described and made, which can be later cut&pasted or scaled, multiplied and upgraded with additional number of sensor nodes, gateways, appliances, data storage and similar.

Data storage in this pilot project was not such a problem due to small amount of sensors and data throughput but in the case where a large city would come into account, some data warehousing should be taken into consideration, especially in cases with increasing amount of dynamic ICT installations.

C. *Expected results*

One of the most important pilot project characteristics is the usage of existing infrastructure in greater scale. This project is expected to be made without additional civil works like foundations, cable trenches, excavation works and similar. Also, the existing right-of-way is used which is often very important in order not to define any new corridors which usually makes a lot of problems with corridor defining. In this way huge CAPEX amount was saved making the project very prone for shorter payment period.

However, one of the main issues was the referent cost estimation, namely energy costs and maintenance costs in order to calculate project payback indicators. The biggest difficulties in reference cost estimation was that the existing system had additional problems described in chapter III-A which reflected in additional reference costs curtailments, namely:

- lighting standard lifting costs
- availability costs
- lower working hours costs

Total reference cost curtailment can be calculated as follows:

$$\sigma = \sigma 1 \times \sigma 2 \times \sigma 3 \qquad (1)$$

- σ – total reference cost curtailment
- $\sigma 1$ – reference cost curtailment due to lighting standard lifting
- $\sigma 2$ – reference cost curtailment due to lower availability
- $\sigma 3$ – reference cost curtailment due to lower working hours number

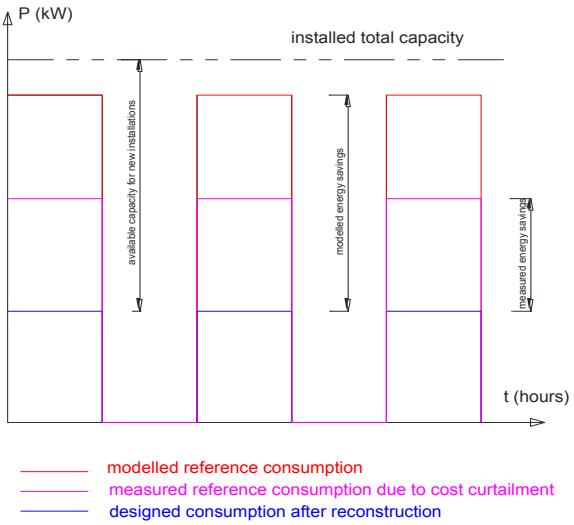


Fig. 8. Energy savings through designed reconstruction

Since each of these reference cost curtailments is in average (100%-20% = 80%) according to chapter III-A, this gives a total reference cost curtailment according to (1):

$$\sigma = 80\% \times 80\% \times 80\% \cong 50\%$$

Reference cost curtailment of 50% gives approximately 2 times longer investment payment period, so from initially planned 4-5 years payback period, realistic values climb up to 8-10 years. This makes additional challenge for decision makers to approve such investments.

However, large energy cost reduction, as well as neglecting future maintenance cost, give results of internal rates of return from around 12%-15% making it suitable for investments when taking into account weighted average cost of capital (WACC) in intervals 8-10%. [8]

Not to neglect is the possibility of additional applications in marketing, public and social services. These services can create additional free cash flows which can be further optimized through Energy Performance Contracting (EPC) or some other kind of energy services [9] like Asset management, PPP, leasing service and similar.

Marketing services can be very lucrative in this area. For example, a tourist can cross the border or enter the city with his electric car and a suitable application can give him most desirable place for his vehicle charging on a highway route or local road to his destination and also a list of restaurants, sights and shops that he can visit while his car or bike is charging [10]. Charging place for his EV can also be reserved at a desired time of arrival with assigned parking place, billing method etc.

Further possibilities are to make freely available some parts of newly produced data to API programmers for building applications that are available to the members of the social networks. It will allows applications to use the social connections and profile information to make applications more involving, and to publish activities to the news feed and profile pages of social networks, subject to individual users privacy settings. Since most of these databases use interfaces like SOAP, REST, JDBC and similar, it can easily be connected with most important social network services.

IV. Conclusions

In this paper one solution of cost cutting is described through reconstruction and expansion of urban infrastructure systems connected with several static and dynamic ICT installations. Adaptive and modular (so called adaptidular) ICT installations were suggested and designed for the subject reconstruction of public lighting systems and making of the new smart city backbone. It can be seen that existing infrastructures of public lighting, water supply and energy grid can be used as a basic platform where all additional ICT installations can fast and easily be done modularly and adaptively thus making a basic new infrastructure for all future expansions.

In this paper one example of such pilot project design is presented. It can be seen that, although there are problems with detecting true energy costs in existing situations in many projects, future benefits from these kinds of installations can achieve savings and additional revenues which are big enough for investment returns. Also, in this way it can be very attractive area for private investment boost in infrastructure through various kinds of PPP, EPC or Asset Management contracting.

Also, this kind of infrastructure is very appropriate for all other utilities. Maybe the most dominant role will be in electric grid in general where it will lead to a new paradigm of power system management and control where substitute models already evolved from deterministic to stochastic, and now will evolve further to data mining models.

All these activities can start the process of integration of Internet of Things (IoT) which would lead to Internet of services (IoS) and later to Internet of Everything (IoE).

References

[1] C. Manville, G. Cochrane, J. Cave, J. Millard, J.K. Pederson, R.K. Thaarup, A. Liebe, M. Wissner, R. Massink, B. Kotterink, "Mapping Smart Cities in the EU" January 2014, European Union., This document is available on the Internet at: http://www.europarl.europa.eu/studies

[2] S. Dirks, C. Gurdgiev, M. Keeling, , "Mapping Smarter cities for smarter growth", IBM Institute for Business Value., May 2010, IBM Corporation

[3] M. Claudel, C. Ratti, "Dimensions of the Future" August 2016, Senseable City Lab, Massachusetts Institute of Technology, Cities in the 21st Century: Academic Visions on Urban Development, Abingdon: Routledge

[4] F. Cleveland, "Use of Wireless Data Communications in Power System Operations," in Proc. 2006 Power System Conf. and Expo., pp. 631-640." unpublished.

[5] A. Kanwar, A. Khazanchi, "ZigBee: The New Bluetooth Technology" International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume1 Issue 2 Nov 2012 Page No. 67-74

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Low Throughput Networks for the IoT: Lessons learned from industrial implementations", Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on, pp. 181 - 186, Milan, December 2015.

[7] N. Holjevac, T. Capuder, I. Kuzle, "Adaptive Control for Evaluation of Flexibility Benefits in Microgrid Systems", Energy journal, Vol. 92, Part 3, December 2015, pp. 487-504

[8] Aswath Damodaran Database, last updated January 2016. http://pages.stern.nyu.edu/~adamodar/

[9] Graz Energy Agency Ltd with input from Berlin Energy Agency, "Comparison and evaluation of financing options for Energy Performance Contracting Projects", Graz Energy Agency Ltd and EC authors, 2008, Available at https://ec.europa.eu/energy/intelligent/projects/sites/iee-projects/files/projects/documents/eurocontract_epc_financing_manual_en.pdf

[10] I. Pavić, T. Capuder, I. Kuzle, "Value of Flexible Electric Vehicles in Providing Spinning Reserve Services", Applied Energy, Vol. 157, November 2015, pp. 60-74

Automatic Pathole and Speed Breaker Detection using Android System

Vinay Rishiwal*, Hamshan Khan **
Department of CSIT, IET, MJP Rohilkhand University, India.
Tejas Networks, Bangalore
Email: *rishi4u100@gmail.com, **hamshankhan@gmail.com

Abstract - Recent studies have shown that road surface monitoring is essential for the municipal corporations as well as for travelers for choosing the best road possible. Such schemes provide the comfort and security to the vehicle travelers. This paper presents a vibration based approach for automatic detection of potholes and speed breakers along with their co-ordinates. In this approach, a database is maintained for each road, which is made available to the public with the help of global database or through a portal. Potholes and speed breakers are detected along with their severity using android's built-in accelerometer. The results of the proposed approach are tested over a 4 km flat road and compared to manual inspection of pothole and speed breakers on the same considered road. The accuracy of the proposed approach came out to be 93.75% for detection of potholes and speed breakers. This approach is cost efficient and very effective for road surface monitoring.

Index Terms: pothole, speed breaker, road surface monitoring, accelerometer

I. INTRODUCTION

Road surface monitoring is essential for municipal corporations for quick detection and maintenance of potholes, detecting the other abnormalities of the road such as speed breaker rail road crossing etc and making them accessible to the travelers will provide a better driving experience. If such information of all roads is put on a central server which could be accessed by anyone freely, the drivers can choose the best possible road from source to destination.

Road surfaces can be classified into different categories such as smooth roads, potholes, bumps, contraction joints, man holes, expansion joints etc [1]. The surfaces where one has to slow down his speed are potholes and bumps (also known as speed breakers). So to detect a speed breakers and putting them on to the server can provide the driver prior information about the roads so they could be able to roughly figure out what time would be taken to complete the journey.

Currently road surface monitoring is carried out manually by means of employing persons for this purpose only, or by volunteer drivers locating any pothole and reporting it to the concerned authorities. But this approach of road condition monitoring is slow as well as costly. Alternative to this there are two approaches for automatic detection of potholes, one is a responsive system and other is unresponsive system. The unresponsive system makes use of optical and acoustic sensors [2]. These sensors are costly and they also need a complex algorithm to detect the potholes. Special purpose vehicles would have to be employed because devices have to be mounted over the vehicle so the condition of road is clearly visible. Another approach is responsive system which uses inertial sensors (like accelerometer and gyroscope) to access and analyse the irregularities.

To make the system of road monitoring attractive to the users, this has been implemented as a service similar to Waze [3] which makes use of little system resources and does not add extra overhead to the volunteers. This service of Road monitoring would run on devices operated by Android OS [4]. Android OS is extensively used system now a days, which makes our system attract more volunteers for the better road monitoring.

This paper discusses about the monitoring of road condition using android's built-in accelerometer, which records 3-axis acceleration [5]. The data obtained by the accelerometer is processed with suitable approach to detect the type of event along with their severity. Location co-ordinates of the place where the pothole or the speed breaker has been detected recorded using android's built-in GPS system [6].

These events along with their severity and the location co-ordinates are stored in the local database of the mobile device. This local database is synchronized with the global database as soon as the session ends. The synchronization is done in such a way that the previous road surface condition is replaced by the current one

II. RELATED WORKS

A lot of work has been done in this area of automatic pothole detection using 3-axis accelerometer and a GPS system. Some of them use embedded accelerometer sensor and GPS device. This section discusses various researches performed and different algorithms used to detect the potholes and their drawbacks.

MIT provided a road surface monitoring system using Smartphone technology called P^2 (Pothole Patrol) in 2008 [1], which uses smart phones accelerometer sensor to monitor the changes in acceleration while driving and then analyzing the accelerometer data to find

out potholes. They used 7-taxis in Boston area as the experimental vehicles. After collecting the on-board data of accelerometers and GPS sensors, P^2 passes the data to different data classifiers to extract the exact pothole information. Their algorithm is based on simple machine learning approach. They passed the x and z axis acceleration data through the different filters namely: *speed, high-pass, z-peak, xz-ratio ad speed vs. xz ratio*. Each of them removes one or more type of road abnormalities causing changes in acceleration such as speed bumps, rail road crossings etc.

Chan Kang et al. [7] improvised the idea of P^2 by introducing extra attitude information from gyroscope sensor to distinguish between manholes and potholes and it is referred as PDGPA (Pothole Detection using On-board gyroscope, accelerometer and GPS).

Microsoft Research India also developed Nericell [8] and TrafficSense [9] using Windows Mobile OS powered Smartphone as hardware/software platform with an array of accelerometer (sampling rate 310 Hz), and GPS. Their algorithms are based on the simple threshold approach *z-sus* (for speeds <25km/h) and *z-peak* (for speeds ≥ 25 km/h). They also used virtual re-orientation to adjust the orientation of the Smartphone during driving.

Mendis et al [10] have presented a Real Time Pothole Detection system using Android Smartphones with Accelerometer. This algorithm is similar to *z-peak* algorithm used in P^2. Algorithm has been extended using a factor *Z-DIFF* as the main classifying factor for potholes and other road abnormalities. This is a measurement of change in acceleration between two consecutive records of accelerations. Although this algorithm was able to address many issues for accurate detection of potholes but it does not give any idea about the severity of potholes. Authors only identify that whether the pothole is large or it is a cluster of potholes.

Researchers from University of Jyväskylä designed an offline data mining technique for pothole detection [11]. They pre-processed raw data collected by accelerometer using band-pass filters, further they used the feature detection algorithms such as mean, peak-to-peak ratio, standard deviation, root mean square, variance wavelet packet decomposition and power spectrum density. Further to reduce the features backward and forward selection and genetic algorithms were used. Although this approach shows a good performance in detecting the potholes but this is too complex to be implemented over small devices like Smartphone because of less processing power and small memory size.

Although all of the techniques discussed in above researches provided a good mechanism to detect the potholes but they did not given any idea about the severity of the potholes. Our algorithm would mainly focus on following two aspects:

i. Real time detection of potholes and speed breakers along with their location co-ordinates.
ii. Severity of Potholes as well as speed breakers on the scale of 3.

III. PROPOSED WORK

Primarily, the 3-axis acceleration data was collected from the built-in accelerometer of an android device and applied a continuous monitoring of z-axis acceleration. When any pothole or a road breaker is present it results into sudden changes in the acceleration along z-axis, which is compared against z_{th} and appropriately classified into different categories.

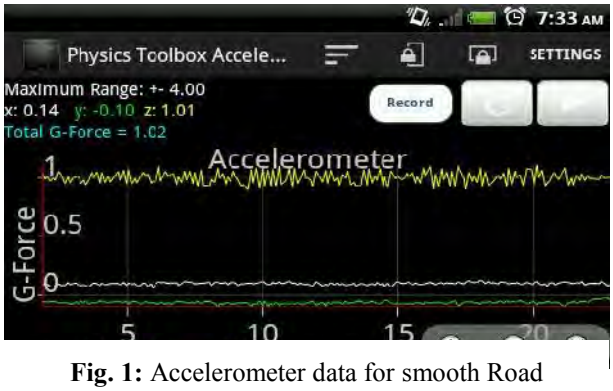


Fig. 1: Accelerometer data for smooth Road

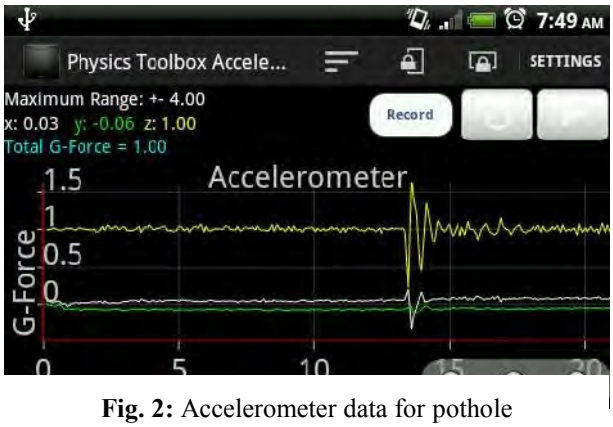


Fig. 2: Accelerometer data for pothole

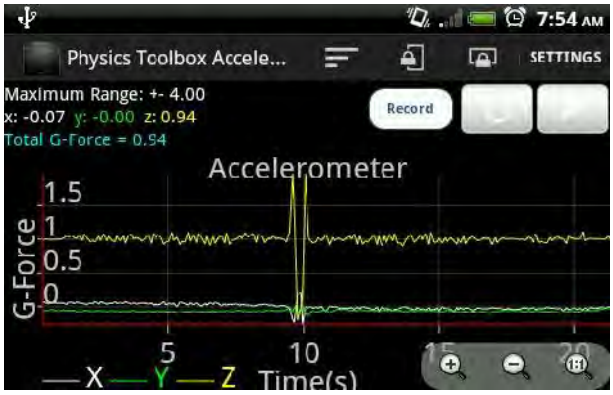


Fig. 3: Accelerometer for Speed Breaker Data

The road abnormalities in this paper are classified into two categories *viz* potholes and speed breaker as given in Table I. Severity is calculated on a scale of 1 to 3. 1 refers a normal severity, 2 to high severity and 3 refers to very high severity.

Table I: Event Categorization

Abnormality	Severity
Pothole	On the scale of 1-3
Speed breaker	On the scale of 1-3

After analyzing the graphs obtained from data collected over various experimental drives it was formulated that we can use z_{th} (threshold value of acceleration along z-axis) to distinguish between smooth road, Pothole and speed breakers. There are some graphs shown through Fig. 1 to Fig. 3 generated for various events categorized in Table I.

Graph for the smooth road is given in Fig. 1, which shows that there is no sudden change in acceleration. While Fig. 2 shows the graph for Pothole. At 14th second it shows a sudden change in downward acceleration, which marks the presence of a pothole. Similarly, Fig. 3 is the graph of a speed breaker. In this graph there is sudden change in aceelaration in upwads direction which signify the presence of a speed breaker.

Threshold acceleration is being set to identify the severity of pothole and speed breakers, which is the maximum permissible acceleration in the vertical direction. Further, analysis shows that if the vehicle moved on a same pothole or a breaker with a high speed and a slow speed the change in acceleration for high speed will me more than that of while the vehicle is slow. Therefoe, two thresholds are being set, one for slow speeds (speed ≤40) and one for high speeds (speed >40).

A. Scenario of Low Speed

If the acceleration is less that 0.5g (4.9 ms^{-2}) then a speed breaker has detected, and if it is greater than 1.5g (14.7 ms^{-2}) then a speed breaker has detected.

To measure the severity classes of threshold are created which are given in the tables below.

Table II: Table for severity level of Potholes

Z_{th}	Severity Level
$0 \le Z_{th} < 0.5g$	1
$-0.5g \le Z_{th} < 0$	2
$-g \le Z_{th} < -0.5g$	3

Table III: Table for severity level of Speed Breakers

Z_{th}	Severity Level
$1.5g \le Z_{th} < 2g$	1
$2g \le Z_{th} < 2.5g$	2
$2.5g \le Z_{th} < 3g$	3

B. Scenario of High Speed

A vehicle travelling at a high speed exhibits severe jerks as compared to the slow speeds over the same abnormality. So different threshold have been set which are shown in the table below.

Table IV: Table for severity level of Potholes

Z_{th}	Severity Level
$-0.5g \le Z_{th} < 0$	1
$-g \le Z_{th} < -0.5g$	2
$-1.5g \le Z_{th} < -g$	3

Table V: Table for severity level of Speed Breakers

Z_{th}	Severity Level
$2g \le Z_{th} < 2.5g$	1
$2.5g \le Z_{th} < 3g$	2
$3g \le Z_{th} < 3.5g$	3

Table 5: Table for severity level of speed breakers

Once the event has detected along with its severity, current location co-ordinates are being fetched from the built-in GPS system of android device, it is stored in the local database specially designed for this application which synchronizes with the global database after the finish of journey.

IV. RESULTS AND DISCUSSION

To evaluate the described algorithm the authors have gone through the following steps.

i. Counting the potholes manually and recording their locations using Walking GPS approach [12]
ii. Test drive over a 4 km road
iii. Examining the database for the correctness

The selected road is of 4 km over which the test was performed. An android phone running an application made for this same purpose was kept in a car on the flat horizontal surface, and the car was made to run on the road with slow speeds. After the finish of the journey the number of potholes and breakers detected during the journey were counted by looking into the database. The following table shows the potholes and speed breakers recorded by the application.

Table VI: Number of Potholes and speed breaker as recorded by the application

Severity Level	Pothole	Speed Breaker
1	20	1
2	8	1
3	2	0

The number of potholes and speed breakers counted by manual inspection is shown in the following table.

Table VII: Number of Potholes and speed breaker as recorded by the application

Severity Level	Pothole	Speed Breaker
1	18	1
2	8	1
3	2	0

Table VI and Table VII shows that the proposed automated approach obtained better results as compared to manual inspection. The correctness of the proposed approach comes out to be 93.75%.

The proposed approached has been evaluated with the certain constraints as given below:

1) Proposed methodology has been tested on an android phone keeping the phone in a four wheeler.
2) Phone is kept on a flat horizontal surface, virtual orientation would be considered as an extra feature.
3) Proposed method uses real time accelerometer sensor data for the processing and detection of events. Recording the accelerometer data along with the time stamps in a file for further processing is classified as an extra feature.
4) Data synchronization is done using phone's GPRS/3G service, implementing it on the VANET [13][14] using phone's Wi-Fi, would be considered as an extra feature.

V. CONCLUSIONS

This paper discusses about the importance of road surface monitoring in terms of comfort and safety needed by the road travelers (humans). A vibration based approach has been presented which automatically detects the pothole and speed breakers with their severity levels. This approach can be very beneficially for secure travelling especially in unknown road conditions. Proposed approach can be easily deployed on any android based smart phone. Results of the paper indicate the acceptance of the proposed approach. In future, the algorithm can be tested over different road conditions for more complex scenarios.

REFERENCES

[1] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Maden, and H. Balakrishnan, "The pothole patrol: using a mobile sensor network for road surface monitoring," in *Proc. of the 6th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys'08, pp. 29–39, 2008.

[2] Mohammad R. Jahanshahi, Farrokh Jazizadeh, Sami F. Masri, Burcin Becerik-Gerber, "A Novel System for Road Surface Monitoring using an Inexpensive Infrared Laser Sensor", *Proc. SPIE 8345*, Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems, 2012.

[3] "Real-time maps and traffic information based on the wisdom of the crowd." [Online] Available: http://world.waze.com/

[4] "Android OS"[Online].Available: http://android.com

[5] G. Strazdins, A. Mednis, G. Kanonirs, R. Zviedris, and L. Selavo, "Towards Vehicular Sensor Networks with Android Smartphones for Road Surface Monitoring." In *CONET'11, CPSWeek'11, 2011.*

[6] "Gps Status| Android Developers" [Online]. Available:http://developer.android .com/reference/android/location/GpsStatus.html

[7] C. Kong, Y. Zhang, X. L. Mingming, "Data Mining System for Road Pothole Detection with On-board Sensors" Bulletin of advanced technology research, Vol. 5, No. 12, 2011.

[8] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: using Mobile Smartphones for Rich Monitoring of Road and Traffic Conditions," in *Proc. of the 6th ACM conference on Embedded Network Sensor Systems*, ser. SenSys '08., pp. 375-358, 2008.

[9] "Trafficsense: Rich monitoring of road and traffic conditions using mobile smartphones," Microsoft Research, Tech, Rep. MSR-TR-2008-59, April 2008.[Online].Available:http://research.microsoft.com/pubs/70573/tr-2008-59

[10] A. Mednis, G. Strazdins, R. Zviedris, G. Kanonirs, L. Selavo, "Real Time Pothole Detection using Android Smartphones with Accelerometers", IEEE, DOI: 10.1109/DCOSS.2011.5982206, 2011.

[11] H. Hautakangas and Nieminen, "Data mining for pothole detection," Presented at the Pro gradu seminar, University of Jyväskylä, February 2011.

[12] R. Stoleru, T. He, and J. A. Stankovic, "Walking gps: A practical solution for localization in manually deployed wireless sensor networks," in *Proc. of the 29th Annual IEEE International Conference on Local Computer Networks*, ser. LCN '04. IEEE Computer Society, pp. 480-489, 2004.

[13] W. Liang, "Vehicular Ad Hoc Networks: Architectures, Research Issues, Methodologies, Challenges, and Trends," IJDSN, pp. 2015.

[14] M. Sivasakthi and S. Suresh, "Research on vehicular ad hoc networks (VANETs): an overview," Journal of Applied Sciences and Engineering Research, vol. 2, no. 1, pp. 23–27, 2013.

The Influence of the CAPTCHA Types to Its Solving Times

D. Brodić*, S. Petrovska *, M. Jevtić* and Z. N. Milivojević**
* University of Belgrade, Technical Faculty in Bor, V.J. 12, 19210 Bor, Serbia
* College of Applied Technical Sciences, Aleksandra Medvedeva 20, 18000 Niš, Serbia
dbrodic@tf.bor.ac.rs

Abstract - In this paper, we present a survey of different CAPTCHA types. Then we address the problem of the CAPTCHA userability. In the experiment, the Internet user response to solve different types of CAPTCHA is tested. The obtained results are given and statistically processed. They are discussed leading to the conclusion different types of CAPTCHA usability and suitability for Internet users.

I. INTRODUCTION

CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) is a program created in order to differentiate between humans and bots during the logging to the website [1]. In this sense, the bot is a software robot, which tries to emulate human users. Hence, it includes elements of artificial intelligence as well as ability of automated reasoning. In fact, CAPTCHA represents a test program which gives a task to be solved. If the correct answer is obtained, then the program classifies the user as a human. The aim of the program is to stop the attacks of bots. Today's research on CAPTCHA is focusing on the development of the test program, which will be easily solved by people and represents a heavy problem to bots. The reasons for using CAPTCHA are as follows [2]:

- Prevention of spams on forums.
- Prevention to open a large number of orders by users on sites that offer free services like Gmail, etc.
- User accounts protection from attacks through which bots are discovering user passwords.
- Validity of online surveys by determining whether the humans or bots answering the questionnaire.

To develop a CAPTCHA that incorporates a high level of security, it has to meet the following requirements [3]:

- The solution must not be conditional, which depends on the user's language and age. It means that it should be intuitive.
- It has to be hard to solve the CAPTCHA test except for the humans in order to differentiate humans from bots.
- It has to be created to not disturb the user privacy. Hence, it has to be not user related.

First CAPTCHA was designed by Broder's team in 1997 for Altavista, to prevent automatic adding URL to a database of a web browser [4]. CAPTCHA may be based on the [5]: (i) Image elements, (ii) Text elements, and (iii) Audio and video elements.

Text-based CAPTCHA is the most common form. This type of CAPTCHA asks the user to decrypt the text which is usually distorted in some way [6]. Unfortunately, this type of CAPTCHA can be successfully attacked by bot due to the existence of good decoders. Figure 1 shows the example of the text-based CAPTCHA.

Image-based CAPTCHA is usually considered as the most advanced and safest one. This type of CAPTCHA requires users to find a desired image between the list of images. Because it is based on image details, it represents an almost impossible task to be solved by bots. Figure 2 shows an example of the image-based CAPTCHA.

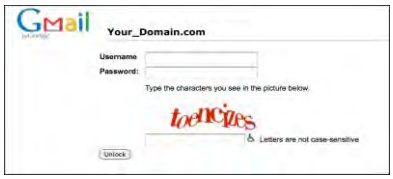


Figure 1. An example of text-based CAPTCHA

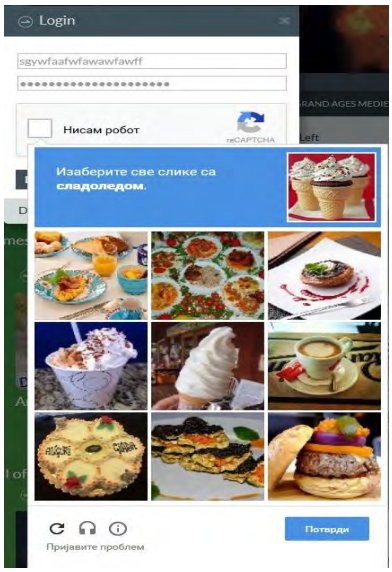


Figure 2. An example of image-based CAPTCHA



Figure 3. An example of FaceDCAPTCHA

Although, CAPTCHA protects user accounts and passwords, it often represents a firm obstacle not just the bots, but to the humans too. As an extension to image-based CAPTCHA the FaceDCAPTCHA is used [7]. It is a CAPTCHA that incorporates the elements of a face detection. It is one of the newer CAPTCHA types that includes a high level of security. It exploits a research about the human brain, which is very effective in the process of natural face segmentation in spite of used complex backgrounds. Figure 3 shows an example of the FaceDCAPTCHA.

Video and audio-based CAPTCHA refers to the auditory reproducible characters that the user have to input. Although, this type of CAPTCHA is typically attacked in approximately 70% of cases, its development and innovation is essential for the blind users [8].

All articles about CAPTCHA have researched the safety and security standpoint ignoring the difficulties of users to solve its task. In this paper, we explore the complexity of CAPTCHA tasks from the user's viewpoint. Hence, we conducted the experiment based on four different types of CAPTCHA, which are tested in the community of 100 Internet users differentiated by the level of Internet use and gender. The aim of our research was to identify the following:

- The user's response rate of solving different types of CAPTCHA.
- The suitability of different types of CAPTCHA to the certain group of Internet users.

As the final results, we suggest the use of certain types of CAPTCHA and their implementation in different software environment.

The paper is organized as follows. Section 2 describes the elements of the experiment. Section 3 presents the results and discusses them. Section 4 gives the conclusions.

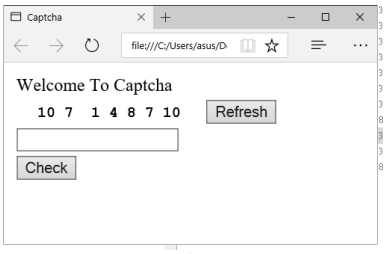
II. EXPERIMENT

The experiment is performed on four different CAPTCHAs. Figure 4 illustrates these four CAPTCHAs.

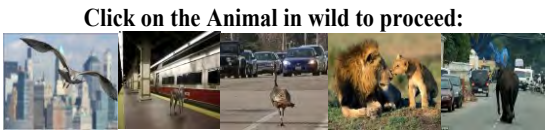
The aim was to solve these types of CAPTCHAs and to measure the user's response rate (in seconds) to successfully solve the task of each CAPTCHA. In this



(a)



(b)



(c)



(d)

Figure 4. Four different CAPTCHAs: (a) Text based CAPTCHA, (b) Text-number based CAPTCHA, (c) Image-based CAPTCHA – Animals in Wild, (d) Image-based CAPTCHA – Picture of CAPTCHA

sense, the response rate of each user for each of four CAPTCHAs represents the dependent variable. It is given depending on the Internet years of use as well as the gender. The experiment is carried out on the sample of 100 Internet users. The Internet users are classified according to: (i) years of Internet use (from 1 to 9 years) and (ii) gender (male of female). The population has 50 female and 50 male users.

Furthermore, the tested population is chosen to have a Gaussian distribution. Figure 5 shows the distribution of testing population, according to the years of Internet use.

Three hypotheses are in the focus of our experiment:

- Hypotheses 1: Subjects who use the Internet more time will have a shorter response rate.
- Hypotheses 2: Faster response rate is presumable for image-based compared to text-based CAPTCHA.
- Hypothesis 3: Does the gender have any influence on solving the CAPTCHA?

III. RESULTS AND DISCUSSION

First of all, it is important to know that all users successfully solve all four CAPTCHAs. However, the

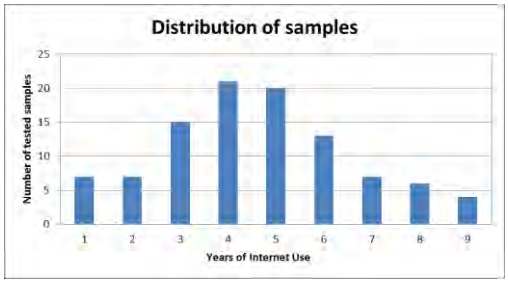


Figure 5. The distribution of testing population, according to the years of Internet use

time of solving CAPTCHA is quite different between them. Figures 6-7 show the results representing the time to solve the CAPTCHA in response to the years of Internet use or gender, respectively.

The measures that characterize each CAPTCHA are: (i) Minimum time to solve a CAPTCHA, (ii) Maximum time to solve a CAPTCHA, (iii) Mean time to solve a CAPTHA, (iv) Standard deviation, and (v) Correlation coefficient *R*. The statistical measures like standard deviation, variance and correlation coefficient are defined respectively as follows [9]:

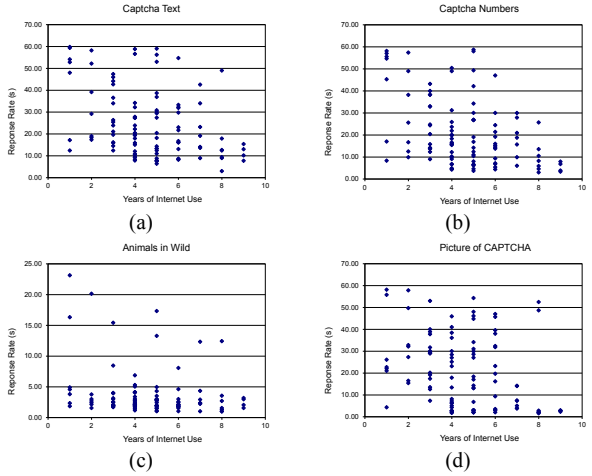


Figure 6. The distribution of the CAPTCHA solving times in accordance to the years of Internet use

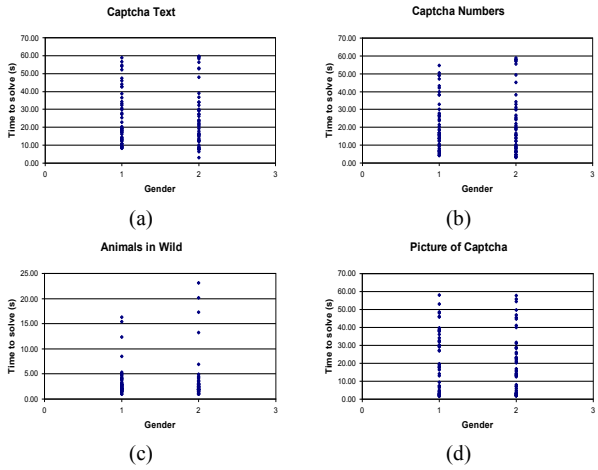


Figure 7. The distribution of the CAPTCHA solving times in accordance to the gender

$$\sigma = SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}\ , \tag{1}$$

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}\ . \tag{2}$$

The correlation coefficient *R* expresses the strength and direction of a linear relationship between two variables *x* and *y*, where x_i is the value of the variable *x* for the instance *i*, $\bar{x}$ is the mean of *x* value, y_i the value of the variable *y* for the instance *i*, and $\bar{y}$ is the mean of *y* value [9]. Consequently, *R* measures the strength and direction of a linear relationship between two variables on a scatterplot [10]. It can receive the value from -1 to +1. The larger the absolute value of the coefficient, the stronger the relationship between the variables. An absolute value of 1 indicates a perfect linear relationship. A correlation close to 0 indicates no linear relationship between the variables. If *R* is positive, then the two variables tend to increase or decrease together. In contrast, if *R* is negative, then one variable increases as the other decreases.

The experimental result of these measures is given in Tables I-II. Figure 8 shows the level of the correlation coefficient *R* in accordance to the years of Internet use or gender of each examinee.

TABLE I. SOLVING CAPTCHA VS. INTERNET USE

Measures	Internet use	Text	Number	Animals in Wild	Picture of CAPTCHA
Population	100	100	100	100	100
Minimum	1	3.00 s	3.00 s	1.00 s	1.56 s
Maximum	9	59.78 s	58.71 s	23.14 s	58.14 s
Mean	-	24.71 s	21.40 s	3.61 s	20.78 s
SD	-	15.56	15.68	3.94	16.63
R(Internet use)	-	-0.42	-0.45	-0.27	-0.34

TABLE II. SOLVING CAPTCHA VS. GENDER (1-MALE, 2-FEMALE)

Measures	Gender	Text	Number	Animals in Wild	Picture of CAPTCHA
Population	100	100	100	100	100
Minimum	1-male	8.28 s	4.30 s	1.00 s	1.56 s
Maximum	1-male	58.80 s	54.67 s	16.32 s	58.14 s
Minimum	2-female	3.00 s	3.00 s	1.00 s	1.99 s
Maximum	2-female	59.78 s	58.71 s	23.14 s	57.81 s
Mean	1-male	25.09 s	21.28 s	3.48 s	21.85 s
Mean	2-female	25.01 s	22.20 s	3.83 s	19.74 s
R(Gender)	-	0.00	0.03	0.04	-0.01

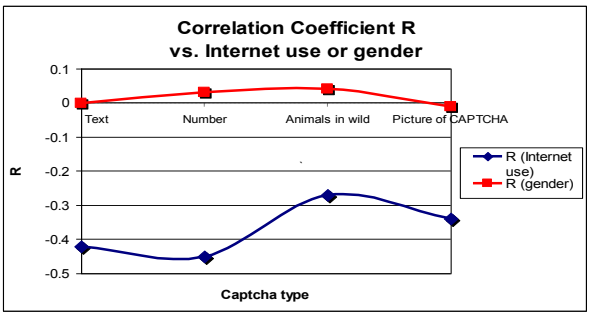


Figure 8. Correlation coefficient *R* for all four CAPTCHAs in accordance to: (a) years of Internet use, (b) gender

In our case, the *R* is negative for all three CAPTCHA. It means that if the users have more years of Internet experience, then CAPTCHA will be solved in shorter time. However, the *R* is -0.42 for text based CAPTCHA and -0.45 for number based CAPTCHA. It can be qualified as a moderate downhill or negative relationship. On the contrary, the image based CAPTCHA has *R* equal to -0.27, which represents a weak downhill or negative linear relationship. From the given *R* values, it is obvious that hypotheses 1 is confirmed.

If we take into account aforementioned hypothesis about CAPTCHA that its solution must not be conditional and it should be intuitive, then image based CAPTCHA is better qualified for the right, i.e. ideal CAPTCHA choice.

The obtained time to successfully solve CAPTCHA is as follows: (i) Text based CAPTCHA from 3.00 s to 59.78 s, (ii) Number based CAPTCHA from 3.00 s to 58.71 s, (iii) Image based CAPTCHA (Animals in the wild) from 1.00 s to 23.14 s, and (iv) Image based CAPTCHA (Picture of CAPTCHA) from 1.56 s to 58.14 s. The mean time of solving CAPTCHA brings great advantage of image based CAPTCHA, but only if the CAPTCHA image is carefully chosen. It means that image in CAPTCHA should be clearly defined and chosen. Hence, image based CAPTCHA (Animals in the wild) receives 3.61 s compared to 24.71 s and 21.40 s for text and number based CAPTCHA, respectively. However, badly chosen image CAPTCHA receives 20.78 s. It is obvious that image based CAPTCHA has clear advantages compared to text or number based CAPTCHA, if it is chosen adequately. In this way, hypotheses 2 were also confirmed. Furthermore, it is clear that solving time of CAPTCHA in accordance to gender choice are quite similar with the margin inside 10%. Hence, hypotheses 3 were not confirmed, because there is no evident difference between solving time of a male or female. It is also confirmed with correlation coefficient *R*, which receive the values between -0.01 and 0.04.

In some application, the time to solve a CAPTCHA can represent a critical value. Hence, the quick solving an easily differentiate humans from bots. Also, image-based CAPTCHA is more universal, because it is more appropriate to different types of electronic devices like computer, tablet or smartphone. From all aforementioned, the results of the experiment and comparison of four types of CAPTCHA give the clear winner in the competition of three different CAPTCHA. Image based CAPTCHA has a clear advantage in the time of its successfully solving. Furthermore, it is more intuitive and less conditional to be solved compared to other two tested CAPTCHA.

IV. CONCLUSION

The paper described the research conducted to evaluate different CAPTCHA according to the experience of Internet users (years of Internet use) and gender (male or female). The result of the experiment showed that users more easily solve (in less time) image-based CAPTCHA compared to text or number based CAPTCHA. Furthermore, the value of correlation coefficient *R* (negative value) proved that the years of Internet use can help users to solve CAPTCHA in a less time. However, one of the CAPTCHA postulate is that CAPTCHA should be intuitive. Using this premise, the ideal model of CAPTCHA can be equally easily solved by experience and inexperienced user. Because the correlation coefficient *R* of the text or number based CAPTCHA is approximately 50% higher than in the image based CAPTCHA, it proved that the image-based CAPTCHA is much less dependent of the user experience with the Internet. Hence, image based CAPTCHA better satisfied the aforementioned premise. The study also proved that there is small difference in CAPTCHA's solving time between male or female users.

Further research will be toward inclusion of higher number of different CAPTCHA's samples in the experiment as well as the exploration of education level and age of users as a parameter, too.

REFERENCES

[1] L. von Ahn, M. Blum, J. Langford, "Telling Humans and Computers Apart Automatically", Communications of the ACM, vol. 47, no. 2, pp. 56–60, 2004.

[2] www.google.com,http://en.wikipedia.org/wiki/CA PTCHA

[3] L. von Ahn, M. Blum, N. Hopper, and J. Langford, "CAPTCHA: Using hard AI problems for security", Proceedings of Eurocrypt, Warsaw, Poland, May 4-8, 2003, pp. 294-311

[4] M. Lillibridge, M. Abadi, K. Bharat, A. Broder, "Method for Selectively Restricting Access to Computer Systems", United States Patent 6195698. Applied 1998 and Approved 2001.

[5] K. Kukade1, M.S. Deshmukh, "CAPTCHA Problems on AI Based For Protection from Automated Hacking Tool", International Journal of Computer Science and Mobile Computing, vol.4, no. 5, 2015, pp. 714-719.

[6] X. Ling-Zi and Z. Yi-Chun, "A Case Study of Text-Based CAPTCHA Attacks", Proceedings of International Conference on Cyber Enabled Distributed Computing and Knowledge Discover, Sanya, China, 2012, pp. 121-124.

[7] G. Goswami, B. M. Powell, M. Vatsa, R.a Singh, A. Noore, "FaceDCAPTCHA: Face detection based color image CAPTCHA", Future Generation Computer Systems, vol. 31, no. 2, 2014, pp. 59-68.

[8] H. Gao, H. Liu, Dan Yao, X. Liu, "An audio CAPTCHA to distinguish humans from computers," Proceedings of IEEE Symposium on Security Privacy, Oakland, U.S.A., May 16-19, 2010, pp. 399–413.

[9] J. Higgins, The Radical Statistician: A Beginners Guide to Unleashing the Power of Applied Statistics in The Real World (5th Ed.), Jim Higgins Publishing, 2006.

[10] Chi Yau, R Tutorial with Bayesian Statistics Using OpenBUGS, Kindle Book, 2015.

Techniques and Applications of Emotion Recognition in Speech

S. Lugović*, I. Dunđer** and M. Horvat*
* Zagreb University of Applied Sciences, Department of Computer Science and Information Technology, Vrbik 8, Zagreb, Croatia
** University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information and Communication Sciences, Ivana Lučića 3, Zagreb, Croatia
slugovic@tvz.hr, idundjer@ffzg.hr, marko.horvat@tvz.hr

Abstract - Affective computing opens a new area of research in computer science with the aim to improve the way how humans and machines interact. Recognition of human emotions by machines is becoming a significant focus in recent research in different disciplines related to information sciences and Human-Computer Interaction (HCI). In particular, emotion recognition in human speech is important, as it is the primary communication tool of humans. This paper gives a brief overview of the current state of the research in this area with the aim to underline different techniques that are being used for detecting emotional states in vocal expressions. Furthermore, approaches for extracting speech features from speech datasets and machine learning methods with special emphasis on classifiers are analysed. In addition to the mentioned techniques, this paper also gives an outline of the areas where emotion recognition could be utilised such as healthcare, psychology, cognitive sciences and marketing.

Keywords - emotion recognition, speech analysis, machine learning, acoustic signal processing, linguistic speech features, affective computing, human-computer interaction

I. INTRODUCTION

Humans verbally communicate by speech and language. This enables faster sharing of messages, conveying of ideas and spreading of inventions. Communication between humans is actually not just what humans say, but also how they say it. Furthermore, facial expressions, as a part of non-verbal communication, are responsible for about 55%, voice intonation for about 38% and actual words for 7% of the message perception [1].

According to [2], "human beings are entities born into a languaging community. Language acts to fundamentally sever the human species from the rest of existence. Our human nature is not therefore a 'natural nature'. Rather it is a humanly constructed nature which is generated through languaging conventions. It places us radically and finally across a distinguishing boundary from all else that is not 'human'".

Simply writing "go home", without producing any accompanying voice intonation or facial expression, could be understood as an imperative statement of very positive or very negative polarity: e.g. if a child is feeling sick and the teacher sends him home; or a child that is not behaving well in the classroom and the teacher sends him home. With facial expressions and intonations humans are communicating emotions, which could be related to judgments and subjective perception. Emotions influence judgment by how one feels about the object that affects judgment [3].

The rational model of emotion theory provides a description of the process how humans generate emotions, and its base is the complex reaction that involves both mind and body [4].

According to [5], an emotion can be defined as "a mental state that arises spontaneously from external stimuli that is often accompanied with physical expressions, which is heavily influenced by culture to serve a purpose of a particular situation". So how humans perceive emotions will affect how the make judgments in particular situations.

In some circumstances, such as professional driving or flying, people could utilise computers to facilitate emotion recognition, in order to enhance the perception process and to use such information to help make better judgments and to boost the decision making process. So far, results arising from machine emotion recognition showed better performance when compared to human emotion recognition.

Reference [6] provided results that researchers applying a fuzzy rule based system for emotion recognition from natural speech and a speaker-independent mode achieved emotion recognition rates ranging from 55% to 95%, while humans could reach emotion recognition rates close to 60% from unknown speakers.

A research [7] stated that rates are up to 81% when it comes to automatic emotion detection and recognising which of eight emotions an actress expresses through four physiological channels. They also developed an online version of the algorithm, which scored 8% less than the offline version. Such a level of physiological recognition is comparable to facial and vocal expressions [8].

Machines could help people in making right decisions by recognising emotions, especially in irrational situations where decisions have to be made faster than a rational-performing mind. Sometimes it is valuable to artificially influence mental and emotional states to get a better

individual performance in stress-related occupations [9] and prevent mental disorders from happening [10]. Recent research has shown that under certain circumstances multimodal emotion recognition is feasible even in real-time [11].

Directions to explore emotions more deeply were for the first time proposed more than 60 years ago, by [12] who stated "it is interesting to know that the sort of phenomenon which is recorded subjectively as emotion may not be merely a useless epiphenomenon of nervous action, but may control some essential stage in learning, and in other similar processes. I definitely do not say that it does, but I do say that those psychologists, who draw sharp and uncrossable distinctions between man's emotions and those of other living organisms and the responses of the modern type of automatic mechanisms, should be just as careful in their denials as I should be in my assertions".

Probably on the echo of such a call, [13] developed a concept of "ways to think", as a response to different emotional states. "Each of our major 'emotional states' results from turning certain resources on, while turning certain others off – thus changing the way one's brain behaves". The author also proposed that the peoples' mind is organised around the following "ways to think":

- instinctive reactions as the lowest level,
- learned reactions,
- deliberative thinking,
- reflective thinking,
- self-reflective thinking,
- self-conscious emotions as the highest level.

Every emotional state triggers one "way to think" and switches off another state.

II. Brief Overview of Affective Computing

According to [8] "affective computing and Human Computer Interaction (HCI) research target four broad areas: a) reducing user frustration, b) comfortable communication of user emotion, c) infrastructure and applications to process affective information and d) building tools to support development of socio emotional skills". Without information about emotions, "it is difficult to achieve a harmonic and natural man-machine interface for applications such as patient care, geriatric nursing, call centres, psychological consultation, and human communication" [14].

In [15], the author is defining affective computing as "measuring observations of the motor system behaviour that correspond with high probability to an underlying emotion or combination of emotions". So, one could say that by analysing expressions of emotions it is possible to identify emotional states of the subject that is being observed. To find a particular state, the subject has to be exposed to a specific experience. It is also possible to influence the user experience by a machine [9]. Such a generic process in the framework of affective computing is presented in Fig. 1.

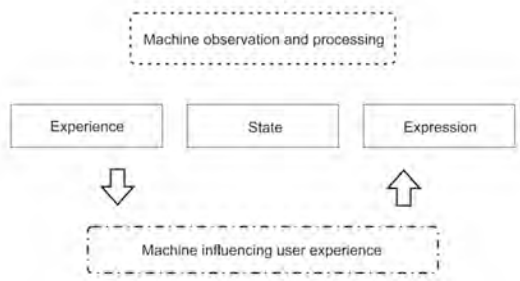


Figure 1. The framework of affective computing.

The experience will result in a new state and such a new state could be measured through the expression of the subject. This process could be observed and at the same time artificially influenced. So there are present basic cybernetic principles of feedback, communication and control at work. As such a loop is possible, where a machine actually influences a human based on the emotions of an individual, ethical questions arise. Such sort of computing could be called Non-Turing computing [16], which "is based on the simulation of neurons and requires complex models and technical implementations… this technology is not deterministic and does not follow a clearly described programme".

Ethical issues in ICT are well addressed in the Etica (Ethical Issues of Emerging ICT Applications) project funded by the EU[1]. Extensive research [17] provided detailed guidelines for approaching ethical issues. Recommendations for policy makers are: provide a regulatory framework to support Ethical Impact Assessment; set up an ICT ethics observatory and set up a forum for stakeholders. Recommendations for the industry are: incorporate ethics into ICT research and development, and facilitate ethical reflexivity in ICT projects and practice. It is important for everybody involved in research and development of implementation of human cognitive functions in machines to get involved in ethical discussions.

In order to design machines to response to human emotional states, one has to recognise user emotional states, and to do so different affect recognition options can be applied [18], *including the use of psychophysiological measures (e.g. heart rate); diagnostic tasks; self-reports; facial expressions; and knowledge-based methods to derive likely affective state based on factors from current task context (e.g. type, complexity, time of day, length of task), personality (extraversion, aggressiveness, obsessiveness, etc.) and individual history (past failures and successes, affective state associated with current task, etc.)* [19].

To achieve better results one could integrate the aforementioned methods and such multimodal system could depend on the following factors: the types of emotions to be detected, the instrumentation necessary to collect the required data, the real-time requirements and data availability [18].

[1] http://www.etica-project.eu

If user-machine interface is examined as one unified system [20], then the impact of this new artefact on the society in a wider sense must also be considered. Such shift in research may lead to development of new socio-technical theories and models that could be implemented in the economy and management, health care systems and social sciences.

III. EMOTION RECOGNITION IN SPEECH

Emotion recognition in speech can be divided into two main aspects of research. The first, where a machine is designed to artificially produce emotional sounds, while the second aspect deals with a machine that recognises emotional states of a user. In order to succeed, a machine has to learn about human emotions out of speech, using different classifiers that discern from speech-based data from which features are extracted.

An abstract conceptual architecture scheme of a generic system that is capable of recognising emotions in speech could be composed of a number of processes and elements as presented in Fig. 2. It provides an overview of the information system functions and processes that are necessary to implement the recognition of emotions in speech. The following paragraphs present some results from published experiments with related techniques used to recognise emotions in speech.

The first step in getting machines to recognise emotions is to gather data from which a machine will learn. This could be done by using primary and secondary inputs. The primary input could be recordings of actors which express different emotions by reading the same text, while the secondary input could be using already existing databases, which were developed by other researchers. Some hybrid approaches are also possible.

In some situations, the best input is the one from real scenarios – e.g. in one study a baby voice was analysed [21]. Another one is using input learning sets from movie databases [22], using a machine to create voices [23], or collecting speech samples from various sources on the internet and then evaluating emotions by humans [24]. A very extensive overview of databases consisting of emotions in speech can be found in [25]. Some of the most widely used databases are The Danish Emotional Speech Database (DES) [26], The Berlin Emotional Speech Database (BES) [27], The Speech Under Simulated and Actual Stress (SUSAS) Database [28], and The Emotional Prosody Speech and Transcripts acted Database [29].

The next step would be to extract features from the collected or acquired data. Features can be acoustic or linguistic [30].

Acoustic features are intonation, intensity, TF-transformation, linear prediction, cepstral coefficients, formants, spectrum, harmonicity and perturbation. Linguistic feature can be divided into three subcategories: a) linguistics such as phonemes or words, b) paralinguistics such as laughter or sighs, and c) disfluencies such as pauses.

Literature, as in [31], defines linguistic information as symbolic information that is represented by a set of discrete symbols and rules of their combination. Paralinguistic features are related to how a sentence is uttered in different ways according to intentions, attitudes and speaking styles of the speaker and related to the speaker's emotional state. Non-linguistic factors are related to age, gender, physical and emotional states of the speaker and are not directly related to the linguistic and paralinguistic content.

Voice features can also be successfully analysed by scientific computer software. One of the most widely used freely available software packages in phonetics for speech analysis is Praat from University of Amsterdam[2].

Some of useful supported features are: generation of waveforms, wide and narrow band spectrograms and pitch tracks; speech signal filtering and processing; enhancement and manipulation of specific frequency regions; segmentation and labelling of words, syllables and individual phonemes; display of intensity contours; extraction of individual sounds; and measurement of voice onset time (VOT).

After the extraction of features, the machine learning part of the generic system should analyse features and find statistical relationships between particular features and emotional states. This part of the system is also called "classifier".

Most commonly used classification algorithms are Artificial Neural Networks (ANN) [32], k-Nearest Neighbour (k-NN) and Support Vector Machines (SVM) [33], decision trees [34]. Furthermore, probabilistic models such as the Gaussian mixture model (GMM) [35] (or stochastic models such as Hidden Markov Model (HMM) [36] can be applied.

Some researchers use different classifiers to define different emotions [37]. An interesting approach was taken by [38], where a probabilistic semantic classifier was developed, which tried to emulate the way humans perform classification tasks instead of applying the standard mathematical, i.e. statistical approach.

When the machine learning subsystem becomes able to statistically recognise emotions from features, the system could be provided with real speech input from which features are extracted and compared with existing preferences and settings. Results could then be evaluated again and sent to re-learning and/or extracted for use in another system. Such a system could be used in different real-life situations. In [39], authors proposed a system that helps to detect abnormal driver behaviour from a dataset of driver speech recordings.

A previously published research [40] analysed automatic anger recognition in speech as an important factor to improve the user satisfaction in call centres. Results in this work showed the possibility of using such a technique as part of a real-life decision support system in call centres.

In [41] a method was proposed for detecting baby voice and using such a recording to recognise emotions.

[2] http://www.fon.hum.uva.nl/praat

Such information is then used to improve baby care support in the baby's home.

Researchers also proposed a method to distinguish deceptive speech from truthful speech [42]. Such a system could have a practical application in law and other government agencies, to evaluate speech reports from informants or analyse suspects while being interviewed.

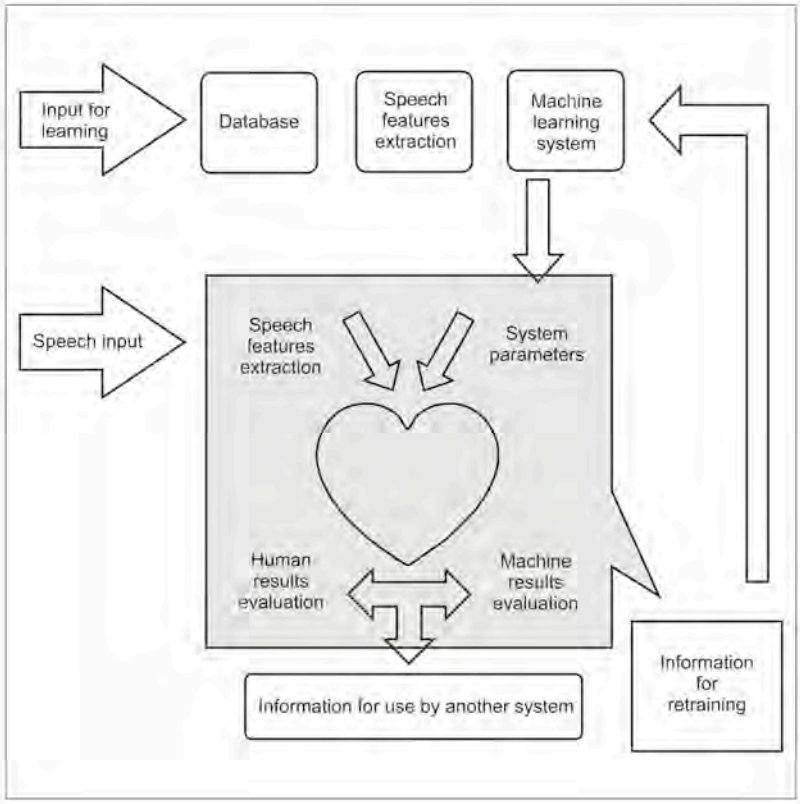


Figure 2. Abstract conceptual architecture scheme of a generic system for automated recognition of emotions in speech.

IV. DISCUSSION

From the authors' view, being able to recognise emotions by machines could be an important step in the advancement of socio-technical systems research. This research program aims to understand the complexity of real situations rather than analysing its separate aspects [43].

Emotions emerge from the agent internal property, changes in the environment and process of interaction with environment and are dependent variable of experience.Observing them could provide deep insights into the complexity of real situations. And when it is done by machines, the output of such analyses is not limited by the number of human agents that are doing it, but by the processing power of computers.

For example, from the perspective of the management, insights into the emotional states of employees are limited by the number of the human agents conducting an emotional analysis of employees. But if one could employ machines to do such an analysis, then there are possibilities to increase the capacity and to reduce the price of conducting it.

Research shows that in corporations exist two parallel cultures, one cognitive and one emotional (thinking vs. feeling). The former is conveyed verbally and later through nonverbal cues such as body language and facial expressions. Emotional culture influences the satisfaction of employees, the burnout, teamwork, and even hard measures such as financial performance reports and absenteeism [44].

The authors of this paper will propose that by analysing emotions in speech we could get insight into both cultures extracting two signals from the same source. One representing what is said and another representing how agents feel in the process of communication.

In particular, by using recognition of emotions in speech, which is not an invasive technology, one could place a small black box in every board room and analyse how managers emotionally react and respond to a particular topic. Such information could even be used for reporting towards stakeholders.

There are two main types of communication, one that creates new information and one that distributes such new information [45]. Information itself can be defined as a difference that makes a difference [46], and by having in place such an emotion detection and measurement system, the authors will argue that there are existing tools needed to tackle the challenges which confront the contemporary civilisation at the moment.

Such techniques could also be applied in observing companies' interactions with customers through call centres. Currently to analyse emotions in such conversations, a human specialist with limited capacities has to be included. But if one employs machines to do the task, then it will be much cheaper and output will be more consistent.

Public services could also benefit from such an approach. Namely, it is possible to analyse emotions in the voices, i.e. speeches, of parliament members. Such information could be of high interest and value for the society, as attitudes and honesty of politicians could be investigated.

Also, emotion recognition could be used in different NGOs dealing with civil society issues. For example, by using emotion recognition in speech it is possible to track emotional states and behaviour of different social groups.

Also, in the academia such techniques could be used in order to achieve higher granularity, especially in social science research. For example, one could detect emotions in speech while conducting interviews. Such an approach could provide another signal that could be analysed from the same data collection procedures.

V. CONCLUSION

Selective review of the literature presented in this paper gives an overview of affective computing and emotions in speech recognition techniques in particular. By presenting existing researches, experiments and techniques used in this domain the authors hope to bring them one step closer to other domains, such as social sciences and humanities. If intelligence is the capacity of understanding and the ability to perceive and comprehend meaning, then implementing emotions in speech recognition could increase the intelligence of socio-technical systems that people are part of. By doing so socio technical system increase it's capacity of perception procedures and as its done by machines, benefit to cost ratio is high. Such implemented techniques could have a high impact on the society, but there are still lots of open questions related to ethics, who pays for the associated costs, how is gathered information shared, who makes decisions based on the gathered data and how are success and the value of the implementation evaluated? The following decades will most certainly provide possibilities to observe progress in this matter.

REFERENCES

[1] C. C. Chibelushi and F. Bourel, "Facial expression recognition: A brief tutorial overview," CVonline: On-Line Compendium of Computer Vision, vol. 9, 2003.

[2] P. Boxer and V. Kenny, "The economy of discourses: a third order cybernetics?," Human Systems Management, vol. 9(4), pp. 205–224, 1990.

[3] G. L. Clore and J. R. Huntsinger, "How emotions inform judgment and regulate thought," Trends in cognitive sciences, vol. 11(9), pp. 393–399, 2007.

[4] R. S. Lazarus and B. N. Lazarus, "Passion and reason: Making sense of our emotions," Oxford University Press, New York, 1994.

[5] N. Kamaruddin and A. Wahab, "Heterogeneous driver behavior state recognition using speech signal," in Proceedings of The 10th WSEAS International Conference on System Science and Simulation in Engineering, pp. 207–212, 2011.

[6] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Fuzzy emotion recognition in natural speech dialogue," IEEE International Workshop on Robot and Human Interactive Communication 2005, ROMAN 2005, IEEE, pp. 317–322, 2005.

[7] E. Vyzas and R. W. Picard, "Offline and Online Recognition of Emotion Expression from Physiological Data," in Proceedings of the Workshop on Emotion-Based Agent Architectures at the Third International Conference on Autonomous Agents, pp. 135–142, 1999.

[8] R. W. Picard, "Affective Computing for HCI," In HCI (1), pp. 829–833, 1999.

[9] K. Ćosić, S. Popović, M. Horvat, D. Kukolja, B. Dropuljić, B. Kovač, and M. Jakovljević, "Computer-aided psychotherapy based on multimodal elicitation, estimation and regulation of emotion," Psychiatr. Danub., vol. 25, pp. 340–346, 2013.

[10] K. Ćosić, S. Popović, M. Horvat, D. Kukolja, B. Dropuljić, I. Kostović, M. Judaš, M. Radoš, M. Radoš, L. Vasung, B. B. Spajić, S. Doričić, and D. Mesić, "Virtual reality adaptive stimulation in stress resistance training," In HFM-205 Symposium on "Mental Health and Well-Being across the Military Spectrum", 2011.

[11] D. Kukolja, S. Popović, M. Horvat, B. Kovač, and K. Ćosić, "Comparative analysis of emotion estimation methods based on physiological measurements for real-time applications," International journal of human-computer studies, vol. 72(10), pp. 717–727, 2014.

[12] N. Wiener, "The Human Use of Human Beings," Houghton Mifflin, 1954.

[13] M. Minsky, "The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind," Simon and Schuster, 2007.

[14] F. Ren, "From cloud computing to language engineering, affective computing and advanced intelligence," International Journal of Advanced Intelligence, vol. 2(1), pp. 1–14, 2010.

[15] R. W. Picard, "Affective Computing," M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 321, 1995.

[16] B. C. Stahl, "Virtual suicide and other ethical issues of emerging information technologies," Futures 50, pp. 35–43, 2013.

[17] Pathways towards Responsible ICT Innovation Policy brief of STOA on the ETICA project, http://www.etica-project.eu/deliverable-files, retrieved on 21.03.2014.

[18] E. Hudlicka, "To feel or not to feel: The role of affect in human–computer interaction," International Journal of Human-Computer Studies, vol. 59(1), pp. 1–32, 2003.

[19] E. Hudlicka, "Increaing Sia Architecture Realism by Modeling and Adapting to Affect and Personality," In Socially Intelligent Agents, Springer, pp. 53–60, 2002.

[20] E. Hollnagel, "From human factors to cognitive systems engineering: human machine interactions in the 21st century," In Anzen-no-Tankyu (Researches on Safety), M. Kitamura, I. Kimura (Eds.), ERC Publishing, Tokyo, 2001.

[21] S. Yamamoto, Y. Yoshitomi, M. Tabuse, K. Kushida, and T. Asada, "Detection of baby voice and its application using speech recognition system and fundamental frequency analysis," in Proceedings of 10th WSEAS International Conference on Applied Computer Science, pp. 341–345, 2010.

[22] R. Sharma, U. Neumann, and C. Kim, "Emotion Recognition In Spontaneous Emotional Utterances From Movie Sequences," in Proceedings of the WSEAS International Conference on Electronics, Control & Signal Processing, 2002.

[23] M. Kurematsu, H, Jun, and F. Hamido, "An extraction of emotion in human speech using speech synthesize and classifiers for each emotion," in Proceedings of the WSEAS Transaction on

Information Science and Applications, vol. 5(3), pp. 246–251, 2008.
[24] B. Dropuljić, S. Popović, D. Petrinović, and K. Ćosić, "Estimation of emotional states enhanced by a priori knowledge," in Proceedings of the 4th International Conference on Cognitive Infocommunications (CogInfoCom), IEEE, pp. 481–486, 2013.
[25] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J. C. Martin, L. Devillers, S. Abrilian, A. Batliner, and N. Amir, "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," Affective computing and intelligent interaction, Springer Berlin Heidelberg, pp. 488–500, 2007.
[26] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a danish emotional speech database," Eurospeech, 1997.
[27] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," Interspeech, vol. 5, pp. 1517–1520, 2005.
[28] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting started with SUSAS: a speech under simulated and actual stress database," EUROSPEECH, vol. 97(4), 1997.
[29] J, Hirschberg, J. Liscombe, and J. Venditti, "Experiments in emotional speech," in Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.
[30] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," Speech Communication, vol. 53(9), pp. 1062–1087, 2011.
[31] J. Kleckova and B. Mahdian, "Nonverbal Communication in Spontaneous Speech Recognition", WSEAS Transactions on Electronics, vol. 1(3), pp. 531–536, 2004.
[32] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in Proceedings of Artificial Neural Networks in Engineering, 1999.
[33] M. Khan, T. Goskula, M. Nasiruddin, and R. Quazi, "Comparison between k-nn and svm method for speech emotion recognition," International Journal on Computer Science & Engineering, vol. 3(2), 2011.
[34] C.-C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," Speech Communication, vol. 53(9), pp. 1162–1171, 2011.
[35] X. Cheng and Q. Duan, "Speech Emotion Recognition Using Gaussian Mixture Model," Proceedings of The 2nd International Conference on Computer Application and System Modeling,. 2012.
[36] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech communication, vol. 41(4), pp. 603–623, 2003.
[37] J. Hakura, M. Kurematsu, and H. Fujita, "An exploration toward emotion estimation from facial expressions for systems with quasi-personality," International Journal of Circuits, Systems and Signal Processing, vol 1(2), pp. 137–144, 2008.
[38] M. Stanojević, N. Tomašević, and S. Vraneš, "Probabilistic semantic classifier," in Proceedings of the 5th International Symposium on Data Mining and Intelligent Information Processing, pp. 260–265, 2009.
[39] N. Kamaruddin and A. Wahab, "Heterogeneous driver behavior state recognition using speech signal," in Proceedings of The 10th WSEAS International Conference on System Science and Simulation in Engineering, pp. 207–212, 2011.
[40] F. M. Lee, F. M. Li, and R. Y. Huang, "Recognizing low/high anger in speech for call centers," in Proceedings of the International Conference on Signal Processing, Robotics and Automation, World Scientific and Engineering Academy and Society (WSEAS), pp. 171–176, 2008.
[41] S. Yamamoto, Y. Yoshitomi, M. Tabuse, K. Kushida, and T. Asada, "Detection of baby voice and its application using speech recognition system and fundamental frequency analysis," in Proceedings of 10th WSEAS International Conferenc on Applied Computer Science, 2010.
[42] M. Sanaullah and K. Gopalan, "Distinguishing Deceptive Speech from Truthful Speech using MFCC," in Proceedings of the 7th International Conference on Circuits, Systems and Signals, WSEAS, pp. 167–171, 2013.
[43] G. Ropohl, "Philosophy of socio-technical systems. Techné: Research in Philosophy and Technology," vol. 4(3), pp. 186–194, 1999.
[44] S. Barsade and O. A. O'Neill, "Manage your Emotional culture," Harvard Business Review, vol. 94(1–2), pp. 58–66, 2016.
[45] V. Flusser, "Kommunikologie," Frankfurt am Main, Fischer, 1996.
[46] G. Bateson, "Steps to an ecology of mind. Collected essays in anthropology, psychiatry, evolution, and epistemology," University of Chicago Press, 1972.

Word Occurrences and Emotions in Social Media: Case Study on a Twitter Corpus

I. Dunđer*, M. Horvat** and S. Lugović**
*University of Zagreb, Faculty of Humanities and Social Sciences, Department of Information and Communication Sciences, Ivana Lučića 3, Zagreb, Croatia
**Zagreb University of Applied Sciences, Department of Computer Science and Information Technology, Vrbik 8, Zagreb, Croatia
idundjer@ffzg.hr, mhorvat1@tvz.hr, slugovic@tvz.hr

Abstract - Twitter is currently the most popular tool for social interaction and real-time information exchange. Outreach and importance of individual accounts is measured by the number of their followers. The aim of this paper is to investigate the applicability and usefulness of corpora containing textual and visual information for the purpose of machine observation of Twitter activities. The results in the presented analytic research are based on a data set of more than 16000 tweets collected from 22 startup founders' Twitter accounts with a large number of followers over a four-month period. Word usage in tweets was examined with natural language processing (NLP) techniques, applying word occurrence analyses and a manual qualitative evaluation of frequent words within the data set, primarily focusing on the distribution of words. Furthermore, profile pictures of Twitter accounts were collected in order to conduct a facial emotion analysis and emotion mining. Estimated basic emotional states were statistically compared with the number of tweets posted and the number of new followers gained during the observed timespan.

Keywords - word occurrences, emotion recognition, social media analysis, corpus analysis, natural language processing, Twitter

I. INTRODUCTION

Twitter is currently one of the most popular social interaction tool and is used for real-time information exchange. The number of Twitter followers hints the importance of individual accounts.

In this paper the authors would like to explore two techniques that could be utilised for the purpose of tweet analysis: statistical frequency analysis of words used in tweets and emotion mining from Twitter profile pictures of the startup founders.

As this is an exploratory study of mentioned techniques, the main focus is on techniques and the analyses are done on a smaller data set that was extracted from a larger data set, which is being collected since July 2015 [1].

From the authors' perspective this is a necessary step before applying the mentioned techniques on the whole collected data set. The statistical analysis of word occurrences is a language-independent technique and provides critical descriptive statistics of the content of messages, i.e. tweets. The authors of this paper were incrementally collecting tweets from all 28 EU member states startup founders. This paper also explains the overall data retrieval process, whereas the discussion of results and the conclusion are given after the presentation of the two different techniques and the related methodologies.

II. RESEARCH METHODOLOGY AND DATA COLLECTION

The necessary corpora, i.e. data set, is related to EU startup founders and was retrieved by the authors from one of the most widely used (2 million unique visitors a month) databases, F6S.com, which lists more than 110000 companies in its database [2]. In total, data for 50433 founders from 29 EU countries (including Norway) was retrieved, with the assistance of four freelance workers hired through the Upwork platform [3].

Out of those 50433 startup founders, 15912 had submitted private Twitter account information into the F6S.com database. Then, those 15912 accounts were manually checked. Furthermore, additional data was collected: the date the founders joined Twitter, the amount of tweets they posted, the amount of accounts they were following and the amount of their corresponding followers. Afterwards, Twitter accounts that had more than one hundred tweets and more than one hundred followers were selected.

Those chosen accounts, when filtered, were entered into the Twitter Archiving Google Spreadsheet (TAGS) retrieval system [4], which the authors of this very paper set to retrieve data from those filtered accounts on a daily basis. In total, there were 9696 startup founders that had more than one hundred tweets and more than one hundred followers, and those accounts were used as the input for the TAGS system.

For the purpose of the exploratory analysis presented in this paper, the authors decided to extract a smaller, not necessarily representative, set of Twitter accounts that were investigated. The authors decided to analyse only accounts from United Kingdom's startup founders and selected the first 40 accounts that were stored in the authors' database of collected tweets.

Then, those accounts were checked manually, i.e. the authors checked if the accounts were real private accounts of startup founders or if they were formal company accounts. Startup founders, when registering with F6S.com, have the option to submit company and

private Twitter accounts, but some of them choose only to put their company Twitter account in the field intended for private Twitter accounts.

Out of those 40 accounts there were 22 that belonged to real founders, while others were Twitter accounts of companies or did not exist anymore.

After the identification of real startup founders' accounts, their tweets were extracted automatically with TAGS in the period from August 3rd until November 17th 2015. In total, 16332 tweets were acquired.

The authors also collected profile pictures from those 22 founders in order to conduct an emotion analysis of their facial expressions.

III. Analysis of Research Results

The following two subsections will present more detailed information on techniques that were applied in order to acquire valuable feedback on tweet characteristics of UK's startup founders. The first subsection presents statistical natural language processing (NLP) approaches, whereas the second subsection explains the process of estimating emotions from pictures.

A. Statistical analysis of word occurrences

The collected tweets belonged to 22 startup founders from the UK. They were chosen according to the number of followers. The initial data set consisted of 16322 tweets that were sent in a timeframe from August 3rd and November 17th 2015. All the tweets were, at first, stored in a textual database with UTF-8 encoding, after which the database was preprocessed, i.e. normalised. This included the removing of, for this experiment unnecessary, variables, such as:

- ordinal numbers,
- dates,
- language codes,
- country codes,
- Twitter usernames,
- number of followers on Twitter,
- all retweets.

Language and country codes were unessential, as this exploratory study was focusing only on the English language and the startup phenomenon in the UK. The remaining tweets were lowercased in order to eliminate various instances of the same word. Also, links to web pages, Twitter usernames and keywords/topics (marked with a hashtag symbol) in a tweet were removed. Blank or duplicate tweets were also deleted. In total, 12834 tweets consisting of 130696 words, of which 22111 words were unique (distinct), remained after the normalisation process.

An analysis of the distribution of words over frequencies revealed that almost 11% (14325) of all words in the data set appeared only once (hapax legomena), which constituted cca. 65% of all unique words. This indicates that relatively rich vocabulary appears in the analysed tweets, which is reflected in the large number of different content words, i.e. open class words, such as nouns, verbs, adjectives, and adverbs that ensure semantics.

Still 1287 words appear more than 10 times and make up 6% of all unique words, i.e. 71.56% of all words in the Twitter data set. This is mainly due to the frequent usage of function words such as conjunctions, articles, pronouns etc., which are primarily used to form syntactically correct sentences or phrases by expressing syntactic relationships between different types of words.

Even though the analysed Twitter corpus was not a true representative of the English language, Fig. 1 still indicates a Zipfian distribution, i.e. a distribution of word occurrence probabilities that follows the Zipf's law. An elaborate examination is planned for future work, nevertheless, the figure shows that the data set contains rich vocabulary and that specific terminology is used, since most of the distinct words appear only once. This occurrence is delineated by a histogram that applies a nonlinear scale which presents a large range of quantities.

In this case, it is a histogram which uses a logarithmic scale to plot the frequency which covers a large range of values, while at the same time the bins have only a restricted range. The histogram presents log-frequencies of words with respect to the following specific bins:

- 1 (i.e. 6425 words)
- 2-10 (i.e. 4870 words),
- 11-100 (i.e. 1170 words),
- 101-1000 (i.e. 156 words),
- 1001-5000 (i.e. 14 words),
- more than 5000 (i.e. none).

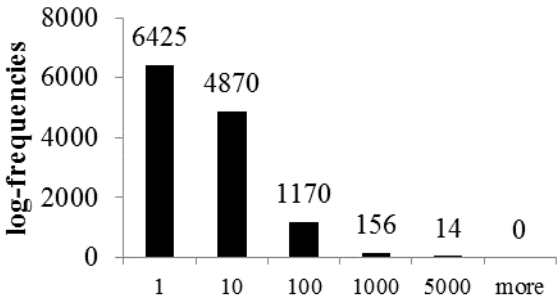


Figure 1. Histogram of log-frequencies of words with respect to specific bins.

This data visualisation responds very efficiently to skewness towards large values [5] and therefore points out data features that would not be easily observable if both variables had been plotted in a linear way.

Application of more task-specific natural language processing techniques, such as sentiment analysis [6] or statistical and linguistically-based terminology extraction [7], and a detailed analysis of specific concordances are left for future work, as such assessments might provide context-related information and explain certain word occurrences, multi-word expressions and collocations in

domain-related Twitter corpora.

A deeper assessment of the frequencies of open and closed class words appearing in the corpus is also planned. Furthermore, some aspects of this research will be repeated, but with different languages, and then compared with regard to a variety of different socio-economic and cultural surroundings of startups.

B. Emotion recognition from facial expressions

In this study the Twitter profile pictures of private startup founder accounts were collected for the purpose of estimating emotions from facial expressions. Based on the analysis of facial features and expressions, the profile pictures were classified according to six discrete emotion states or emotion norms: anger, disgust, fear, happiness, sadness and surprise. Finally, the mined emotions were compared to the number of posted tweets, the number of new followers and the accounts the startup founders were following during the same time period.

The inherent mechanism of emotion recognition from facial expressions has been shown to be universal across human races and cultures [8]. The technology that uses this phenomenon for automated estimation of emotions has also been demonstrated to be functional and exploitable in practice (for example [9], [10]).

In total 22 Twitter profile pictures were collected and analysed for facial features using the visage|SDK framework and FaceReader application [11]. Overall, the most dominant emotion is happiness (*avg*=0.39; *std*=0.25) and the least important norms are anger (*avg*=0.13; *std*=0.08), fear (*avg*=0.09; *std*=0.04) and surprise (*avg*=0.09; *std*=0.03). The statistical distribution of estimated emotion is provided in a box-plot diagram (Fig. 2).

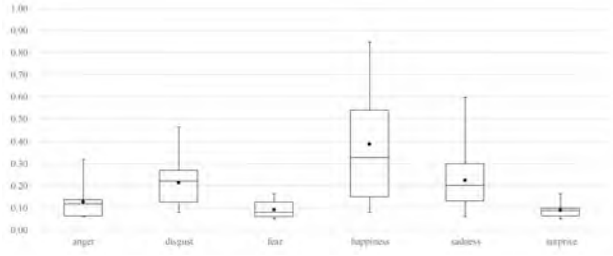


Figure 2. Box-plot diagram showing the distribution of estimated emotions in 22 Twitter profile pictures.

As could be intuitively expected happiness, as a basic emotion with continually positive valence, is the most appealing for a profile picture required to attract large audiences. Maximum value of happiness is 0.85 which is 41.67% more than the second most prevalent emotion and 84.78% more than the third. Interestingly, the second and the third desirable norms are sadness and disgust. At first, this may seem surprising but in different areas of applied psychology, for example such as in content marketing, studies have irrefutably shown that opposing visual cues, which trigger contrasting emotions, motivate the audience to see what you want them to see, feel what you want them to feel, and to do what you want them to do [12]. The use visual cues also affect content usability. Therefore, sadness and disgust can be explained as motivating emotions which together with the most prevalent happiness help to direct and shape the followers' attention. The least desirable emotions embedded in profile pictures are anger, fear and surprise which obviously all have negative valence and trigger unwanted connotations in observers.

Another obvious feature in the data set is strong fluctuation of emotions between different profile pictures. Happiness is the most varying emotion with *std*=0.25, almost twice as much as with sadness. Fear and surprise constantly have low values and vary the least, only 0.04 and 0.03, respectively. Furthermore, minimal values of all emotions are similar between 0.05 and 0.08. Respective maximal values are consistent with the rank which implies strong consistency between mean, maximal and standard deviation of emotions.

Values of statistical correlation (i.e. Pearson's *r*) between the difference in number of tweets, following and follower profiles, and basic emotions estimated in founders' profile pictures are given in Table 1.

TABLE I. CORRELATION BETWEEN THE DIFFERENCE IN TWEETS, FOLLOWING AND FOLLOWER PROFILES AND ESTIMATED BASIC EMOTIONS IN TWITTER PROFILE PICTURES.

	anger	**disgust**	**fear**	**happiness**	**sadness**	**surprise**
Tweets	-0.14	-0.11	-0.19	0.45	-0.07	-0.36
Following	0.07	0.04	0.45	-0.45	0.29	0.19
Followers	-0.13	-0.1	-0.43	**0.55**	**-0.55**	-0.12

The difference in the number of followers is statistically most significantly related with happiness and sadness, although the overall correlation is not strong. The number of followers is positively related to happiness, and negatively with sadness implying that the basic emotion of happiness, at any level, contributes in attracting more followers to a profile while sadness drives them away. Fear functions in a similar way as sadness but at even lesser statistical significance. Furthermore, happiness is positively correlated with the number of tweets the founders posted and negatively with the number of other profiles they were following. The emotion of fear has some relation with the number of following profiles. A more detailed exploratory analysis, including dominant emotions present in each profile picture, is planned for future.

All analysed founders do not follow the same emotional strategy in creating their profiles. From the available data three categories or groups of founders may be tentatively identified. In broad terms, the first group wants to induce happiness as a clearly dominant and strong emotion, while the second group choses a melange of happiness and sadness. Finally, the third group of analysed founders tries to remain neutral and not to provoke emotions at all. Their personal image is controlled with a specific content, intended meaning and connotations. In the third group elicitation of emotion is

secondary, while semantics is the most important. But apparently in all strategies fear and surprise are avoided as undesirable emotions. Each group has 5, 9 and 8 profiles, respectively.

Arguably, in the analysed set the most successful Twitter founders do not succumb to anger or fear, and do not act surprised. They are almost always positive, visibly marry or more often calm, sure of themselves, with a touch of sadness and a noticeable unfocused aversion.

However, an important caveat must be emphasized in this analysis: the set of 22 Twitter profile pictures is statistically inadequate to draw universal conclusions about all founders. Considering 15912 harvested accounts the test set represents <0.2% of the population. Therefore, in this initial estimation of emotions from profile pictures we restrained from categorizing Twitter founders in general. Rather, we quantitatively and, to a point, qualitatively described the analysed set. Whether the conclusions actually extrapolate to a larger proportion of the founders' population, or not, must be tested in subsequent research relying on an expanded corpus that encompasses a statistically significant number of founders.

IV. CONCLUSION

Twitter provides a real stream of messages that are not manipulated by algorithms (as it is the case in other social media platforms, such as Facebook). Such data are more valuable for the purpose of the analysis of human information behaviour and can provide better insights into different characteristics of Twitter corpora. Also, by analysing different signals from the same social media platform, such word occurrences and emotion estimations, as presented in this paper, could serve as a triangulation technique.

In social media research, cross-platform analyses are important for understanding data, but within the same platform one could have different types of signals recognised and analysed. By doing so, researchers could increase the granularity of the observation process of a socio-technical system.

This paper presents an initial exploratory study focusing on available techniques and leaves no space for any bold conclusions. What the authors can conclude is that the statistical analysis of word occurrences can be used to overcome language barriers in analysing tweets, and that it is possible to discover insights through correlations between emotions expressed in profile pictures and the account performance in terms of the number of followers. Twitter usage among private companies' employees is used as a performance indicator which impacts their income through different incentives. But there are strong ethical issues that should be considered. For example, where is the trade between privacy concerns and public benefits generated by learning from Twitter (or other social media) data?

REFERENCES

[1] S. Lugović, and W. Ahmed, "An Analysis of Twitter Usage Among Startups in Europe," In 5th International Conference The Future of Information Sciences (INFuture), pp. 299–308, 2015.

[2] F6S.com, https://www.f6s.com/f6s, 2015, online, last accessed on 19.09.2015.

[3] Upwork.com, https://www.upwork.com/, 2015, online, last accessed on 19.09.2015.

[4] M. Hawksey, "Twitter Archiving Google Spreadsheet (TAGS) v5. JISC CETIS MASHe: The Musing of Martin Hawksey (EdTech Explorer)," https://tags.hawksey.info/, 2013, online, last accessed on 19.09.2015.

[5] S. Th. Gries, "Useful statistics for corpus linguistics," In: A mosaic of corpus linguistics: selected approaches A. Sánchez and M. Almela (Eds.). Peter Lang: Frankfurt am Main, Germanny, pp. 269–291, 2010.

[6] A. Di Ciaccio, and G. M. Giorgi, "Statistical analysis of social networks," Rivista Italiana di Economia Demografia e Statistica, vol. LXVII(3/4), pp. 103–110, 2013.

[7] S. Seljan, I. Dunđer, and A. Gašpar, "From Digitisation Process to Terminological Digital Resources," In 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2013), pp. 1329–1334, 2013.

[8] P. Ekman, W. V. Friesen, and P. Ellsworth, "Emotion in the human face: Guidelines for research and an integration of findings," Elsevier, 2013.

[9] V. Terzis, C. N. Moridis, and A. A. Economides, "Measuring instant emotions based on facial expressions during computer-based assessment," Personal and ubiquitous computing, vol. 17(1), pp. 43–52, 2013.

[10] K. Ćosić, S. Popović, M. Horvat, D. Kukolja, B. Dropuljić, B. Kovač, and M. Jakovljević, "Computer-aided psychotherapy based on multimodal elicitation, estimation and regulation of emotion," Psychiatr. Danub., vol. 25, pp. 340–346, 2013.

[11] Visage Technologies, http://visagetechnologies.com/, 2015, online, last accessed on 19.09.2015.

[12] M. M. Aslam, "Are you selling the right colour? A cross‐cultural review of colour as a marketing cue," Journal of marketing communications, vol. 12(1), pp. 15–30, 2006.

The Application of Parameterized Algorithms for Solving SAT to the Study of Several Discrete Models of Collective Behavior

Stepan Kochemazov, Alexander Semenov, Oleg Zaikin
Matrosov Institute for System Dynamics and Control Theory SB RAS, Irkutsk, Russia
Email: veinamond@gmail.com, biclop.rambler@yandex.ru, zaikin.icc@gmail.com

***Abstract*—In this paper we present the computational study of one class of discrete models of collective behavior. In the context of these models a set of agents, that form a collective, is represented by a network. Each agent is assigned a special weight function. The behavior of a collective in discrete time moments is specified with a vector function, the coordinates of which are defined by values of agents weight functions at corresponding time moments. We study phenomena concerning the so-called conforming behavior: when an agent at some time moment decides to act or not to act depending on the similar decisions, made by agents from its neighborhood at the previous moment. We consider the problem how to dispose a relatively small number of always acting agents called activators in the network so that the majority of agents soon become active. We apply state-of-the-art SAT solvers to this problem. To tune the solver we use special parameterization techniques. In our computational experiments we show that the solving of considered problems of large dimension can be significantly sped up by finding effective combinations of the SAT solver parameters values on test instances from the same class but of much lower dimension.**

I. Introduction

Hereinafter, by collective V we mean a finite set of entities called agents. The notation $V = \{v_1, \ldots, v_m\}$ means that we consider a collective consisting of m agents. By collective behavior in the context of V we mean coordinated actions of some or all agents from V that take place at consecutive time moments. The research of phenomena regarding collective behavior is relevant in a number of areas, such as medicine, psychology, economics, robotics, etc.

In this paper we present the results of computational study of one class of discrete models of collective behavior that was introduced in [1]. The discrete nature of the model means that actions of agents happen at discrete time moments. In the models considered below we assume that at each time moment each agent chooses one of the variants of its behavior: action (1) or inaction (0). All agents make their decision simultaneously (synchronously). Such formulation makes it possible to apply to the study of collective behavior the so-called synchronous Boolean networks (SBN), also known as Kauffman networks [2].

In [1] for the proposed class of models we studied several combinatorial problems related to conforming and anticonforming behavior. To solve these problems we used state-of-the-art algorithms for solving Boolean satisfiability problem (SAT). In the present paper we develop the results obtained in [1] in the direction of significant increase of dimension of studied collectives. It became possible thanks to employing up-to-date parameterization techniques for SAT solvers.

Let us give a brief outline of the paper. In the next section we briefly describe the class of models from [1] and studied phenomena of collective behavior. Then we give an overview of state-of-the-art algorithms for solving SAT and describe the known parameterization techniques for SAT solvers. The section 4 is the main section of the paper. In it we apply the parameterization technique to SAT instances encoding the studied phenomena of conforming behavior. In accordance with this technique there are outlined sets of instances called training set and test set. The first set is used to find the values of SAT solver parameters with which it shows the best effectiveness. The second set is essentially the set of instances we want to solve. We show that for the studied phenomena of collective behavior we can use the training sets with instances from the same class as in test set, but of much lower dimension. The application of this parameterization technique made it possible to solve instances encoding the study of collective behavior of collectives with several thousand vertices.

II. Discrete dynamical models of collective behavior

So, assume that $V = \{v_1, \ldots, v_m\}$ is some collective. Let us associate with each agent from V a vertex of a directed graph G_V. Graph G_V should not have multiple arcs. According to the common notation [3], [4] we will refer to G_V as to a network. The orientation of arcs in G_V defines how different agents influence each other: in particular, the arc (v_j, v_i) interprets the fact that agent v_j influences the behavior of v_i. For an arbitrary agent v_i we will refer to the set of agents that influence its behavior in the aforementioned sense as neighborhood of v_i and denote it as V_i.

Let us assume that an arbitrary agent can choose between two types of behavior: 0 and 1. Informally, the 0-type corresponds to inaction and the 1-type corresponds to action. Hereinafter we consider only models with discrete time, i.e. $t \in \{0, 1, 2, \ldots\}$. Also we assume that at the initial time moment $t = 0$ an arbitrary agent demonstrates some type of behavior. At time moment $t + 1$ the behavior of an arbitrary

agent v_i is completely defined by behavior of agents $a_j \in V_i$ at time moment t.

We associate with each time moment $t \in \{0, 1, \ldots\}$ the set of all behavior types of all agents of the considered collective at that moment: it means we have Boolean vector $W(t) = (w_1, \ldots, w_m)$, $w_i \in \{0, 1\}$, $i \in \{1, \ldots, m\}$, to which we refer as to the state of network G_V at time moment t. For all $t > 0$ the bit w_i is produced via calculating a value of a special function that defines the way agents from V_i influence v_i. Such function is denoted as $f_{v_i}(t)$ and referred to as weight function of agent v_i. One can specify weight functions in a different manner. What is important is that the rules according to which they are constructed are fixed and do not depend on time moment t. Network G_V with weight functions associated with its vertices and with discrete dynamics defined as it was described above is called synchronous Boolean network (SBN).

It is easy to see that SBN G_V has 2^m different states (and the number of time moments here does not matter). The transitions of network G_V from one state to another are usually depicted in the form of directed graph Γ_{G_V}, that is called State Transition Graph (STG). Essentially Γ_{G_V} defines the function

$$F_{G_V} : \{0, 1\}^m \to \{0, 1\}^m$$

(by $\{0, 1\}^m$ we denote the set formed by all 2^m possible different Boolean vectors of length m).

Because the number of SBN states is finite, it means that network can traverse the same state in different moments of time. The corresponding situations define the so-called cycles. In some sources (for example [2], [5], [6]) they are referred to as attractors. More formally, the cycle of length $l - k$ is a sequence of network states $W(k), \ldots, W(l)$, in which $W(l) = W(k)$ (here k, l are some time moments, and $k < l$). The cycle of length 1 is called a stationary state or a fixed point.

In [1] we applied the apparatus of Boolean networks to model the so-called conforming behavior. The agent v_i demonstrating such behavior makes the decision 1 (action) at time moment $t + 1$ only if not less than fixed percent of agents from V_i (its neighborhood) at time moment t made decision 1. Otherwise at time moment $t+1$ agent v_i makes the decision 0. The percentage of the agents in V_i neighborhood, that has the described effect on the decision of v_i, defines the so-called conformity threshold. It should be noted that for the first time the threshold models of collective behavior were proposed in the famous work by Mark Granovetter [7]. Unfortunately, for the models from that paper at the moment of its writing there did not exist an effective computational apparatus that could have been applied to study its dynamical properties. For the models proposed by us in [1] we used state-of-the-art SAT solvers in the role of such apparatus. In particular, in [1] we considered SBN modeling the collective with the use of the following weight functions:

$$f_{v_i}(t+1) = \begin{cases} 1, \sum\limits_{v_j \in V_i} f_{v_j} \geq \theta_i \cdot |V_i| \\ 0, \sum\limits_{v_j \in V_i} f_{v_j} < \theta_i \cdot |V_i| \end{cases} \quad (1)$$

According to (1) agent v_i makes the decision to act at time moment $t + 1$ only if the number of active agents in its neighborhood V_i at time moment t is at least $\theta_i \cdot |V_i|$. Here $\theta_i \in [0, 1]$ specifies the conformity threshold of agent v_i.

We allow some agents to have conformity threshold equal to 0. These agents are always active (we additionally assume that they act at time moment $t = 0$). M. Granovetter suggested in his work to call such agents instigators (mainly assuming the sociological component of considered model). From our point of view it is more correct to call them activators since the processes that they start have a lot of similarities in areas not related directly to sociology (development of epidemics, dynamical processes in gene networks, development of economical crisis, etc.).

In [1] we have shown that the speed of network activation depends significantly on how well are activators positioned in it (assuming that at time moment $t = 0$ only activators are active). It can be seen on an example presented in Fig. 1. Activators correspond to crimson vertices. Blue vertices interpret agents that are inactive at this time step and orange vertices interpret active agents. Arc is marked as green if it goes from active vertex and as red if it goes from inactive vertex. The figure clearly shows that thanks to a better positioning 1 activator can activate network faster than 8 activators (note, that in the right part of the figure it can be seen that after 4 steps the system enters stationary state).

Thus, we had the following combinatorial problem: to find a disposition of relatively small number of activators in the network in such a way that almost all network agents become active in a relatively small number of time moments. This and closely related problems in [1] were reduced to Boolean satisfiability problem (SAT) and solved using state-of-the-art SAT solvers. Using parallel solvers at that moment we managed to solve considered problems for collectives with up to 600 agents. All experiments were made in application to random networks, generated in accordance with Erdos-Renyi [8], Watts-Strogatz [9] and Barabasi-Albert [10] models. In the next section we will describe the parameterization technique for SAT solvers that made it possible for us to achieve significant success in the direction of increasing dimension of considered collectives.

III. Boolean Satisfiability Problem (SAT) and Parameterization Techniques for SAT Solvers

The Boolean satisfiability problem (SAT) for an arbitrary Boolean formula in the conjunctive normal form (CNF) is to answer the question if this formula is satisfiable, i.e. if there exists such assignment of Boolean variables from this formula that makes it true. SAT is classical NP complete problem. It means that to SAT one can effectively reduce various combinatorial problems. Recent years witnessed a remarkable progress in the efficiency of SAT solving algorithms. Up-to-date algorithms successfully cope with SAT instances encoding combinatorial problems from various areas (ranging from symbolic verification to cryptanalysis) [11].

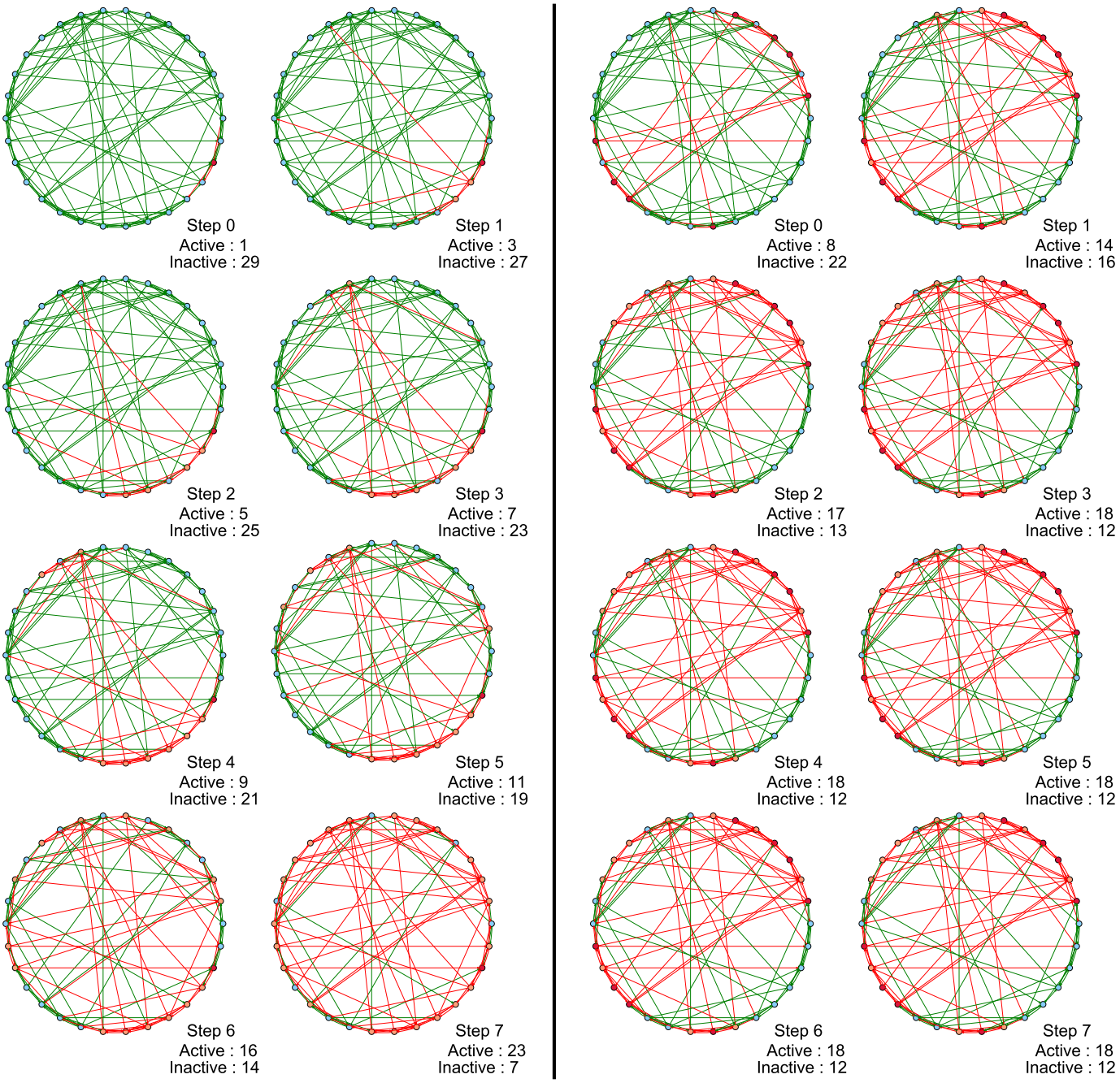


Fig. 1. The behavior of the network with 30 vertices, constructed according to Watts-Strogarz model, under the influence of two different dispositions of activators: 1 activator (left); 8 activators (right).

State-of-the-art SAT solvers are programs with a large number of various parameters. Meticulous tuning of these parameters for a particular class of problems may significantly improve average effectiveness of the solver: sometimes it is even possible to achieve more than 100 times better performance.

There are some generally accepted parameterization techniques for SAT solvers. According to papers [12], [13], [14] this parameterization assumes that we select from the considered set of SAT instances some subset called training set. On this training set (using parallel computing environment) the SAT solver with different combinations of parameter values is launched. For each combination of parameter values the value of special function is computed to measure the effectiveness of SAT solving with these parameter values. The parameterization process itself is essentially a local search in a Cartesian product of domains containing all possible values of considered parameters. To jump from local optimums sometimes various metaheuristic procedures are used. When the local search scheme finished its work either due to reaching

the optimality condition or because some time limit was exceeded, the resulting combination of parameters is then applied to solve all problems from the so-called test set. By test set it is usually meant the set of test instances used to measure the effectiveness of the parameterization procedure. Usually it is assumed that test set and training set do not intersect. In practice, given some test set one employs reasonable heuristics to form training set using problems of the same nature and similar dimension as in test set. In our paper we show that for some classes of problems it is possible to include in the training set the problems of much lower dimension than that in test set and achieve consistently good results.

In the next section we apply SAT solver parameterization techniques to the problems of study of collective behavior outlined above. Hereinafter we consider only networks generated according to Watts-Strogatz model (see [1], [9]) with parameters $k = 8$ and $\beta = 0.2$. The problem considered was to find such disposition of at most 10% activators that after 15 time steps at least 95% agents in the network become active, assuming that at the initial time moment only activators are active.

IV. Computational experiments

There are three state-of-the-art tools for tuning the parameters of SAT solvers: ParamILS (Iterated Local Search in Parameter Configuration Space [12]), GGA (Gender-based Genetic Algorithm [14]) and SMAC (Sequential Model-based Algorithm Configuration [13]). Using these tools one can speed up both local search and tree search SAT algorithms. The only requirement is that all the solver parameters must be discrete (each non-discrete parameter should be discretized). With the help of these tools two Configurable SAT Solver Challenges (CSSC) in 2013 and 2014 were held. According to [13] SMAC shows better efficiency compared to ParamILS and GGA. That is why in our computational experiments we used SMAC.

We chose LINGELING SAT solver [15] for parameters optimization, because at CSSC 2014 it won the gold medals in all categories (Industrial SAT+UNSAT, Crafted SAT+UNSAT, Random SAT+UNSAT, Random SAT). There are 323 parameters in LINGELING available for tuning.

In our experiments we used "Academician V.M. Matrosov" computing cluster of ISC SB RAS[1]. Each node of this cluster is equipped with two 16-core AMD Opteron 6276 CPUs and 64 GB of RAM. The LINGELING solver is a single-threaded application, so we launched it on one CPU core. We considered tests for collectives of 6 different dimensions (from 200 to 2000). For each variant we generated 100 different graphs and produced 100 corresponding SAT instances encoding the problem of search for good disposition of activators. In the Table I the characteristics of CNFs (on average) from the considered test sets are shown.

We launched solving SAT instances from all test sets using LINGELING with default values of parameters (see Table II).

[1]http://hpc.icc.ru

TABLE I
CHARACTERISTICS OF CNFS FROM THE TEST SETS

Dimension	Variables	Clauses	Size in MB
200	287780	684952	12
400	579412	1383256	26
600	929944	2234968	42
800	1175764	2817040	54
1000	1421052	3397936	66
2000	2942108	7056640	143

TABLE II
RESULTS FOR SOLVING SAT INSTANCES FROM THE CONSIDERED TEST SETS

	Default		Tuned	
Dimension	Solved	Mean	Solved	Mean
200	100	13.68	100	0.88
400	100	59.07	100	11.16
600	100	172.68	100	40.15
800	100	460.39	99	149.13
1000	98	1003.23	95	311.55
2000	19	3229.849	45	2057.32

In this experiment we used the time limit of 5000 seconds for every launch. After this we launched 64 instances of SMAC (with different random seeds) for 1 day (SMAC is single-threaded application too). In the role of training set we used 100 SAT instances corresponding to dimension 200. We compared the resulting combinations of parameters values for all 64 SMAC instances and chose the best one (according to the value of objective function in seconds). Then we checked the performance of the solver with found combination of parameters in application to SAT instances from the remaining 5 test sets. The results are summed up in Table II. For this purpose we used the same time limit (5000 seconds) as with default parameters.

In the Table II 'Mean' stands for mean time in seconds for the solved SAT instances in a test set. Value in the cell which corresponds to mean time for test set for 200 vertices is in fact the best value of the SMAC objective function (because we used the tests of dimension 200 for training). It can be seen that on the test sets for 400 and 600 vertices the tuned variant performs significantly better than the default one in terms of mean time (number of solved is the same in both cases). When considering the tests of dimension 800 and 1000 the tuned variant is still much better in terms of mean time, but the default one solved more instances within the time limit. On the test set encoding the study of graphs with 2000 vertices the tuned variant performed better both in terms of mean time and number of solved. From the observed results we conclude that the combination of parameter values found using tests with 200 vertices made it possible to significantly improve overall effectiveness of the solver in application to tests of much greater dimension.

V. Conclusion

In this paper we present the results of computational study of one class of discrete models of collective behavior, that was introduced by us earlier. To solve combinatorial problems

arising in the context of the considered models we used parameterized algorithms for solving Boolean satisfiability problem (SAT). The main novelty of our results lies in experimental justification of the fact that the tuning of SAT solver parameters can be performed on the set of randomly generated instances from the considered class with dimension significantly lower than the dimension of the tests we want to solve fast. In our experiments tuning the parameterized solver LINGELING in application to considered problems made it possible to increase its performance several times over that on default parameters values. The implementation of the described technique on a computing cluster made it possible to solve combinatorial problems of finding good dispositions of activators for networks with several thousand vertices.

ACKNOWLEDGMENT

The research was funded by Russian Science Foundation (project No. 16-11-10046).

REFERENCES

[1] S. Kochemazov and A. Semenov, "Using synchronous boolean networks to model several phenomena of collective behavior," *PLoS ONE*, vol. 9, no. 12, p. e115156, 12 2014.

[2] S. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, 1969.

[3] M. E. J. Newman, "The structure and function of complex networks," *Siam Review*, vol. 45, pp. 167–256, 2003.

[4] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes, "Critical phenomena in complex networks," *Rev. Mod. Phys.*, vol. 80, no. 4, pp. 1275–1335, Oct. 2008.

[5] E. Dubrova, M. Teslenko, and A. Martinelli, "Kauffman networks: Analysis and applications," in *Proceedings of the 2005 IEEE/ACM International Conference on Computer-aided Design*, ser. ICCAD '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 479–484.

[6] E. Dubrova and M. Teslenko, "A SAT-based algorithm for finding attractors in synchronous boolean networks." *IEEE/ACM Trans. Comput. Biology Bioinform.*, vol. 8, no. 5, pp. 1393–1399, 2011.

[7] M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[8] P. Erdös and A. Rényi, "On random graphs," *Publications Mathematicae*, vol. 6, pp. 290–297, 1959.

[9] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[10] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[11] A. Biere, M. J. H. Heule, H. van Maaren, and T. Walsh, Eds., *Handbook of Satisfiability*, ser. Frontiers in Artificial Intelligence and Applications. IOS Press, February 2009, vol. 185.

[12] F. Hutter, H. H. Hoos, and T. Stützle, "Automatic algorithm configuration based on local search," in *Proc. of the Twenty-Second Conference on Artifical Intelligence (AAAI '07)*, 2007, pp. 1152–1157.

[13] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Proc. of LION-5*, 2011, p. 50723.

[14] C. Ansótegui, M. Sellmann, and K. Tierney, "A gender-based genetic algorithm for the automatic configuration of algorithms," in *Principles and Practice of Constraint Programming - CP 2009, 15th International Conference, CP 2009, Lisbon, Portugal, September 20-24, 2009, Proceedings*, ser. Lecture Notes in Computer Science, I. P. Gent, Ed., vol. 5732. Springer, 2009, pp. 142–157.

[15] A. Biere, "Lingeling essentials, A tutorial on design and implementation aspects of the the SAT solver lingeling," in *POS-14. Fifth Pragmatics of SAT workshop, a workshop of the SAT 2014 conference, part of FLoC 2014 during the Vienna Summer of Logic, July 13, 2014, Vienna, Austria*, ser. EPiC Series, D. L. Berre, Ed., vol. 27. EasyChair, 2014, p. 88.

Logical-Algebraic Equations Application in Discrete-Event Systems Studying

Nadezhda Nagul
Matrosov Institute for System Dynamics and Control Theory,
Siberian Branch of the Russian Academy of Science,
Lermontov str., 134, 664033 Irkutsk, Russia
Email: sapling@icc.ru

***Abstract*—Discrete-event systems (DESs) are a wide range of models used primarily to deal with complex technological objects, e.g. communication networks and computer systems. To regulate DES behavior, Ramadge-Wonham supervisory control theory is often used. As usual, only a part of DES's event set is available for supervisor to obtain control pattern. However, sometimes information about the system state is provided by sensors. The issue of correlation between one essential property of supervisors for a DES with state and event observation and for a DES with event observation only is studied using the method of logical-algebraic equations, proved to be very useful in studying various dynamical systems.**

I. Introduction

This paper presents an application of the method of logical-algebraic equations (LAE-method) to the field of discrete-event systems (DESs). Lying at the intersection of system dynamics, algebra and logic, the LAE-method is a method of mathematical systems theory which serves to synthesize criteria for preservation properties of systems connected by special mappings called *morphisms*. One of the main applications of the preserving criteria obtained is the reduction of studying some complex system, say S, to studying a much simpler one, say S'.

To exploit the LAE-method [1], a property of the system under consideration is treated as a property of an algebraic system, name it $\mathfrak{A}$. Then we consider either a logical-algebraic equation $\mathcal{X}\&\mathcal{P} \Rightarrow \mathcal{P}'$ or $\mathcal{X}\&\mathcal{P}' \Rightarrow \mathcal{P}$. Here $\mathcal{P}$ is a formula predicate which describes the studied property of the system $\mathfrak{A}$, $\mathcal{P}'$ is a property of another algebraic system $\mathfrak{A}'$ corresponding to the system S', and $\mathcal{X}$ is the subject for searching. The second equation corresponds to the case of preservation of the property in the direction which is opposite to mappings acting from $\mathfrak{A}$ to $\mathfrak{A}'$. A solution, name it $\mathcal{R}$, of the chosen LAE is constructed algorithmically. Such a solution in place of $\mathcal{X}$ guarantees preserving the truth values of formula predicate $\mathcal{P}$ under the mappings of many-sorted algebraic system $\mathfrak{A}$ to $\mathfrak{A}'$. $\mathcal{R}$ is constructed in the form of traditional morphisms, i.e. it is of the meaning of preservation operations and relations only. Morphisms of dynamical systems proved to be especially useful, for example, for those procedures of studying stability and other dynamic properties that require changing variables, since we should ensure that the property under consideration in old variables is equivalent to that one in new variables, or at least guarantee its unidirectional preservation.

The author was supported by the Russian Science Foundation (project no. 16-11-00053)

In this paper the LAE-method is applied to studying the properties of partially observed DESs under control. Widely used, discrete-event systems describe system evolution by considering the occurrence of some event sequences. The development of DESs theory is driven by the rapid progress of manufacturing systems and communication networks, technological processes, transportation networks, automated and robotic systems, and others, primarily man-made systems. To regulate DES behavior, the Ramadge-Wonham framework of supervisory control is commonly used, which adopt ideas from logic, language and automaton theory. The concept of observability plays an important role in supervisory control theory. Based on practical constraints, a subset of observable events is distinguished from all events. Observability was initially defined for event observation only, no state observation was explicitly considered. In [2] it was first shown that the studying DES with event and state observation can be reduced to the same problem in the context of event observation only. Using the LAE-method, in [3] it was easily shown a connection between properties, which are basic for control implementation, of the language generated by DESs with event and state observation and the language generated by a common DES. In the same way here supervisor properties preservation is considered.

II. DES with State Observation

Let $\mathcal{G} = (Q, \Sigma, \delta, q_0, Q_m)$ be a discrete event system modeled as a generator of a formal language [4]. Here Q is the set of states q; Σ the set of events; δ: $\Sigma \times Q \to Q$ the transition function; $q_0 \in Q$ the initial state; $Q_m \subset Q$ the set of marker states. As usual, Σ^* denote the set of all strings over Σ, including the empty string ε. In general, δ is only a partial function. Language generated by $\mathcal{G}$ is $L(\mathcal{G}) = \{w : w \in \Sigma^*$ and $\delta(w, q_0)$ is defined$\}$, while language marked by $\mathcal{G}$ is $L_m(\mathcal{G}) = \{w : w \in L(\mathcal{G})$ and $\delta(w, q_0) \in Q_m\}$.

The Ramadge–Wonham supervisory control framework assumes the existence of a means of control $\mathcal{G}$ presented by a *supervisor* [4]. Let Σ_c be a controllable event set, $\Sigma_{uc} = \Sigma \setminus \Sigma_c$, $\Sigma_c \cap \Sigma_{uc} = \emptyset$. The supervisor switches control patterns so that the supervised discrete event systems achieve a control

objective described by some regular language K. Formally, a supervisor is a pair $\mathcal{J} = (\mathcal{S}, \phi)$ where $\mathcal{S} = (X, \Sigma, \xi, x_0, X_m)$ is a deterministic automaton with input alphabet Σ. $\mathcal{S}$ is considered to be driven externally by the stream of event symbols (words) generated by $\mathcal{G}$ (i.e. words from $L(\mathcal{G})$), while $\phi : X \to \Gamma$ is a (total) function that maps supervisor states x into control patterns $\gamma \in 2^{\Sigma}$. If $\mathcal{S}$ is in state x, the events $\sigma \in \Sigma_c$ of $\mathcal{G}$ are subject to control by $\phi(x)$. If $\sigma \in \phi(x)$, then σ is enabled, while if $\sigma \notin \phi(x)$ then σ is disabled (prohibited from occuring). Note that, unlike DES models with *forced events* [5], enabled events should not necessary occur. It is obvious that ϕ is the state feedback map. Because uncontrollable events cannot be disabled, it is required $\Sigma_{uc} \subseteq \gamma = \phi(x)$.

Construct the function $\xi \times \delta_c : \Sigma \times X \times Q \to X \times Q$, where $(\xi \times \delta_c)(\sigma, x, q) = (\xi(\sigma, x), \delta_c(\phi(x), \sigma, q))$ is defined iff $\delta(\sigma, q)$ is defined, $\sigma \in \phi(x)$ and $\xi(\sigma, x)$ is defined. Denote $L(\mathcal{J}/\mathcal{G})$ a language generated by the closed-looped behavior of the plant and the supervisor: $L(\mathcal{J}/\mathcal{G}) = \{w : w \in \Sigma^*$ and $(\xi \times \delta_c)(w, x, q)$ is defined$\}$. Let $L_m(\mathcal{J}/\mathcal{G})$ denote the language marked by the supervisor: $L_m(\mathcal{J}/\mathcal{G}) = \{w : w \in L(\mathcal{J}/\mathcal{G})$ and $(\xi \times \delta_c)(w, x_0, q_0) \in X_m \times Q_m\}$. The main goal of supervisory control is to construct such supervisor that $L_m(\mathcal{J}/\mathcal{G}) = K$.

Let $L \subset \Sigma^*$. The *closure* of L is the set of all strings that are prefixes of words of L, i.e. $\overline{L} = \{s | s \in \Sigma^*$ and $\exists t \in \Sigma^* : s \cdot t \in L\}$. Symbol $\cdot$ denotes string concatenation and is often omitted. A language L is *closed* if $L = \overline{L}$. If $\mathcal{G}$ is any generator then $L(\mathcal{G})$ is closed.

Let $L_c(\mathcal{J}/\mathcal{G}) \triangleq L(\mathcal{J}/\mathcal{G}) \cap L_m(\mathcal{G})$ denote the language controlled by $\mathcal{J}$ in $\mathcal{G}$, i.e. the set of those (marked) strings of the uncontrolled process language that "survive" in the presence of supervision [4].

Definition 1: A supervisor $\mathcal{J}$ is called *non-rejecting* if

$$\overline{L}_c(\mathcal{J}/\mathcal{G}) = \overline{L}_m(\mathcal{J}/\mathcal{G}).$$

A non-rejecting supervisor guarantees that every string $s \in \overline{L}_c(\mathcal{J}/\mathcal{G})$ which can be completed to "task" in $\mathcal{G}$, can also be completed to a task marked by $\mathcal{J}$.

Let $\mathcal{G}$ be partially observable, i.e. a set Σ_o of observable events is distinguished from all events, $\Sigma_{uo} = \Sigma \setminus \Sigma_o$, $\Sigma_c \cap \Sigma_{uo} = \emptyset$ [7]. The supervisor observes only events from the observable event set and, basing on this information, controls events in the controllable event set, by disabling them. The observation function is defined as the natural projection $P : \Sigma^* \to \Sigma_o^*$.

In [2] the supervisory control problem of DES with state and event observation was introduced. We borrow from there the problem formulation with some changes and necessary corrections. Assume that in addition to event observation, supervisor also possesses some information on the states of $\mathcal{G}$. It observes $y = h(q)$, where $h : Q \to Y$ is an output mapping from the set Q to the output set Y provided by some sensors or so on. In [2] other way then in [6] of taking into account state observation is suggested. In order to include state observation, the observation mapping is extended as follows. Since for a given string generated by $\mathcal{G}$ there is a unique sequence of states visited by the string and the supervisor observes the corresponding state output as well as the observable events, let such observations be characterized by the extended projection $P_s : L(\mathcal{G}) \to ((\Sigma_o \cup \{\varepsilon\}) \times Y)^*$ defined as

$$P_s(\varepsilon) = (\varepsilon, h(q_0)),$$

$$P_s(s\sigma) = P_s(s)(P(\sigma), h(\delta(s\sigma, q_0))).$$

Here, unlike [2], we do not admit multiple occurring of the artificial events corresponding to the output. Only one pair $(\sigma, y) \in \Sigma \times Y$ is added to the observation sequence, since adding more than one output is redundant for control purposes. Since $\mathcal{S}$ is now driven by the words of $L(\mathcal{G})$ along with restrictions on observation, a supervisor $\mathcal{J}$ is characterized by a mapping $\gamma_e : P_s(L(\mathcal{G})) \to 2^{\Sigma_c}$. Here $\gamma_e(P_s(s))$ is interpreted as a set of events enabled by γ_e after observing $P_s(s)$, $s \in L(\mathcal{G})$. The goal now is to construct such supervisor that $L_m(\gamma_e/\mathcal{G}) = K$.

We modify $\mathcal{G}$ by extending its states and events as follows:

$$\mathcal{G}_s = (Q \cup Q', \Sigma \cup Y, \delta_e, q_0', Q_m),$$

where Q' is the "double" of Q, that is, for each $q \in Q$, there is a double q' corresponding to q, and Q' is the set of all doubles; δ_s: $(\Sigma \cup Y) \times (Q \cup Q') \to Q \cup Q'$ is the extended transition function defined as follows:

$$\delta_s(\sigma, q) = \delta(\sigma, q)' \text{ if } \sigma \in \Sigma, q \in Q,$$

$$\delta_s(y, q) = q \text{ if } y = h(q), q \in Q,$$

$$\delta_s(y, q') = q \text{ if } y = h(q), q' \in Q',$$

and everything else is undefined. Thus, the supervisor for $\mathcal{G}_s$ is able to take the information about the output $y = h(q)$ into consideration since an output y is inserted before each occurrence of event. The artificial events in Y are considered to be uncontrollable and observable.

Now we can define the language generated and marked by $\mathcal{G}_s$ via the language generated and marked by $\mathcal{G}$. Let us define $e : \Sigma^* \to (\Sigma \cup Y)^*$ as follows:

$$e(\varepsilon) = h(q_0),$$

$$e(s\sigma) = e(s)\sigma h(\delta(s\sigma, q_0)),$$

which is different from [2] but we believe is more correct. Again, unlike [2], we do not repeat insertion of the output, since it is redundant. The following lemma may be easily proved in the same way as in [2].

Lemma 1: If $L(\mathcal{G})$ is not empty, then

1) $L(\mathcal{G}_s) = e(L(\mathcal{G}))$,
2) $T(L(\mathcal{G}_s)) = L(\mathcal{G})$,
3) $L_m(\mathcal{G}_s) = e(L_m(\mathcal{G}))$,

where $T : (\Sigma \cup Y)^* \to \Sigma^*$ is the projection.

Let the desired language K is extended to K_s as

$$K_s = T^{-1}(K) \cap L_m(\mathcal{G}).$$

Lemma 2: [2] Assume that $K = \overline{K} \cap L_m(\mathcal{G})$. Then

1) $\overline{K}_s = T^{-1}(\overline{K}) \cap L(\mathcal{G}_s)$,
2) $K_s = \overline{K}_s \cap L_m(\mathcal{G}_s)$,
3) $T(\overline{K}_s) = \overline{K}$.

A supervisor for $\mathcal{G}_s$ is a mapping $\gamma_s : P(L(\mathcal{G}_s)) \to 2^{\Sigma_c \cup Y}$. In [2] it was proved that non-blocking γ_s such that $L_m(\gamma_s/\mathcal{G}_s) = K_s$ exists iff γ_e such that $L_m(\gamma_e/\mathcal{G}) = K$ exists and an existence condition for γ_s is expressed in terms of controllability and observability of K_s. Moreover, given γ_e we can synthesize γ_s by letting $\gamma_s(P_s(t)) = \gamma_e(P(e(t))) \setminus Y$ for any $t \in L(\mathcal{G})$. Backward, for any $w \in L(\mathcal{G}_s)$ $\gamma_e(P(w)) = \gamma_s(P_e(s)) \cup Y \supseteq \Sigma_{uc}$ where s is a string such that $e(s) = w$. However, no properties of γ_s and γ_e other then non-blockness were discussed. This issue is considered in what follows, or more exactly, the method of logical-algebraic equations is used to show non-rejecting property preservation while transition from γ_s to γ_e and vice versa.

III. The Method of Logical-Algebraic Equations

To exploit the LAE-method we treat the property of the system under consideration as a property of an algebraic system. Due to the complex nature of dynamical systems, the process of algebraizing of their models usually leads to many-sorted algebraic systems (MASs) where the basic sets have the meaning of a state space, a time scale, etc. Moreover, in [1] we introduced the notion of a *general many-sorted algebraic system of finite type* (GMAS) $\mathfrak{A} = \langle A, \Omega_F, \Omega_P, \Omega_E \rangle$, where $A = \{A_\lambda | \, \lambda = \overline{1,k}\}$ is a family of basic sets, $\Omega_F = \{\mathbf{F}_\beta^{n_\beta} | \, \mathbf{F}_\beta^{n_\beta} : S_{1\beta}[A] \times S_{2\beta}[A] \times \ldots \times S_{n_\beta \beta}[A] \to S_{n_\beta+1,\beta}[A], \, \beta = \overline{1,k_F}\}$ a set of functions, $\Omega_P = \{\mathbf{P}_\gamma^{n_\gamma} | \, \mathbf{P}_\gamma^{n_\gamma} \subseteq T_{1\gamma}[A] \times \ldots \times T_{n_\gamma \gamma}[A], \, \gamma = \overline{1,k_P}\}$ a set of relations, $\Omega_E = \{\mathbf{E}_\delta | \, \mathbf{E}_\delta \in U_\delta[A], \, \delta = \overline{1,k_E}\}$ a set of distinguished elements. The elements of the set $\Omega_F \cup \Omega_P \cup \Omega_E$ are defined on extended Bourbaki *steps* over the family A. To give a notion of a step, consider the set of k symbols $\sigma_a = \{a_\lambda | \lambda = \overline{1,k}\}$ $(k > 0)$ and the signature $\sigma_0 = \{\sqcup \times \sqcup, \mathcal{P}, \mathcal{N}\}$. We define a *step scheme* over σ_a (or just *scheme* for brevity sake) as follows:

1) for all $\lambda = \overline{1,k}$ a_λ is a scheme;
2) if S_1, S_2 are schemes then $(S_1 \times S_2)$ is a scheme;
3) if S is a scheme then $\mathcal{P}(S)$ and $\mathcal{N}(S)$ are schemes;
4) expression is a scheme iff it follows from the rules 1–3.

$St[\sigma_a]$ will denote the *set of all scheme steps* S over σ_a. Given a family of sets $A = \{A_\lambda | \, \lambda = \overline{1,k}\}$, for any step scheme $S \in St[\sigma_a]$ by induction on the construction of the scheme S we can naturally and unambiguously define the set $S[A]$ called *a step over the family A* (built with the scheme S). In this case:

1) symbols $a_\lambda \in \sigma_a$ are interpreted as sets $A_\lambda \in A$;
2) the binary symbol $\sqcup \times \sqcup \in \sigma_0$ is interpreted as the operation of constructing Cartesian product, i.e. for any schemes S_1, $S_2 \in St[\sigma_a]$ we assume $(S_1 \times S_2)[A] = S_1[A] \times S_2[A]$;
3) the function symbol $\mathcal{P} \in \sigma_0$ denotes Boolean constructing, i.e. for the scheme $\mathcal{P}(S) \in St[\sigma_a]$ we assume $\mathcal{P}(S)[A] = 2^{S[A]}$;
4) the symbol $\mathcal{N} \in \sigma_0$ denotes sequences set constructing, i.e. for the scheme $\mathcal{N}(S) \in St[\sigma_a]$ we assume $\mathcal{N}(S)[A] = (S[A])^{\boldsymbol{N}}$.

Hence, the classical concepts of a scheme and a Bourbaki step [8] are extended with the operation of sequence forming, by introducing to σ_0 the extra character $\mathcal{N}$ and the corresponding interpretation of schemes. The set of all such steps, generated as previously described, is called *generalized scale of steps* over *A* and denoted by $St[A]$. Due to the generalization of a scheme and step concepts, it is easy to represent in the form of GMAS various models of dynamic systems and, in particular, DES dealing with events sequences.

We describe the considered system property with a formula predicate $\mathcal{F}(\overline{x}) \stackrel{df}{=} \mathcal{F}(x_1, \ldots, x_p)$ of the signature σ of the chosen GMAS $\mathfrak{A}$, where x_μ is a free variable, $\mu = \overline{1,p}$, $p \geq 0$. Without loss of generality, the formula $\mathcal{F}$ is considered to be formed of *literals* (*concluding statements*, or *c-formulas*) $\mathcal{F}^\nu$, that is, atomic formulas $\mathcal{F}_+^\nu$ or their negations $\mathcal{F}_-^\nu$, with the help of connectives &, $\vee$, and type quantifiers $\hat{\omega}_\alpha \stackrel{df}{=} \forall z_\alpha : Z_\alpha \stackrel{df}{=} \forall z_\alpha (Z_\alpha \Rightarrow \sqcup)$ (universal type quantifier), $\check{\omega}_\alpha \stackrel{df}{=} \exists z_\alpha : Z_\alpha \stackrel{df}{=} \exists z_\alpha (Z_\alpha \& \sqcup)$ (existential type quantifier), $\alpha = \overline{1,n}$, $\nu = \overline{1,M}$. Such formulas, unlike positive formulas used, for example, in general algebraic system theory, are referred to as *generalized positive formulas*. The notion of "negation normal form" (ref., for example, [9]) is the analog of the notion of generalized positive formulae except that the latter is built with type quantifiers instead of classic quantifiers. Z_α are called *type conditions*.

Let us have a family of mappings

$$\varphi = \{\varphi_\lambda | \varphi_\lambda : A_\lambda \to A'_\lambda, \lambda = \overline{1,k}\}, \qquad (1)$$

that map the basic sets of GMAS $\mathfrak{A}$ to the basic sets $A' = \{A'_\lambda | \lambda = \overline{1,k}\}$ of the GMAS $\mathfrak{A}' = \langle A', \Omega'_F, \Omega'_P, \Omega'_E \rangle$ which is of the same type as $\mathfrak{A}$. The "same type" means that the powers of the sets *A* and *A'*, Ω_F and Ω'_F, and so on accordingly match, and the step $S'_{1\beta}[A'](S'_{2\beta}[A']$, $T'_\gamma[A']$, $U'_\delta[A']$ respectively) is formed from the sets A'_λ with the same scheme as $S_{1\beta}[A]$ ($S_{2\beta}[A]$, $T_\gamma[A]$, $U_\delta[A]$ respectively) from the sets A_λ. The aim is to find a way of synthesizing the formula predicate $\mathcal{F}(\overline{x})$ preserving conditions in respect to mappings (1).

Let us add to the signature σ a set of symbols $\sigma' = \sigma'_a \cup \sigma'_P \cup \sigma'_F \cup \sigma'_E$ which doubles the signature σ, $\sigma' \cap \sigma = \emptyset$. We build the formula $\mathcal{F}' = \mathcal{F}(P'/P, F'/F, E'/E, x'_\mu/x_\mu, z'_\alpha/z_\alpha)$ of the signature σ', where expression X'/X means the substitution of all entries of the symbol X with the symbol X', i.e. each symbol $P \in \sigma_P$, $F \in \sigma_F$, $E \in \sigma_E$, …, is replaced with a new corresponding symbol $P' \in \sigma'_P$, $F' \in \sigma'_F$, $E' \in \sigma'_E$, …, $\mu = \overline{1,p}$, $\alpha = \overline{1,n}$. The formula $\mathcal{F}'(\varphi^{|x_1|}(x_1), \ldots, \varphi^{|x_p|}(x_p))$ thus obtained describes the same property of the system $\mathfrak{A}'$ which $\mathcal{F}(\overline{x})$ describe for the system $\mathfrak{A}$. Symbol $\varphi^{|x_i|}(x_i)$ denotes *the canonical expansion of mappings (CEM)* φ on the step $S[A] = |x_i|$ to which variable x_i belong.

Next we construct logic-algebraic equations (LAEs)

$$\mathcal{X} \& \mathcal{F}(x_1, \ldots, x_p) \Rightarrow \mathcal{F}'(\varphi^{|x_1|}(x_1), \ldots, \varphi^{|x_p|}(x_p)), \qquad (2)$$

$$\mathcal{X}\&\mathcal{F}'(\varphi^{|x_1|}(x_1),\ldots,\varphi^{|x_p|}(x_p)) \Rightarrow \mathcal{F}(x_1,\ldots,x_p). \quad (3)$$

Preserving conditions in respect to mappings (1) in a direction which coincides with the direction of the mappings between systems are to seek as a solution of the LAE (2), while preserving conditions in the opposite direction are to seek as a solution of the LAE (3). The detailed algorithm for finding non-trivial solutions of (2) and (3) is presented in [1]. The main steps of the algorithm are the following:

Step 1 Separate the components of the signature σ in concluding statements: on the base of $\mathcal{F}(\overline{x})$ construct a formula $\Psi(\overline{x}) = \Psi(x_1, \ldots, x_p)$ which does not contain functional symbols in c-statements.

Step 2 Generate the preserving conditions in terms of CEMs. Denote them $\mathcal{R}_1$ for LAE (2) and $\mathcal{P}_1$ for LAE (3).

Step 3 Split the formulas obtained on the pervious step to morphism-like conditions. Construct the formula $\mathcal{R}$ as a solution of the LAE (2) or $\mathcal{P}$ as a solution of the LAE (3).

Conditions $\mathcal{R}_1$ and $\mathcal{P}_1$, obtained at step 2, are already sufficient to guarantee preservation of the property $\mathcal{F}$. Yet, as usual, they are too lengthy to deal with therefore at step 3 split into conjunctions of simpler formulas which include in some sense "minimal" number of existential quantifiers and concluding statements. In [10] we introduced the notion of "standard splitting" which assumes the splitting $\mathcal{R}_1$ (or $\mathcal{P}_1$) according to the number of concluding statements, and then each part thus obtained are split according to existential quantifiers. Standard splitting allows one to generate instead of, for example, $\mathcal{R}_1$ a set of formulas collectionwise sufficient for $\mathcal{R}_1$. Each formulae of the set represents a condition of preserving a single symbol of the signature, functional or relation, in accordance with the traditional definitions of morphisms of algebraic systems. Next, the conditions obtained are to be interpreted in the language of application domain. Standard splitting leads to obtaining the simplest structure of the final formulas. This, in turn, allowed us to define the classes of properties which are preserved by the morphisms of the same class [10].

IV. Supervisor Properties Preservation

In [2] it was proved that since state observation provides more information about the system, the existence condition for a supervisor with event and state observation is weaker than the existence condition for a supervisor with event observation only. Moreover, such conditions may be expressed in terms of controllability and observability of K_s. Using the LAE-method, in [3] we examined the connection between properties of the languages K and K_s, basic for control problems. It was proved that controllability is not affected by the state observation and that K is observable then K_s is observable while the opposite is not the case. These facts were previously demonstrated in [2] but the proof considerably differs from the one presented in [3]. In this section we exploit the LAE-method to show non-rejecting property preservation when state observation is added to DES.

Let $\mathcal{G}$ be a DES controlled by a supervisor $\mathcal{J} = (\mathcal{S}, \phi)$ with the state observation included, i.e. $\mathcal{J}$ realize the mapping $\gamma_e : P_s(L(\mathcal{G})) \to 2^{\Sigma_c}$. The non-rejecting property of $\mathcal{J}$ (Definition 1) is informally expressed as

$$\forall s \in \overline{L}_c(\mathcal{J}/\mathcal{G}) \quad \exists t \in L(\mathcal{J}/\mathcal{G})(s\cdot t \in L_m(\mathcal{J}/\mathcal{G}) \Rightarrow \Rightarrow s\cdot t \in L_c(\mathcal{J}/\mathcal{G})). \quad (4)$$

Here the expression "$s \cdot t \in L_m(\mathcal{G})$" is the informal note of the atomic formula $L_m(\mathcal{G})(s \cdot t)$ where $L_m(\mathcal{G})$ is a predicate symbol.The atomic formula $L_m(\mathcal{G})(s)$ is true if $s \in \boldsymbol{L}_m(\mathcal{G})$, i.e. the string s belong to the language marked by $\mathcal{G}$. It forms a literal $\mathcal{F}^{\nu}_{+}$ of the language of the generalized positive formulas. The expression "$s \cdot t \notin L(\mathcal{J}/\mathcal{G})$" is equivalent note of the negation of an atomic formula so it is the literal of the type $\mathcal{F}^{\nu}_{-}$. The expression "$\forall s \in \overline{L}_c(\mathcal{J}/\mathcal{G})$" is the informal note of the type quantifier $\forall s : \overline{L}_c(\mathcal{J}/\mathcal{G})$. Thus (4) will be transformed to a generalized positive formulae taking into account the fact that $s \cdot t \in L_m(\mathcal{J}/\mathcal{G})$ is equivalent to

$$s\cdot t \in L(\mathcal{J}/\mathcal{G})\,\&\,(\xi\times\delta_c)(t,\xi(P_s(s),x_0), \delta_c(\phi(\xi(P_s(s),x_0)),s,q_0)) \in X_m\times Q_m,$$

the equivalence $A \Rightarrow B = \neg A \vee B$ and the definition of the language $L_c(\mathcal{J}/\mathcal{G})$. Then obvious modifications lead to

$$\mathcal{F} = \forall s : \overline{L}_c(\mathcal{J}/\mathcal{G})\,\exists t : L(\mathcal{J}/\mathcal{G})\,(s\cdot t \notin L(\mathcal{J}/\mathcal{G}) \vee \vee\ s\cdot t \in L_m(\mathcal{G})\ \vee\ (\xi\times\delta_c)(t,\xi(P_s(s),x_0), \delta_c(\phi(\xi(P_s(s),x_0)),s,q_0))) \notin X_m\times Q_m). \quad (5)$$

The property (4) of the DES $\mathcal{J}/\mathcal{G}$ may be treated as the property of the triple-sorted algebraic system of finite type $\mathfrak{A} = \langle A, \Omega_F, \Omega_P, \Omega_E\rangle$, which is described with the formula $\mathcal{F}_1$ without free variables, where $A = \{X, Q, \Sigma, Y\}$, $\Omega_F = \{\delta_c, \xi, \phi, \cdot, P_s\}$, $\Omega_P = \{\overline{\boldsymbol{L}}_c(\mathcal{J}/\mathcal{G}),\ \boldsymbol{L}(\mathcal{J}/\mathcal{G}), \boldsymbol{L}_m(\mathcal{G}), \boldsymbol{X}_m \times \boldsymbol{Q}_m\}$, $\Omega_E = \{x_0, q_0\}$. Since $\overline{\boldsymbol{L}}_c(\mathcal{J}/\mathcal{G})$, $\boldsymbol{L}(\mathcal{J}/\mathcal{G})$, $\boldsymbol{L}_m(\mathcal{G}) \subseteq \Sigma^{\mathbb{N}}$, then each of the predicates $\overline{\boldsymbol{L}}_c(\mathcal{J}/\mathcal{G})$, $\boldsymbol{L}(\mathcal{J}/\mathcal{G})$, $\boldsymbol{L}_m(\mathcal{G})$ is connected with the scheme $\mathcal{N}(a_3)$. Schemes a_1 and a_2 correspond to the predicate $\boldsymbol{X}_m \times \boldsymbol{Q}_m$ because of $X_m \times Q_m \subseteq X \times Q$. In what follows we do not distinguish predicates and predicate symbols notations, i.e. $\boldsymbol{L}(\mathcal{J}/\mathcal{G})$ is denoted by $L(\mathcal{J}/\mathcal{G})$.

Let a supervisor for $\mathcal{G}_s = (Q \cup Q', \Sigma \cup Y, \delta_e, q'_0, Q_m)$, realizing a mapping $\gamma_s : P(L(\mathcal{G}_s)) \to 2^{\Sigma_c \cup Y}$, is a structure $\mathcal{J}_s = (\mathcal{S}_s, \phi_s)$ where $\mathcal{S}_s = (X_s, \Sigma_s, \xi_s, x^s_0, X^s_m)$, $\phi_s : X_s \to 2^{\Sigma^s}$, and GMAS $\mathfrak{A}_s$ is of the same type as GMAS $\mathfrak{A}$. We introduce the family of mappings $\varphi = \{\varphi_\lambda | \varphi_\lambda, \lambda = \overline{1,4}\}$ as follows:

$$\varphi_1 : X_s \to X, \varphi_2 : Q_s \to Q, \quad \varphi_3 : \Sigma_s \to \Sigma, \varphi_4 : Y_s \to Y, \quad (6)$$

where $\Sigma_s = \Sigma \cup Y$, $Q_s = Q \cup Q'$, $Y_s = Y$.

Note that according to step 1 of the algorithm, new variables will be added to eliminate entries of the symbols of σ_F

from the concluding statements of $\mathcal{F}$. To omit bulky formal manipulations of steps 2 and 3 we are going to use results obtained in [10]. For any formula $\mathcal{F}(\overline{x})$ let $ex(\mathcal{F})$ ($all(\mathcal{F})$, respectively) denote a set of variables of existential type quantifiers $\breve{w}_\alpha$ of $\mathcal{F}$ (universal $\hat{w}_\alpha$, respectively). Let Q_{ex} (Q_{all}, respectively) denote a set of predicates which form type conditions of quantifiers for the variables from $ex(\mathcal{F})$ ($all(\mathcal{F})$, respectively). In case of (5) $Q_{ex} = \{L(\mathcal{J}/\mathcal{G})\}$, $Q_{all} = \{\overline{L}_c(\mathcal{J}/\mathcal{G}))\}$, A set of predicate symbols which form c-statements $\mathcal{F}^\nu_+$, we denote as $pos(\mathcal{F})$, while a set of predicate symbols which form c-statements $\mathcal{F}^\nu_-$ we denote as $neg(\mathcal{F})$. In (5) $pos(\mathcal{F}_1) = \{L_m(\mathcal{G})\}$, $neg(\mathcal{F}_1) = \{L(\mathcal{J}/\mathcal{G}), X_m \times Q_m\}$. Next we form the sets

$$\Omega_{P+} = Q_{ex} \cup pos(\mathcal{F}) = \{L(\mathcal{J}/\mathcal{G}), L_m(\mathcal{G})\},$$

$$\Omega_{P-} = Q_{all} \cup neg(\mathcal{F}) = \{\overline{L}_c(\mathcal{J}/\mathcal{G}), L(\mathcal{J}/\mathcal{G}), X_m \times Q_m\},$$

$$\Omega_{P^+_-} = \Omega_{P+} \cap \Omega_{P-} = \{L(\mathcal{J}/\mathcal{G})\}.$$

Let $\mathcal{I} \subseteq \{1, \ldots, k\}$ be a subset of the index set of the family (1). The definition below generalizes for GMAS the notion of a morphism of single-sorted algebraic systems and the Lyndon's notion of a Q-morphism of models [11].

Definition 2: Let Θ, Ξ be some sets of relation symbols, $\Theta, \Xi \subseteq \Omega_P$, $\Theta \cap \Xi = \varnothing$. The family of mappings (1) is said to be $\mathcal{I}$-injective $\Theta\Xi$-morphism of GMAS $\mathfrak{A} = \langle A, \Omega_F, \Omega_P, \Omega_E\rangle$ to $\mathfrak{A}' = \langle A', \Omega'_F, \Omega'_P, \Omega'_E\rangle$ if

1) $\langle\varphi\rangle^{S_{n_\beta+1,\beta}[A]}(\mathbf{F}^{n_\beta}_\beta(z_1, \ldots, z_{n_\beta})) = (\mathbf{F}^{n_\beta}_\beta)'(\langle\varphi\rangle^{S_{1\beta}[A]}(z_1), \ldots, \langle\varphi\rangle^{S_{n_\beta\beta}[A]}(z_{n_\beta}))$, $z_1 \in S_{1\beta}[A]$, $\ldots$, $z_{n_\beta} \in S_{n_\beta\beta}[A]$, $\beta = \overline{1, k_F}$;

2) $\langle\varphi\rangle^{U_\delta[A]}(\mathbf{E}_\delta) = \mathbf{E}'_\delta$, $\delta = \overline{1, k_E}$; and for all $\gamma = \overline{1, k_P}$

$$\langle\varphi\rangle^{T_{1\gamma}[A]\times\ldots\times T_{n_\gamma\gamma}[A]}(\mathbf{P}^{n_\gamma}_\gamma) \subseteq (\mathbf{P}^{n_\gamma}_\gamma)',$$

if $\mathbf{P}^{n_\gamma}_\gamma$ correspond to a predicate symbol $P_\gamma \in \Theta$;

$$(\mathbf{P}^{n_\gamma}_\gamma)' \subseteq \langle\varphi\rangle^{T_{1\gamma}[A]\times\ldots\times T_{n_\gamma\gamma}[A]}(\mathbf{P}^{n_\gamma}_\gamma),$$

if $\mathbf{P}^{n_\gamma}_\gamma$ correspond to a predicate symbol $P_\gamma \in \Xi$;

$$\langle\varphi\rangle^{T_{1\gamma}[A]\times\ldots\times T_{n_\gamma\gamma}[A]}(\mathbf{P}^{n_\gamma}_\gamma) = (\mathbf{P}^{n_\gamma}_\gamma)',$$

if $\mathbf{P}^{n_\gamma}_\gamma$ correspond to a predicate symbol $P_\gamma \in \Omega_P \setminus (\Theta \cup \Xi)$; and the mappings φ_λ are injective for all $\lambda \in \mathcal{I}$.

Let $\mathcal{I}_{S[A]}$ denote the index set of the basic sets A_λ which are used to construct the step $S[A]$ with the scheme S.

Theorem 1 ([10]): Let $\mathcal{F}(x_1, \ldots, x_q)$ be a generalized positive formula with free variables x_1, $\ldots$, x_q, $q \geqslant 0$. Let $\mathcal{I} = \bigcup_{S[A]} \mathcal{I}_{S[A]}$, where $S[A] = |v|$, v are all variables z_β or x_μ that enter the equalities $z_\alpha \equiv v$, which are included in c-statements $=_-$ or type conditions of the universal quantifiers $\hat{z}_\alpha$ where z_α is a defining variable. Then satisfiability of $\mathcal{F}$ implies satisfiability of $\mathcal{F}'$ under $\mathcal{I}$-injective $(\Omega_{P+} \setminus \Omega_{P^+_-})(\Omega_{P-} \setminus \Omega_{P^+_-})$-morphism of GMAS $\mathfrak{A}$ to GMAS $\mathfrak{A}'$.

The theorem says that if the mappings (6) are defined and proved to be $\mathcal{I}$-injective $(\Omega_{P+} \setminus \Omega_{P^+_-})(\Omega_{P-} \setminus \Omega_{P^+_-})$-morphism then the non-rejecting property of supervisor $\mathcal{J}_s$ is preserved for the supervisor $\mathcal{J}$. Note that for (5) the set $\mathcal{I}_{S[A]}$ is empty so injectivity is not required.

Let $\varphi_2 : Q_s \to Q$ be defined as $\varphi_2(q_s) = q$ for both $q_s = q$ and $q_s = q'$, and let $\varphi_3 : \Sigma_s \to \Sigma$ be defined as

$$\varphi_3(\sigma_s) = \begin{cases} \sigma, & \text{if } \sigma_s = \sigma \in \Sigma; \\ \varepsilon, & \text{if } \sigma_s = y \in Y, \end{cases}$$

i.e. φ_3 is projection. φ_4 is identity map. To define φ_1 we will use conditions of Definition 2. According to condition 1 for the function $\xi_s : \Sigma_s \times X_s \to X_s$ the following equality is required:

$$\langle\varphi\rangle^{S_3[A]}(\xi_s(\sigma_s, x_s)) = \xi(\langle\varphi\rangle^{S_1[A]}(\sigma_s), \langle\varphi\rangle^{S_2[A]}(x_s)),$$

where $\langle\varphi\rangle^{S_i[A]}$ is the CEM of the family φ over the step $S_i[A]$, and S_i is the step schemes corresponding to the function ξ, $i = \overline{1,3}$. In case of ξ these schemes are $S_1 = S_3 = a_1$, $S_2 = a_2$ therefore $S_1[A] = S_3[A] = X_s$, $S_2[A] = \Sigma_s$ and we obtain the condition

$$\varphi_1(\xi_s(\sigma_s, x_s)) = \xi(\varphi_3(\sigma_s), \varphi_1(x_s)). \tag{7}$$

For the function $\phi_s : X_s \to 2^{\Sigma_s}$ it is required

$$\widehat{\varphi_3}(\phi_s(x_s)) = \phi(\varphi_1(x_s)). \tag{8}$$

According to the definition of φ_3, this means $\phi(\varphi_1(x_s)) = \phi_s(x_s) \setminus Y$ and obviously correlates with [2].

For the function $\delta^s_c : \Gamma_s \times \Sigma_s \times Q_s \to Q_s$

$$\varphi_2(\delta^s_c(\phi_s(x_s), \sigma_s, q_s)) = \\ = \delta_c(\widehat{\varphi_3}(\phi_s(x_s)), \varphi_3(\sigma_s), \varphi_2(q_s)) \tag{9}$$

should be valid. Suppose $\sigma_s = \sigma$. If $q_s = q$ and $\sigma \in \phi_s(x_s)$ then by (8) $\sigma \in \phi(\varphi_1(x_s))$ and $\delta_c(\widehat{\varphi_3}(\phi_s(x_s)), \varphi_3(\sigma_s), \varphi_2(q_s)) = \delta(\sigma, q) = \tilde{q} \in Q$. By definition $\delta^s_c(\phi_s(x_s), \sigma, q) = \delta_c(\phi_s(x_s), \sigma, q)' = \tilde{q}'$ and since $\varphi_2(\tilde{q}') = \tilde{q}$, (9) is valid. If $\sigma \notin \phi_s(x_s)$ or $q_s = q'$ then neither part of (9) is defined. Suppose $\sigma_s = y$. Then for both $q_s = q$ and $q_s = q'$ such that $y = h(q)$ $\delta^s_c(\phi_s(x_s), y, q) = q$. But $\delta_c(\widehat{\varphi_3}(\phi_s(x_s)), \varphi_3(\sigma_s), \varphi_2(q_s)) = \delta_c(\widehat{\varphi_3}(\phi_s(x_s)), \varepsilon, q) = q$ therefore (9) is again valid. Thus, condition (9) do not impose any additional constraints on the systems connection.

The functional symbol $\cdot$ is connected with three schemes $\mathcal{N}(a_3)$ and the following equality should be satisfied:

$$\varphi_3|^N(s_s \cdot_s t_s) = \varphi_3|^N(s_s) \cdot \varphi_3|^N(t_s). \tag{10}$$

Since φ_3 is the projection, it is valid. Note that $\varphi_3|^N$ coincide with the mapping T. In the similar way it is easy to show that $\varphi_3|^N(P_s(t)) = P(\varphi_3|^N(t))$.

According to the condition 2 of Definition 2, for the distinguished elements we demand equalities $\varphi_1(x^s_0) = x_0$, $\varphi_2(q^s_0) = q_0$. Since by construction $q^s_0 = q'_0$, the later equality is true recalling the definition of φ_2.

For (5) we have $\Omega_{P+} \setminus \Omega_{P^+_-} = \{L_m(\mathcal{G})\}$, $\Omega_{P-} \setminus \Omega_{P^+_-} = \{\overline{L}_c(\mathcal{J}/\mathcal{G}), X_m \times Q_m\}$. To satisfy Definition 2, for the predicate $L_m(\mathcal{G})$ the inclusion $\langle\varphi\rangle^{T[A]}(L_m(\mathcal{G}_s)) \subseteq L_m(\mathcal{G})$ is required. As $L_m(\mathcal{G}_s) \in \Sigma^{I\!N}$ then

$$\langle\varphi\rangle^{T[A]}(L_m(\mathcal{G}_s)) = \langle\varphi\rangle^{A_3{}^{I\!N}}(L_m(\mathcal{G}_s)) = \varphi_3|^N(L_m(\mathcal{G}_s))$$

and the inclusion looks like $\varphi_3|^N(L_m(\mathcal{G}_s)) \subseteq L_m(\mathcal{G})$. Obviously, for any $s \in L(\mathcal{G})$ $T(e(s)) = s$. Since $\varphi_3|^N$ coincide with T, Lemma 1 implies $\varphi_3|^N(L_m(\mathcal{G}_s)) = L_m(\mathcal{G})$ therefore the inclusion is satisfied.

According to the theorem, for the predicate $L(\mathcal{J}/\mathcal{G}) \in \Omega_{P^+_-}$ an equality $\varphi_3|^N(L(\mathcal{J}_s/\mathcal{G}_s)) = L(\mathcal{J}/\mathcal{G})$ should be valid. First we show that $\varphi_3|^N(L(\mathcal{J}_s/\mathcal{G}_s)) \subseteq L(\mathcal{J}/\mathcal{G})$. Denote as $L^s_{(1)}$ all the strings of the language $L(\mathcal{J}_s/\mathcal{G}_s)$ of the length 1. Let $s \in L^s_{(1)}$. By construction of $\mathcal{G}_s$, $s = \sigma_s = y_0 \in Y$. Since $\varphi_3(y_0) = \varepsilon$ then $\varphi_3|^N(L^s_{(1)}) \subseteq L_{(1)}$. Assuming $\varphi_3|^N(L^s_{(i)}) \subseteq L_{(i)}$, $i = 1, 2, \ldots, j$, let $s_s\sigma_s \in L^s_{(j+1)}$. We want to prove that $\varphi_3|^N(s_s\sigma_s) \in L_{(j+1)}$. By (10) it is equivalent to $\varphi_3|^N(s_s) \cdot \varphi_3(\sigma_s) \in L_{(j+1)}$. If $\sigma_s = y$ then $\varphi_3|^N(s_s) \cdot \varepsilon \in L_{(j+1)}$ is immediately satisfied. Suppose $\sigma_s = \sigma$. By induction hypothesis $\varphi_3|^N(s_s) \in L_{(j)}$ so $(x, q) \triangleq (\xi \times \delta_c)(\varphi_3|^N(s_s), x_0, q_0) = (\xi(\varphi_3|^N(s_s), x_0), \delta(\varphi_3|^N(s_s), q_0))$ is defined. Let $(x_s, q_s) \triangleq (\xi_s \times \delta^s_c)(s_s, x^s_0, q^s_0) = (\xi_s(s_s, x^s_0), \delta_s(s_s, q^s_0))$. Since $s_s\sigma_s \in L^s_{(j+1)}$ then $\delta_s(\sigma_s, q_s)$, $\xi_s(\sigma_s, x^s)$ are defined and $\sigma_s = \sigma \in \phi_s(x_s)$. Then $\varphi_2(\delta_s(\sigma_s, q_s)) = \delta(\varphi_3(\sigma_s), \varphi_2(q_s)) = \delta(\sigma, q)$ is also defined. By (7) $\varphi_1(\xi_s(\sigma, x_s)) = \xi(\sigma, \varphi_1(x_s))$ is defined while $\sigma \in \phi(\varphi_1(x_s))$. Therefore $\varphi_3|^N(s_s) \cdot \sigma \in L_{(j+1)}$ and $\varphi_3|^N(s_s\sigma_s) \in L_{(j+1)}$. For reverse inclusion $L(\mathcal{J}/\mathcal{G}) \subseteq \varphi_3|^N(L(\mathcal{J}_s/\mathcal{G}_s))$ it should be shown that $t \in L(\mathcal{J}/\mathcal{G})$ imply $t \in \varphi_3|^N(L(\mathcal{J}_s/\mathcal{G}_s))$, i.e. a string $t_s \in L(\mathcal{J}_s/\mathcal{G}_s)$ exists such that $\varphi_3|^N(t_s) = t$. The desired string is obtained by the operation $e(t)$. Indeed, $\varphi_3|^N(t_s) = \varphi_3|^N(e(t)) = T(e(t)) = t$. Thereby, $\varphi_3|^N(L(\mathcal{J}_s/\mathcal{G}_s)) = L(\mathcal{J}/\mathcal{G})$.

For predicates from the set $\Omega_{P-} \setminus \Omega_{P^+_-} = \{\overline{L}_c(\mathcal{J}/\mathcal{G}), X_m \times Q_m\}$ we have to guarantee inclusions

$$\overline{L}_c(\mathcal{J}/\mathcal{G}) \subseteq \varphi_3|^N(\overline{L}_c(\mathcal{J}_s/\mathcal{G}_s)), \tag{11}$$

$$X_m \times Q_m \subseteq \varphi_1 * \varphi_2(X^s_m \times Q^s_m). \tag{12}$$

Consider (11). Recalling $L_c(\mathcal{J}/\mathcal{G}) = L(\mathcal{J}/\mathcal{G}) \cap L_m(\mathcal{G})$, let $s \in \overline{L}_c(\mathcal{J}/\mathcal{G})$, i.e. there exists $t \in \Sigma^*$ such that $st \in L(\mathcal{J}/\mathcal{G})$ and $st \in \cap L_m(\mathcal{G})$. Since $L(\mathcal{J}/\mathcal{G}) = \varphi_3|^N(L(\mathcal{J}_s/\mathcal{G}_s))$ then $st \in \varphi_3|^N(L(\mathcal{J}_s/\mathcal{G}_s)) = T(L(\mathcal{J}_s/\mathcal{G}_s))$. This implies $e(st) \in e(T(L(\mathcal{J}_s/\mathcal{G}_s))) = L(\mathcal{J}_s/\mathcal{G}_s)$ and also $e(st) \in e(L_m(\mathcal{G}))$. By statement 3 of Lemma 1 $e(L_m(\mathcal{G})) = L_m(\mathcal{G}_s)$. Then $e(st) \in L(\mathcal{J}_s/\mathcal{G}_s)$ and also $e(st) \in L_m(\mathcal{G}_s)$ so $e(st) \in L(\mathcal{J}_s/\mathcal{G}_s) \cap L_m(\mathcal{G}_s) = L_c(\mathcal{J}_s/\mathcal{G}_s)$. Therefore $e(s) \in \overline{L}_c(\mathcal{J}_s/\mathcal{G}_s)$ that is $s \in T(\overline{L}_c(\mathcal{J}_s/\mathcal{G}_s)) = \varphi_3|^N(\overline{L}_c(\mathcal{J}_s/\mathcal{G}_s))$. Thus the desired inclusion is obtained. For (12) notice that $Q'_m = Q_m$ therefore it is enough to have the inclusion $X_m \subseteq \varphi_1(X^s_m)$. Thus, based on Theorem 1 we formulate the following statement:

Proposition 1: Let $\mathcal{J}_s = (\mathcal{S}_s, \phi_s)$ is a non-rejecting supervisor for DES $\mathcal{G}_s$. Suppose that $\varphi_1 : X_s \to X$ is such that conditions (7), (8) are satisfied and $\varphi_1(x^s_0) = x_0$, $X_m \subseteq \varphi_1(X^s_m)$. Then supervisor $\mathcal{J} = (\mathcal{S}, \phi)$ for DES $\mathcal{G}$ is non-rejecting as well.

Note that if, given $\mathcal{S}$, consider $\mathcal{S}_s = (X, \Sigma_s, \xi_s, x_0, X_m)$ with ξ_s defined as $\xi_s(x_s, \sigma_s) = \xi(x_s, \sigma)$ in case of $\sigma_s = \sigma$ and $\xi_s(x_s, \sigma_s) = x_s$ in case of $\sigma_s = y$, then $\varphi_1 : X_s \to X$ is identity map and all conditions of Proposition 1 are satisfied.

Due to lack of space we just mention that all inclusions above may be proved to be equalities. Then another theorem from [10], dealing with LAE (3), may be applied saying that if supervisor $\mathcal{J} = (\mathcal{S}, \phi)$ for DES $\mathcal{G}$ is non-rejecting then supervisor $\mathcal{J}_s = (\mathcal{S}_s, \phi_s)$ is non-rejecting as well. Combining these results we obtain

Theorem 2: Supervisor $\mathcal{J}_s = (\mathcal{S}_s, \phi_s)$ for DES $\mathcal{G}_s$ is non-rejecting iff a supervisor $\mathcal{J} = (\mathcal{S}, \phi)$ for DES $\mathcal{G}$ is non-rejecting.

V. Conclusion

In the framework of supervisory control theory we have studied one of the main properties of a supervisor for a discrete-event system. In particular, we address the issue of preservation of the non-rejecting property of a supervisors for DESs with event and state observation and event observation only. The presented results illustrate the way of the LAE-method application and how the theorems, concerning the classes of the properties which are preserved under the similar type morphisms, allow one to easily obtain preservation criteria, basing just on the structure of the formula and omitting numerous formal manipulations. The LAE-method combines algebraic and logical approaches and may be of considerable interest to the audience since it is suitable for generating the properties preservation conditions for different dynamical systems. For example, in [12] $(max, +)$-algebra model of the public railway transportation network was investigated. The interpretation of the conditions, generated by the LAE-method, allowed to obtain criteria of non-deterioration of the properties of the timetable when new railway paths are added to the network. The only stipulation of the LAE-method applicability is the issue of algebraization of a dynamical system model.

References

[1] N.V. Nagul, "The logic-algebraic equations method in system dynamics", St. Petersburg Math. J., vol. 4(24), pp. 645–662, 2013.

[2] C. Changwei, F. Lin, and Z. Lin, "Why event observation: observability revisited", Discrete Event Dynamic Systems, vol.149, pp. 127–149, 1997.

[3] N. Nagul, "Discrete-event systems with state observation properties studying", Computational Technologies, vol. 20, IV, pp. 167–177, 2015.

[4] P.J. Ramadge, and W.M. Wonham, "Supervisory control of class of discrete event processes", SIAM J. Control and Optimisation, vol. 25(1), pp. 206–230, 1987.

[5] C.H. Golaszewski and P.J. Ramadge, "Control of discrete event processes with forced events", Proc. 26th IEEE Conf. on Decision and Control, pp. 247-251, December 1987.

[6] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen, D.C. Teneketzis, "IEEE Transactions on Control Systems Technology", vol. 4, pp. 105–124, 1996.

[7] F. Lin and W.M. Wonham, "On observability of discrete-event systems", Information Sciences, vol. 44(3), pp. 173–198, 1988.

[8] N. Bourbaki, Theory of Sets. Springer Science & Business Media, 2004.

[9] S. Feferman, "Harmonious Logic: Craig's interpolation theorem and its descendants", Synthese, vol.3(164), pp. 341–357, 2008.

[10] N.V. Nagul, "Classes of properties preserved under morphisms of generalizations of many-sorted algebraic systems in studying dynamics", Trudy Instituta Matematiki i Mekhaniki UrO RAN, vol. 1(20), pp. 185–200, 2014. (in Russian)

[11] R.C. Lyndon, "Properties preserved under homomorphism", Pacific J. Math, vol. 9, pp. 143–154, 1959.

[12] N.V. Nagul, "Timetable properties preserving in one model of public transportation network", The modern technologies. The system analysis. The modelling, vol. 4(28), pp. 150–159, 2010.(in Russian)

An Evaluation Framework and a Brief Survey of Decision Tree Tools

N. Vlahovic
University of Zagreb, Faculty of Economics and Business, Zagreb, Croatia
nvlahovic@efzg.hr

Abstract - Over the last 50 years decision trees have been extensively used in numerous application areas by practitioners and academics alike. Decision trees are both a mathematical representation tool and a highly versatile graphic representation tool suited for modelling and visualisation of structured hierarchical concepts. Decision tree construction algorithms belong to supervised learning algorithms in the field of machine learning. Most prominent applications areas of decision trees are managerial sciences, knowledge engineering and data analysis. Variety of specific decision tree types is used: decision analysis trees, knowledge representation trees, classification and regression trees, decision forests, etc. Software tools for each of these types are available often borrowing from one decision tree concept to another. As a consequence, distinct features and capabilities of these tools make direct comparison difficult. Goal of this paper is to propose a framework for evaluation of available decision tree tools in relation to their foundation model. Brief survey of currently available decision tree tools will be given accordingly. Results may motivate more comprehensible studies of decision tree applications. Results may also motivate development of hybrid decision tree tools where sound combination of decision tree concepts can enhance functionalities and provide additional benefits to users.

I. INTRODUCTION

Communicating critical information and complex concepts may present a challenge that can be successfully overcome using various visualisation and abstraction techniques. Trees are one such technique extensively used over the last 50 years. Concepts such as algorithms and processes can be easily represented by trees. Intuitive structured graphical properties alleviate human interpretation. Vast number of similar software tools that implement various types and concepts of trees and decision trees are available.

In current literature there are a few quite extensive overviews of trees in decision analysis (often referred to as decision trees) [11] and decision tree algorithms in machine learning [16]. Nevertheless, there is a lack of systematic overview of decision tree tools and their functional capabilities. Similar software tools such as simulation tools, web extraction tools, etc., have been subject to evaluation and research resulting in extensive evaluation frameworks and models. The omission of decision tree tools may be because decision tree software is not regarded as a coherent group of tools. There are many different and at times incompatible implementations of decision tree concepts. This work is aimed at bringing these differences together.

Goal of this paper is to determine common characteristics that connect various decision tree concepts and implementations allowing the creation of unified evaluation framework for decision tree software tools. This framework will then be used to analyse currently available software tools that implement trees and decision trees. Finally, the results should indicate current availability of software tools across different application areas. Additionally, decision trees may be combined with other decision support methods in an attempt to provide their users with additional functionalities and benefits.

The paper is organized as follows. In Section II a broader background on decision trees is given. Brief history of implementing trees as a tool in computer science, managerial science, artificial intelligence and cognitive sciences is given. In Section III an evaluation framework of decision tree tools is constructed and proposed. In Section IV an overview of most typical software tools is given according to the established evaluation framework. In the following Section V results of the qualitative analysis are given. Areas that have the most and the least support by the current tools are indicated. Finally, Section VI summarizes contributions and conclusions indicating directions for further research.

II. BACKGROUND

In this section we will describe and define decision trees. A brief historical overview of applications of decision trees relevant to the social studies will follow. Major focus will be on the applications in economics, business and information sciences.

Decision trees are based on mathematical concepts of a tree. More specifically, in graph theory tree is defined as any undirected graph where any two vertices are connected by one path or edge. For the purpose of representing various data structures in computer science specific type of trees are used called rooted trees.

A rooted tree is a tree graph with directed edges. If all the edges are pointed away from the root, that type of directed tree graph is called *arborescence*. If all the edges are pointed towards the root it is called *anti-arborescence*. In any case one vertex of the tree is designated as the root of the tree. It becomes the focal point for determining the orientation of the edges. It is important to note rooted trees can distinguish hierarchy between its vertices. Vertices

can be ordered in layers depending on how close to the root vertex they are. Parent node is a vertex that is directly connected to vertices with higher partial ordering. Child nodes are vertices directly connected to a vertex with lower partial ordering of a tree.

All of described features supplemented with additional structure information make tree models suitable for describing hierarchical structures in various areas of interest. From mathematics and logic during late 1950s and 1960s to computer science later on, trees were commonly used to represent hierarchical data, compile complex data structures for more efficient search, describe various algorithms, etc. In 1968 Howard Raffa formulated decision making problems in terms of trees, commonly known as decision trees [13, 11]. With the development of operations research, artificial intelligence, machine learning and decision making, trees were the basis of the development of methods, approaches and techniques for modelling and solving many types of problems.

Application of decision trees in decision analysis is based on the representation of decision options and their effects on the solution of the problem at hand. Key algorithm used in evaluating various options is a type of rollback algorithm. It determines the optimal sequence of decisions and events described by the decision tree. Rollback algorithm can be considered a form of backward chaining inference that uses numerical data types to determine the best solution [10]. This type of decision trees are used in strategic management and other strategic analyses such as strategy analysis in politics, military, etc.

Application of decision trees in intelligent systems pertains to modelling knowledge and inference mechanisms. Decision trees were first used as a technique to represent a set of rules that describe knowledge of expert systems. Decision trees are easily translated into rules that are basic knowledge representation approach in intelligent systems [14]. Each tree can be transformed into a set of rules describing it either horizontally or vertically. Decision trees are also used to represent inference mechanisms based on search algorithms [14]. Two main types of algorithms are used: forward chaining and backward chaining. Forward chaining algorithm uses arborescence structures that follow the inductive logic of the reasoning process. Inference starts of from known facts describing problem situation and moves toward conclusions that may represent problem solution. Backward chaining mechanism can be represented by anti-arborescence structure where hypothesis is being challenged in deductive manner. Goal of the inference is to provide known facts as arguments to accept or reject a given hypothesis [5]. Further developments in the field of intelligent systems have encouraged additional enhancement of the decision tree concepts, such as representing complex rules using AND/OR trees, or representing uncertain or incomplete knowledge.

Application of decision trees in data analysis and data mining is highly intertwined with the application of decision trees in artificial intelligence. Machine learning first used decision trees as a type of inductive supervised learning algorithm. It is a predictive model that uses known observations to make conclusions and learn about target items values.

Some of the first algorithms in data analysis used for the construction of decision trees relied on statistical properties of data. Decision trees in data analysis are used for categorization of data where dependent variable assumes one of predefined categorical values [8]. Decision trees are also used in prediction if the dependent variable is a continuous value defined by its mean and standard deviation. Both of these two types of decision tree models are built based on an algorithm that uses statistical properties of underlying data [2]. This is why they are commonly known as classification and regression trees (C&RT). C&RTs use different statistical metrics to determine branching nodes and split values for each tree. Some of the most commonly used metrics are Chi-square, Gini coefficient and variance reduction.

In other algorithms for construction of decision trees statistical tests that determine each node in decision trees for data analysis are replaced by other measures. Most commonly used measures are entropy and information gain [12]. Some of the most common algorithms used include iterative dichotomizer 3 (ID3), C4.5 and C5.

We can see that there are some similarities between various types of decision trees. Combining properties of different decision tree approaches can produce new "hybrid" types of decision tree implementations. Most obvious example is the connection between classification trees and knowledge representation using rules with a clear connection between data mining and intelligent systems. Common premise that allows valid combination of these two decision tree concepts is the induction principle of case based reasoning. Other combinations may also be feasible.

III. EVALUATION FRAMEWORK FOR CHARACTERIZING DECISION TREE TOOLS

In this section we will analyse the taxonomy of decision trees in terms of software applications and tools created for various types of decision tree implementations. Based on the characteristics and properties of methods and software implementations we will create and propose a conceptual evaluation model for decision tree software tools.

Reference [9] describes a partial taxonomy of artificial intelligence tools. For the purpose of this paper it may serve only as a framework example. This taxonomy is too generally defined for the purpose of this paper as it described tools for artificial intelligence in general. Additionally, managerial aspects of artificial intelligence tools are not adequately represented. In current literature there is a lack of described evaluation models dedicated to decision tree tools. However, there are several methodologies for the evaluation of similar software tools such as simulation tools [1], [6] and knowledge mapping tools [3], [4]. These examples were taken into account during the preparation of the proposed evaluation model for decision tree tools.

Taxonomy of decision trees is the basis for establishing the appropriate model for grouping various

software tools that implement decision trees. We will focus on three major groups of tools that provide decision tree support: (1) tools dedicated to decision analysis, (2) tools for knowledge representation and inference (3) tools for data analysis and data mining.

Various important criteria for the evaluation have to be grouped. Beside the core indicators describing the implementation of decision trees, we propose using the group of criteria that describe general software tool properties and group of criteria that focuses on the application of the tools. Additional group that describes various user benefits and perks is also included. In table 1 we can see a summarized structure of the proposed evaluation model. For the purpose of this paper only the summarized grouping and structure is presented at the second level. Individual criteria within each group and their values are not given due to the length of this paper.

TABLE I. TAXONOMY OVERVIEW

Criterion	Description
T1 Software and hardware considerations	Considerations pertaining to the software architecture of the tool itself.
T1.1 Software type	- desktop application, web application or cloud service or add-in extension or mobile app.
T1.2 Coding interface	- text input, visual imput, text-based or visual interface or combination.
T1.3 Software compatibility and extendability	- OS compatibility, API avaiability, etc.
T1.4 Hardware requirements	- technical features of hardware and special requirements.
T1.5 Portable devices support	- what type of portable devices are supported
T1.6 Ease of use	- user interface features and *user-friendly*-ness level
T1.7 Financial considerations	- availability of free-to-try versions, overall cost of the solution, availability of multiple-scale versions
T1.8 Documentation & Customer support	- availability of user documentation, education, user support, local or online, etc…
T2 Decision tree capabilities	Describes particular implementation of the decision tree within the tool, as well as the scope its functionalities
T2.1 Interpretation and visualization	- level of visualisation support, node alignement check and other user-friendly interface features
T2.2 Decision analysis	- decision analysis support and its extent
T2.3 Data mining	- data mining support and its extent
T2.4 Knowledge representation	- knowledge representation and its extent
T2.5 Inferance capabilities	- inference modelling and its extent
T2.6 Tree forests	- forrest support amd its extent
T2.7 Hybridisation level	- based on T2.2-T2.6 level of incorporating different decision tree concept types
T2.8 Reproducability	- saving and re-using features of developed models
T2.9 Reporting and distribution	- generating reports and distribution of tree models and their definitions
T3 Applicability	Describes intended usage of the software tool:
T3.1 Problem domain	- general purpose or specific problem area
T3.2 Model creation	- enables creation of decision model
T3.3 Model export	- enables export of created model in multiple formats for further distribution
T3.4 Executable model	- creates an executable model that can be run on demand or in real-time
T3.5 Other decision support methods incorporated	- if decsion tree implementation supports other methods for analysis, visualisation, representation or inference.
T3.6 Collaboration	- collaborative functions for group decision making
T4 Provider-user relationship	Describes the providers' approach of offering their services and tools
T4.1 Consultative services	- provider of the tool offers consultative services based on the tool or not
T4.2 User invovlment level	- provider to prepare or create model or application used by user or users themselves use the tool
T4.3 Academy and education	- level of openness to academic institutions and (in)formal education

As we can see in table 1, the first group of criteria pertains to software and hardware considerations. This group of criteria evaluates various compatibility issues concerning the technical characteristics of the hardware required to use the software tool as well as software requirements. Availability of documentation and software access level connected with the financial costs is also an important indicator in this group.

Core group of indicators is concerned with decision tree implementation within the software tools. As we can see in Table 1 beside the decision tree types and level of combining different types, important criteria to assess pertains to interpretation capabilities, as well as the management of created tree models.

Next is the group of criteria that evaluates the convenience of the tool to different practical applications. Here the most important aspect is the intended use of the tool determined by the intended problem domain, and sustainability of modelling. Another important aspect is the level of possible adaptation to wider area of usage either by combining decision tree with other decision support methods or by providing collaboration capabilities to support group decision making.

Final group of the proposed evaluation model examines the provider-user relationship which may be important for maintaining high quality customer relations. Here consultative function of the provider and openness to formal and informal learning by their customers is assesses. Important aspect of this relationship is the level of allowed involvement of users with the provider's tool. Level of user involvement can vary in range from only using final results of the tool over using the tool themselves to using the source code to modify the tool.

IV. OVERVIEW OF DECISSION TREE TOOLS

In this section we overview decision tree tools we studied. The list of tools presented in this Section should not be regarded as a complete list of available tools as this is a brief study that aims at covering the most representative tools. The overview of decision tree tools will follow the evaluation model presented in Section III.

A. Visualisation tools

One of the most prominent features of trees is their easy and intuitive interpretation through visual representation. Abstract concepts and ideas originating in mathematics, logic, and more specifically in other areas such as management, business and production, benefit from visualizing the processes and concepts they deal with. Hierarchically structured nature of trees was directly implemented in order to facilitate complex process in determining and explaining structures and process in various application areas. There is a group of tools that focus on the creation of visual depictions of decision trees rather than utilizing any other features. With the development of graphical capabilities of computer systems number of software tools for drawing decision trees increased to the level that most of application software have some sort of drawing capability. Dedicated software tools for drawing are increasingly becoming obsolete. Most available solutions incorporate various

diagrams, charts and graphic presentations as well as standards in order to provide their users with added value. User interfaces of most of them help with the alignment of graphic constructs (Microsoft Visio). There are tools that allow further customization of drawing elements and provide compliance to standards for particular diagrams in various degrees (eDraw Max). Some software tools provide extensive capabilities for collaborative efforts such as Lucidchart.

This group of decision tree tools mostly contains more general purpose drawing tools that are not dedicated to decision trees so they do not allow any additional functionality regarding different decision tree concepts.

B. Decision analysis tools

In terms of complexity decision analysis tools provide more focus on decision trees as they implement some basic calculations that can be performed over decision tree models. These tools are mostly focused on visualization of mental models of decision making procedures and their outcomes, as they implement rollback algorithm. Many of these software tools are not stand-alone applications but serve as add-ins for standards office applications suites, mostly Excel. There is a good selection of commercial add-ins such as TreePlan or Palisades Precision tree, while there are also open-source solutions such as SimpleDecision.

An example of standalone dedicated software tool for decision analysis is SilverDecisions mostly intended for business decision making while there are other tools that take a more general view of strategic analysis such Gambit 15 that incorporates game theory concepts in the analysis.

For this group of software tools we can conclude that connection to other decision tree concepts is very limited as well as the connection to other decision support methods, even though there is a valid indication that the inclusion of other concepts is a viable option.

C. Data analysis and data mining tools

This is probably the biggest group of software tools for decision trees currently available to practitioners and academics. Main reason is that well established statistical software includes decision trees as a standard predictive method. Examples are well known statistical kits Statsoft's Statistica, IBM SPSS Modeler, etc. Due to high demand and diversity of available algorithms additional decision tree splitting and branching algorithms have been implemented in form of add-ins or special versions of branded software. For example, Statsoft established rather successful version Statistica Dataminer. On the other hand some providers made their focus to create statistical software suites specifically dedicated to data mining, such as RapidMiner, Weka, Dtreg, KNIME Analytics Platform or even Angoss KnowledgeStudio. These dedicated data mining tools support C&RT trees and CHAID trees. They also support improved versions of information-gain-based algorithms such as C4.5 or even C5 which were not that commonly used by mainstream statisticians. Most of the statistical suites also support random forests as standard method for data analysis.

Due to high competition in data analysis software market a few innovative approaches are also available. Some of them combine soft computing methods such as genetic algorithms (for example GAtree) for construction of decision trees. There are also a few software tools in development and research which try to offer alternatives such as SMILES. SMILES introduces logic concepts form artificial intelligence and cognitive disciplines by implementing AND/OR trees instead of more widely available sample forests.

Overall this group of decision tree software tools is the least focused on the decision trees and mostly take wider approach and focus as statistical software. At the same time due to high number of available tools there are lot of options for users in terms of versions, education opportunities and support provided to practitioners. Finally, unlike decision analysis tools this group of software tools incorporates more concepts that borrow from different decision tree concepts and other decision support methods.

D. Knowledge management tools

Representing knowledge using decision trees presented an alternative to expressing knowledge in sets of rules. Decision trees have contributed to knowledge representation by allowing easy verification and validation of rule sets, as well as detecting errors and inconsistencies in knowledge. Furthermore, decision trees can also be used to visualize inference processes and promote important and complex concepts in intelligent systems and machine learning to practitioners. Most of the software providers emphasize that using this type of software tools empowers users and establishes better knowledge management practices.

Two of the most representative software tools that use decision trees to represent and model knowledge as well as inference processes are Attar Xpertrule and Doctus. Knowledge representation is expressed relying heavily on the equivalence of rule sets and decision trees. They support both forward chaining and backward chaining concepts implementing them in their own unique right. Doctus creates operating knowledge base that can be consulted to solve new problem situations and record the case history that can later be used to data mine for hidden knowledge. Xpertrule allows for the creation of standalone knowledge based or expert systems that can be easily distributed and re-used. Both software systems can be used for group decision making. DoctuS incorporates some of the approaches of data analysis. Xpertrule incorporates and benefits from other soft computing decision support method such as genetic algorithms.

Less general software tool in this category is definitely Zingtree. It is a specialized tool for the creation of customer support trouble-shooters that also serves as knowledge management system with high focus on collaborative functionalities.

RuleQuest See5 is declared as a data mining tool but it can be used to create source code that captures inference processes in order to embed their logic and mechanisms in other user software systems and information systems. Final application is based on decision trees and rule sets.

There are a number of proposed systems and concepts in this group of software tools. Some have only reached research stage before they were discontinued. One such example is WebShell that intended to use potential of Web and combine it with decision tree models of knowledge [15]. Other software tools in this group are currently in research stage, such as RAMOLI. It is knowledge and inference shell [7] that also supports decision tree concepts.

Overall this group of software tools for decision trees does not offer many general development tools, but is more focused on particular problem domains (such as customer support). There are many examples of specific application areas that are the results of knowledge modelling using decision tree tools. However, user involvement is restricted to project deliverables instead of using the tool themselves.

E. Hybrid decision tree tools

Final group of tools supports at least two different concepts of decision trees, where neither of the involved concepts is in the focus of the tool. In previous groups we encountered several decision tree tools that also recognize multiple decision tree concepts, but the focus is in favour to one particular decision tree type. In this group there is not such focus, since they try to combine various decision support methods to deliver better value to users and greater efficiency to their customers.

A lot of tools in this group are intended for educational purposes while some of them are used by academics and consultant services. There are two such examples. First is YaDT (abr. Yet another Decision Tree) that draws heavily on data mining and machine learning to create self-sufficient knowledge representations. Second is a group of tools gathered around a brand AIspace that presents and trains its users in inference principles (Graph Search) and data mining (Decision Trees).

Rulex is also an example of hybrid tool that supports creating data mining and knowledge representation models relying on the decision tree visualization capability.

An extreme example of applying decision tree concepts in management and problem solving are various applications presented by Mindtools. Instead of software tools, conceptual tools are explained and offered to potential users. These concepts are characterized with the highest level of combining various decision tree concepts either from decision analysis, data analysis or knowledge and inference models. There are concepts that borrow from all three concepts at the same time.

V. QUALITATIVE ANALYSIS

In this section we analyse how presented decision tree tools corresponded to main features of decision trees as described earlier. We will address following features: applicability of the software tools, software versions available, level of user involvement, user interface and customer care. Specifically, we will consider the amalgamation of some of the described features from various decision tree types and how successful and how useful the combinations of features are. Finally we will consider the potential combination of decision tree tools that are not currently available but may prove to be useful for future applications in various fields of interest.

A. Applicability critera

Software tools for decision trees that we have investigate in this survey mostly pertain to specific problem domains. A few general purpose tools primarily pertain to knowledge and inference representation tools. There are also hybrid tools that combine knowledge representation concept with other decision support methods. Almost all of the tools allow creation of decision tree model that can be reviewed, exported or printed. Some of the tools do not allow reusability of developed models, especially tools dedicated to decision analysis. Smallest portion of tools provides users with an executable version of the created model. There is deficiency of tools that feature collaborative facilities for group decision making. Space for improvements in offering additional methods that can be combined with decision trees, beside statistic methods, is indicated.

Overall availability of software tools for all decision tree concepts is well covered, but innovative approaches still present an opportunity for new tools, especially for specific problem domains that would greatly benefit from these tools.

B. Versioning of software tools

As many other software tools, most of the commercially available tools have different bundles depending on the application area, intention of use or intended population. Academic versions are also available, especially in the segment of data analysis and data mining. Open source tools and unstable research versions are also available especially in hybrid decision tree groups as well as in the knowledge and inference representation groups. Best pricing schemes are present among data analysis and data mining tools due to their greater market presence and higher competition. More expensive solution tend to offer readymade models and applications for specific problem domain. With this approach they try to alleviate cost for end users while retaining tool capabilities for development purposes of the provider and software owner.

C. User considerations

User retention and stabile user foundation is a key to sustaining any software products. Decision trees intrinsically appeal to users due to their intuitive interpretation and visualization properties that enable users to quickly understand and use complicated logical concepts. This is why high investment in intuitive user interface is crucial. All of the investigated tools seem to understand the importance of user interface and the importance of ease of use. Most of them provide high quality interfaces that usually slowly deteriorate with the increasing complexity of the model. Open source and educational software still seem to be overly simplified. Software packages such as statistical software, that support bigger number of functionalities have confusing and less intuitive interface. Software tools that are

specifically designed to support particular type of decision trees seem to have best user interface. More complex software suites provide good user documentations, educative materials, tutorials and video tutorials, instead.

Specific tools with innovative approach to decision trees and tools that combine knowledge and inference representation with other decision support methods benefit from decision trees directly. These tools seem to use decision trees as a facilitating model for better understanding and using somewhat complex and more challenging concepts by average users and business professionals. For instance, concepts originating in cognitive sciences, artificial intelligence or soft computing are easier to explain and use by non-technical users.

D. Decision tree innovations

During the survey of decision tree tools great attention was dedicated to understanding what decision tree concepts have been used and what concepts were combined to provide additional benefits to their intended users. The most common was connection between knowledge representation and classification roles of decision trees. In terms of logic, knowledge representation adheres to the deduction principle while classification adheres in the opposite logic process of induction.

There are also a significant number of software tools that try to combine decision trees with other decision support methods. The most common ones are genetic algorithms, neural networks and fuzzy logic.

However there are additional feasible combinations that are not covered by available decision tree tools. Combination of knowledge representation and data mining trees towards decision analysis trees is not available. We find this omission unjustifiable since rollback algorithm can be relate to both backward chaining inference of intelligent systems and induction based on entropy in machine learning for data mining.

VI. CONCLUSION

In this paper we discussed software tools dedicated to supporting decision trees. Decision trees are both intuitive visual tool and supervised learning algorithm applicable to various types of problems. The original trees have been enriched with powerful concepts and algorithms that have been developed and used in the last 50 years. As numerous software tools are available to practitioners and academics the need to systematically analyse their capabilities is required.

Main contribution of this paper is the proposed evaluation model of decision tree software tools. It was used as a framework for the brief survey of currently available decision tree software tools.

Results have reviled that some of the decision tree concepts have been covered and combined by a number of software tools, especially those concepts that have originated from the same problem domain. Knowledge representation and data mining are one confirmed example.

Managerial use of decision trees for decision analysis has been least challenged and explored as a source of innovative implementations that combine this approach to other decision tree concepts.

We believe and speculate that this innovation is feasible. Initial argument is the rollback algorithm used in decision analysis that can be defined as backward induction that binds this decision tree concept to knowledge and inference representation implementations as well as data mining and machine learning implementations.

REFERENCES

[1] V. Bosilj-Vuksic, V. Ceric, V. Hlupic, “Criteria for the evaluation of business process simulation tools”. Interdisciplinary Journal of Information, Knowledge, and Management, vol. 2, 2007, pp. 73–88.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth International Group, 1984.

[3] R. Dehghani R. Ramsin, “Methodologies for developing knowledge management systems: an evaluation framework”, Journal of Knowledge Management, Vol. 19, Issue 4, 2015, pp. 682 – 710.

[4] S. Driessen, W.-O. Huijsen, M. Grootveld, “A framework for evaluating knowledge-mapping tools”, Journal of Knowledge Management, Vol. 11, Issue 2, 2007, pp. 109 – 117.

[5] J. C. Giarratano, G. D. Riley, Expert Systems: Principles and Programming, 4th Edition, Course Technology, 2004.

[6] M. Jadric, M. Cukusic, A. Bralic, “Comparison of discrete event simulation tools in an academic environment”, Croatian Operational Research Review, vol. 5, 2014, pp. 203–219.

[7] S.B.H. Kacem, A. Borgi, M. Tagina, , "RAMOLI: A generic knowledge-based systems shell for symbolic data," in Computer and Information Technology (WCCIT), 2013 World Congress, 2013, pp.1-6.

[8] C. Kingsford, S. L Salzberg, “What are decision trees?”, Nature Biotechnology, vol. 26, Issue 9, 2008, pp. 1011–1013.

[9] J. K. C. Kingston, “Knowledge based system development tools”; Encyclopedia of Life Support Systems”, in Technology, Information and Systems Management Resources, chapter of Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO, Eolss Publishers, Paris, France, 2002-2015.

[10] N. Marfinean, J. Sirbu, D. Racovitan, “Decision Trees – A Perspective of Electronic Decisional Support”, Annales Universitatis Apulensis Series Oeconomica, vol. 12, no. 2, 2010, pp.631–637.

[11] L. Buchanan, A. O'Connell, “A Brief History of Decision Making”, Harward Business Review, vol. 84, no. 1, 2006, pp. 32 – 41.

[12] J. R. Quinlan, “Induction of decision trees”, Machine Learning, Vol. 1, no. 1, 1986, pp. 81 – 106.

[13] H. Raffa, “Decision Analysis: A Personal Account Of How It Got Started And Evolved”, Operational Research, vol. 50, Issue 1, 2002, pp. 179–185.

[14] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 3rd edition, Pearson, 2009.

[15] A. Stranieri, J. Zeleznikow, “WebShell: The development of web based expert systems”, Research and Development in Intelligent Systems XVIII, Proceedings of ES2001, the Twenty-first SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence, Cambridge, December 2001, Springer, London, 2002, pp. 245-258.

[16] S. B. Kotsiantis, “Decision trees: a recent overview”. Artificial Intelligence Review, vol. 39, no. 4, Springer, 2013, pp. 261-283.

Positive Constructed Formulas Preprocessing for Automatic Deduction

Evgeny Cherkashin†, Artem Davydov* and Alexander Larionov*
*Institute for System Dynamics and Control Theory at Siberian Branch of Russian Academy of Sciences
134, Lermontov str., Irkutsk, Russia
Emails: artem@icc.ru,bootfrost@zoho.com
†Irkutsk National Research Technical University, 83, Lermontov str., Irkutsk, Russia
Email: eugene@irnok.net

***Abstract*—The preprocessing of positively–constructed formulas (PCFs) for automatic deduction search algorithms is considered in this paper. A new efficient algorithm for conversion of the predicate calculus language formulas to the language of PCFs and equivalent rules of reducing PCFs, preserving the original heuristic structure of knowledge represented by the formulas of the predicate calculus language, are presented.**

I. Introduction

The calculus of positively constructed formulas (PCF), originally [1], [2], was developed by Russian scientists S.N. Vassilyev and A.K. Zherlov by an evolutionary way in describing and solving control theory (CT) problems. The PCF calculus is presented in [2] as first-order logical formalism (further development presented in [3]), providing examples of CT problems described and solved by the PCF calculus (elevator group control, mobile robot action planning and telescope guidance), as well as proof of soundness and completeness.

The PCF calculus is both machine-oriented, and also human-oriented, naturally aimed at solving the problems of dynamic systems control thanks to its features such as follows: unique inference rule and simple scheme of axioms; modifyability of semantics (constructive, monotonic, temporal, etc.) and besides it is possible to construct intuitionistic inferences of some non-Horn formulas; the explicit usage of ∀– and ∃–quantifiers, the scolemization procedure is not required.

However, in this paper, we will not cover the human–orientation properties of the PCF calculus and its dynamic systems control properties as well. This and the calculus capabilities in action planning is described in [2]. The problems of an automatic theorem proving software (provers) design and implementation are considered briefly in [4].

We describe the language of PCF's, conversion of first–order predicate calculus formulas (FOF) to a PCF's form and the preprocessing of PCFs for automatic deduction search algorithms. The conversion is not a simple problem. The resulting PCF's must obey the main qualitative criteria that the conversion should preserve the heuristic structure of the original formulas. The heuristic structure reflects the properties of the problems being modelled with the logical calculus. The preservation of the structure during conversion allows one to pass the problem properties to the inference engine, resulting in performance of the inference search improving by application of the so called domain–spec strategies. The proposed new algorithm of conversion of FOF into PCF's improve the heuristic preservation quality as compared to the our early used [4] and developed by A. K. Zherlov and E. A. Cherkashin.

II. Preliminaries

Let us consider the basic ideas behind PCF's and its calculi. The language of PCF is a restricted variant of the language of first-order logic (FOL), which consists of first–order formulas (FOFs) built out of atomic formulas with $\&, \vee, \neg, \rightarrow, \leftrightarrow$ operators, $\forall$ and $\exists$ quantifier symbols and constants **True** and **False**. The concepts of *term*, *atom*, *literal* we define in the usual way.

Let $X = \{x_1, \ldots, x_k\}$ be a set of variables, $A = \{A_1, \ldots, A_m\}$ be a set of atomic formulas, and $F = \{F_1, \ldots, F_n\}$ be a set of subformulas. Then the following formulas $((\forall x_1)\ldots(\forall x_k)(A_1\&\ldots\&A_m \rightarrow (F_1 \vee \ldots \vee F_n)))$ and $((\exists x_1)\ldots(\exists x_k)(A_1\&\ldots\&A_m\&(F_1\&\ldots\&F_n)))$ are denoted as $\forall_X A\colon F$ and $\exists_X A\colon F$ respectively, keeping in mind that the ∀–quantifier corresponds to $\rightarrow F^{\vee}$, where $F^{\vee}$ means disjunction of all subformulas from F, and ∃–quantifier corresponds to $\&F^{\&}$, where $F^{\&}$ means conjunctions of all subformulas from F.

If $F = \varnothing$, then the formulas have the form $\forall_X A\colon \varnothing \equiv \forall_X A \rightarrow$ **False** and $\exists_X A\colon \varnothing \equiv \exists_X A\&$**True**, since the empty disjunction is identical to $false$, whereas the empty conjunction is identical to **True**. The form $\forall_X A$ and $\exists_X A$ are abbreviations of such formulas. If $X = \varnothing$, then $\forall A\colon F$ and $\exists A\colon F$ are analogous abbreviations.

The set of atoms A is called *conjunct*. The empty conjunct is identical to **True** as it was already mentioned.

Variables from X are bound by corresponding quantifiers and called ∀–variables and ∃–variables, respectively. In $\forall_X A$, a variable from X that does not appear in conjunct A is called *unconfined* variable.

$\forall\varnothing \equiv \forall\varnothing\colon \varnothing \equiv \forall$**True** $\rightarrow$ **False** $\equiv$ **False**

Construction $\forall_X A$ and $\exists_X A$ are called positive *type quantifiers* (TQ), because A is a conjunction of only positive atoms and referred to as also as *type condition* for X. In practice, this constructions denote phrases such as follows: "for all X satisfying A there is..." or "there exists X satisfying property

A such that..."; "for all integer x, y, z and $n > 2$ there is $x^n + y^n \neq z^n$"".

Originally, the term "type quantifier" was introduced by N. Bourbaki [5] as part of notation for formalization of mathematics. But type quantifiers are stable in languages of another applied fields.

A. PCF Language Explicit Definition

Definition 1 (Positively–constructed formulas (PCF)). *Let, X be a set of variables, and A be a conjunct.*

1) *$\exists_X A$ and $\forall_X A$ are $\exists$–PCF and $\forall$–PCF respectively.*
2) *If $F = \{F_1, \ldots, F_n\}$ are $\forall$–PCF, then $\exists_X A\colon F$ is a $\exists$–PCF.*
3) *If $F = \{F_1, \ldots, F_n\}$ are $\exists$–PCF, then $\forall_X A\colon F$ is a $\forall$–PCF.*
4) *Any $\exists$–PCF or $\forall$–PCF is a PCF.*

This form of logical formulas is referred to as positively constructed formulas (PCFs), as they are written with only positive type quantifiers. The formulas contain no explicit logic negation sign. Any FOF can be represented as PCF [2].

PCF starting from $\forall\varnothing$ is called PCF in a *canonical form*. Any PCF can be represented in the canonical form. Let F is a non–canonical $\exists$–PCF, then the canonical PCF $\forall\varnothing\colon F \equiv$ **True** $\to F \equiv F$. If a PCF F is a non–canonical $\forall$–PCF, then the canonical PCF $\forall\varnothing\colon \{\exists\varnothing\colon F\} \equiv$ **True** $\to \{$**True**$\&F\} \equiv F$. Type quantifiers $\forall\varnothing$ and $\exists\varnothing$ are called *fictitious*, since they do not influence truth value of an original PCF and do not bind any variables. Thy are used to regularize PCFs to a canonical ones.

The PCFs are usually represented as trees for more ease reading, i.e. $Q_X A\colon \{F_1, \ldots, F_n\}$ is represented as

$$Q_X A \left\langle \begin{array}{l} F_1, \\ \cdots \\ F_n; \end{array} \right.$$

where Q is a quantifier. Tree elements have conventional names: *node*, *root*, *leaf*, *branch*, etc. As the quantifiers $\forall$ correspond to disjunctions of formulas $\{F_1, \ldots, F_n\}$ (quantifiers $\exists$ correspond to conjunctions), then each $\forall$–node is considered to be a *disjunctive branching*, and each $\exists$–node is correspond to be a *conjunctive branching*.

Some parts of canonical PCF are denoted as follows:

1) PCF root $\forall\varnothing$ is called PCF *root*.
2) Each PCF root child $\exists_X A$ is called PCF *base*, conjunct A is called *base of facts*, and PCF rooted from base is called *base subformula*.
3) PCF base children $\forall_Y B$ are called *questions* to its parent base. If a question is a leaf of a tree then it is called *goal question*.
4) Subtrees of questions are called *consequents*. If a question has no consequent then the question is referred to as *goal question*, and it is identical to **False**.

Example 1. *Let us consider a PCF representation of a FOF.*

$$F = \neg\big(\forall x \exists y P(x, y) \to \exists z P(z, z)\big).$$

An image F' of F in the PCF language is $F' = \forall\colon \varnothing\{\exists\colon \varnothing\{\forall x\colon \varnothing\{\exists y\colon P(x, y)\}, \forall z\colon P(z, z)\{\exists\colon \boldsymbol{False}\}\}\}$. The tree–like form of the latter is as follows:

$$\forall\colon \varnothing \text{ - } \exists\colon \varnothing \left\langle \begin{array}{l} \forall x\colon \varnothing \text{ ——— } \exists y\colon P(x, y) \\ \forall z\colon P(z, z) \text{ – } \exists\colon \boldsymbol{False} \end{array} \right.$$

III. The transformation algorithm

The algorithm's input is a FOF. At the first stage, the algorithm performs some general transformations:

- elimination of $\leftrightarrow$ and $\to$ logical connections according to well known equivalent transformations $F_1 \leftrightarrow F_2 = (F_1 \to F_2)\&(F_2 \to F_1)$ and $F_1 \to F_2 = \neg F_1 \vee F_2$;
- application of DeMorgan laws $\neg(F_1\& \ldots \&F_n) = \neg F_1 \vee \ldots \vee \neg F_n$, $\neg(F_1 \vee \ldots \vee F_n) = \neg F_1\& \ldots \&\neg F_n$, $\neg\forall x F = \exists x \neg F$, $\neg\exists x F = \forall x \neg F$ and elimination of consequent negations $\neg\neg F = F$; these transformations shift connections $\neg$ to leaves of the FOF (atoms).

The above mentioned transformations will break the original heuristic structures, which was given in the FOF, however, reduction rules described below will reconstruct some of the heuristics to make resulting PCF look similar to the original FOF. Elimination of the connections $\leftrightarrow$ and $\to$ allows us to define the conversion algorithm laconically as a tree–like FOF syntactic representation transformations.

Let F be a FOF processed according above mentioned transformations. A *syntactic tree* of a FOF F is a tree whose root is either

1) the formula F itself, if F a literal (an atom or an atom negation).
2) a quantifier $\forall x$ or $\exists x$, if F is of a form $\forall x G$ or $\exists x G$ correspondingly; syntactic tree of G is the only subtree of syntactic tree of F.
3) a logical connection $\circ$, if F is of a form $G_1 \circ \ldots \circ G_n$, where $\circ$ is connection $\&$ or $\vee$, all subtrees of this node are syntactic trees of FOF's $G_i, \forall i = \overline{1, n}$ correspondingly. The connection $\&$ in a FOF as usual has higher priority than $\vee$, the brackets are used to adjust the priorities (as usual); e.g., if a FOF $\neg A\&B \vee C$, there are two subformulas $G_1 = \neg A\&B$ and $G_2 = C$.

The constructed syntactic tree has literals as a leafs. For each node N of three T, G_N denotes a subformula of FOF F, corresponding to tree T' rooted at node N. Notation $N'_i, \forall i = \overline{1, n}$ denotes the children nodes of the node N.

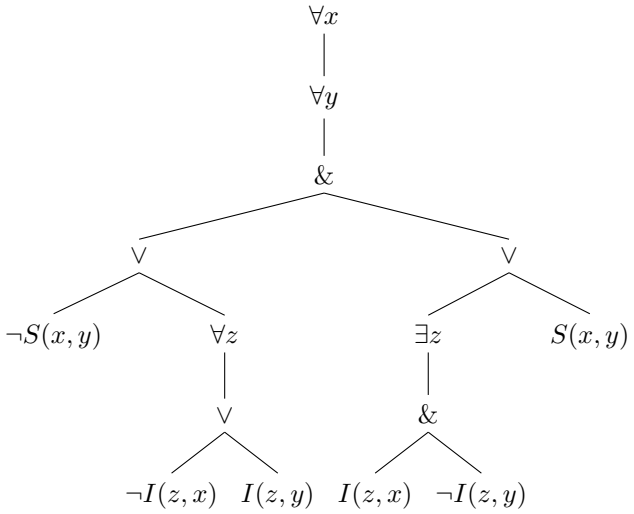
Consider an example of the FOF transformation at the described preparation stage for the formula

$$F = \forall x \forall y (S(x, y) \leftrightarrow \forall z (I(z, x) \to I(z, y))).$$

The sequence of elimination of the connections $\leftrightarrow$ and $\to$ is as follows:

$$\begin{array}{l} F = \forall x \forall y (F_1 \& F_2); \\ F_1 = \neg S(x, y) \vee \forall z (\neg I(z, x) \vee I(z, y)); \\ F_2 = \neg\forall z (\neg I(z, x) \vee I(z, y)) \vee S(x, y) = \\ \exists z (I(z, x) \& \neg I(z, y)) \vee S(x, y). \end{array}$$

The resulting syntactic tree for F is



Let us introduce another denotation. Let $P, Q \in \{\forall, \exists\}, P \neq Q$, and A be a conjunct, then

$$F^Q = \begin{cases} F, \text{if } F = Q_X A\colon \{F_1, \ldots, F_n\}, \\ Q\varnothing\colon F, \text{if } F = P_X A\colon \{F_1, \ldots, F_n\}, \end{cases}$$

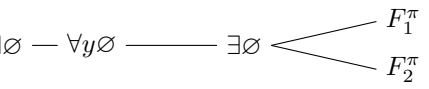
In the following we present the algorithm of conversion of FOF formulas to PCF ones. The above denotation is used in the algorithm. The algorithm obviously terminates in a finite number of steps that is equal to the number of nodes of the syntactic tree of the input FOF F. It is correct, i.e., the output is a PCF image of F, as on the each step the input syntactic tree node is converted into a PCF form.

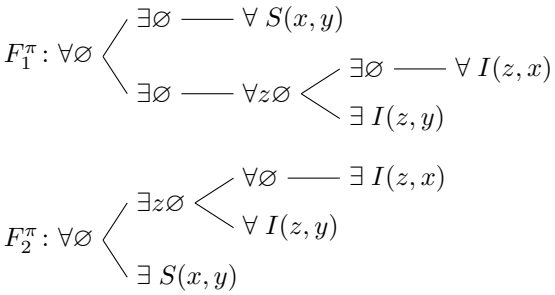
Algorithm 1 FOF to PCF conversion algoritm

Input: Node N is the root of a tree T for F.
Output: F^π is a PCF image of FOF F.

if $N = Qx$ **then** { $P \neq Q$; Node N has $k > 0$ childs }
  return $F^\pi = Qx\varnothing\colon \{(G^\pi_{N'_1})^P, \ldots, (G^\pi_{N'_k})^P\}$
end if
if $N = \vee$ **then** { Node N has $k > 0$ childs }
  return $F^\pi = \forall\varnothing\colon \{(G^\pi_{N'_1})^\exists, \ldots, (G^\pi_{N'_k})^\exists\}$
end if
if $N = \&$ **then** { Node N has $k > 0$ childs }
  return $F^\pi = \exists\varnothing\colon \{(G^\pi_{N'_1})^\forall, \ldots, (G^\pi_{N'_k})^\forall\}$
end if
if $N = R$ **then** {R is an atom}
  return $F^\pi = \exists R$
end if
if $N = \neg R$ **then** {R is an atom}
  return $F^\pi = \forall R$
end if

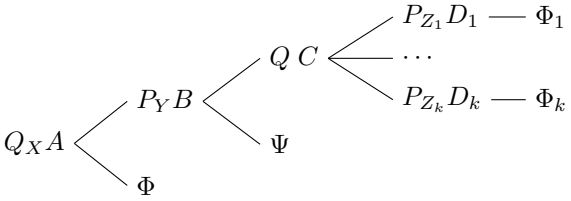
The formula from the previous example converted by the algorithm 1 will have the following form:



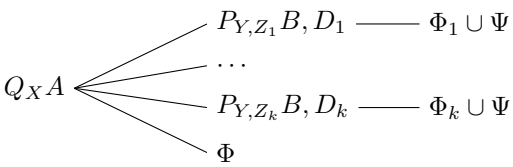


As we can see, the output of the algorithm (a PCF) contains lot of fictitious quantifiers $\forall\varnothing$ and $\exists\varnothing$. Moreover, there can be generated redundant branches that will have negative influence to the deduction search productivity of inference engine. The following transformation will contract the output PCF.

Theorem. (PCF reduction rule)
If in a PCF F:



$P, Q \in \{\forall, \exists\}, P \neq Q$, $\Phi, \Psi, \Phi_i, i = \overline{1,k}$ are sets of subformulas, A, C, B are conjuncts, for which the condition $C \subseteq B$ holds, then F is identical to F' that have the following form:



Proof.
Let the conjunct C be of a form $C^1\& \ldots \&C^n$. If $C \subseteq B$, then $B = B'\&C^1\& \ldots \&C^n$, where B' is a conjunct, which can be empty. Convert the PCF F to the FOF language.

1. If F rooted with $\forall$ quantifier, then F will be of the form

$$F = \forall_X A \to \big(\Phi \vee \exists_Y(\Psi\&B\&(\neg C \vee \exists_{Z_1}(D_1\&\Phi_1) \vee \ldots \vee \exists_{Z_k}(D_k\&\Phi_k)))\big)$$

Opening parenthesis results in a disjunction of conjunctions, one of which will be of form $\Psi\&B\&\neg C = (\Psi\&B'\&C^1\& \ldots \&C^n)\&(\neg C^1 \vee \ldots \vee C^n) =$ **False**, i.e., it is insignificant. Therefore,

$$F = \forall_X A \to \big(\Phi \vee \exists_Y \exists_{Z_1}(B\&D_1\&\Psi\&\Phi_1) \vee \ldots \vee \exists_Y \exists_{Z_k}(B\&D_k\&\Psi\&\Phi_k))\big),$$

which has the same PCF–form as F'.

2. If F rooted with $\exists$ quantifier, its transformation into the FOF language will be of the form:

$$F = \exists_X A\&\Phi\&(\forall_Y B \to (\Psi \vee C\&\forall_{Z_1}(D_1 \to \Phi_1)\& \ldots \&\forall_{Z_k}(D_k \to \Phi_k))).$$

Transform expression in parenthesis.

$$\forall_Y B \to (\Psi \vee C \& \forall_{Z_1}(D_1 \to \Phi_1) \& \ldots \& \forall_{Z_k}(D_k \to \Phi_k)).$$

Eliminate implication and group disjunctive elements into a separate bracket.

$$\forall_Y((\neg B' \vee \neg C^1 \vee \ldots \vee \neg C^n \vee \Psi) \vee$$
$$C^1 \& \ldots \& C^n \& \forall_{Z_1}(D_1 \to \Phi_1) \& \ldots \& \forall_{Z_k}(D_k \to \Phi_k)).$$

Using the distributive law, we obtain:

$$\forall_Y((\neg B' \vee \neg C^1 \vee \ldots \vee \neg C^n \vee \Psi \vee C^1) \& \ldots$$
$$\&(\neg B' \vee \neg C^1 \vee \ldots \vee \neg C^n \vee \Psi \vee C^n) \&$$
$$(\neg B' \vee \neg C^1 \vee \ldots \vee \neg C^n \vee \Psi \vee \forall_{Z_1}(D_1 \to \Phi_1)) \& \ldots$$
$$\&(\neg B' \vee \neg C^1 \vee \ldots \vee \neg C^n \vee \Psi \vee \forall_{Z_k}(D_k \to \Phi_k))).$$

The first n brackets contain tautology $\neg C^i \vee C^i, \forall i = \overline{1,n}$, therefore:

$$\forall_Y(\mathbf{True} \& \ldots \& \mathbf{True} \& (\neg B \vee \Psi \vee \forall_{Z_1}(D_1 \to \Phi_1)) \& \ldots$$
$$\&(\neg B \vee \Psi \vee \forall_{Z_k}(D_k \to \Phi_k))) =$$
$$\forall_Y \forall_{Z_1}(\neg(B \& D_1) \vee \Psi \vee \Phi_1) \& \ldots$$
$$\& \forall_Y \forall_{Z_k}(\neg(B \& D_k) \vee \Psi \vee \Phi_k).$$

That is

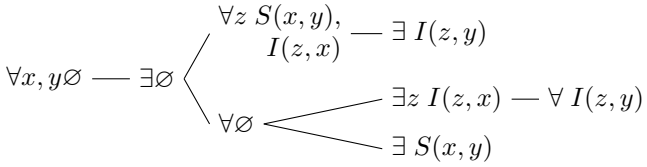
$$F = \exists_X A \& \Phi \& \forall_Y \forall_{Z_1}((B \& D_1) \to (\Psi \vee \Phi_1)) \& \ldots$$
$$\& \forall_Y \forall_{Z_k}((B \& D_k) \to (\Psi \vee \Phi_k)),$$

which being converted to PCF will result in a formula that coincides to F'.

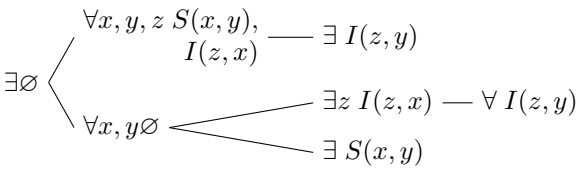
Proof is complete.

Quantifier QC, which is located in F of PCF reduction rule, is referred to as *trivial quantifier*. Also, if formula F (of the rule) has the root of the quantifier form $P_Y B$, then the PCF must be adjusted by adding new fictitious quantifier $Q\varnothing$ as new root. PCF reduction rule should be applied to node pairs and for the "straight" parts (with no branching nodes) inside a PCF's tree structure starting from leaves of the PCF. Application to the branching parts must be avoided unless the resulting contracted PCF will contain less number of nodes.

Let us drive our example forward and contract its PCF, this will result in the following form:



Here we have two directions, either quantifier $\exists\varnothing$ to be reduced, or $\forall\varnothing$. The last case results in a greater number of nodes, as subtree $\forall z\ S(x,y), I(z,x) - \exists\ I(z,y)$ have to be copied in PCF reduction rule application. Finally, the resulting PCF tree of the example formula is as follows:



IV. Conclusion

A new efficient algorithm for conversion of the predicate calculus language formulas to the language of positively constructed formulas (PCF's) and equivalent rules of reduction of the PCFs is considered in this paper. The algorithm preserves the original heuristic structure of knowledge represented by the formulas of the predicate calculus language.

One of the further development of the conversion module is an implementation of the algorithm in a new version of prover, adopted to syntax and properties of TPTP problem set library [6]. Another direction is to adapt the algorithm to the disjunctive language widely used in provers based on the resolution method, as most of the problems of the library are represented in the conjunctive normal forms. The main problem here is to reconstruct the heuristically structure broken during scolemization and conversion (from first–order language) stages. We already have a version of the presented algorithm for conjunctive normal form formulas, which have no existential variables.

Acknowledgment

The research is supported by the Russian Science Foundation (project no. 16-11-00053).

References

[1] S.N. Vassilyev *Machine Synthesis of Mathematical Theorems*. The Journal of Logic programming, V.9, No.2–3, pp. 235–266,1990.

[2] S.N. Vassilyev, A.K. Zherlov, E.A. Fedunov, B.E. Fedosov *Intelligent Control of Dynamic Systems*. Moscow, Russia: Fizmatlit, 2000. (in Russian)

[3] A.V. Davydov, A.A. Larionov, E.A. Cherkashin *On the calculus of positively constructed formulas for automated theorem proving*. Automatic Control and Computer Sciences (AC&CS), V.45, No.7, pp. 402–407, 2011.

[4] E.A. Cherkashin, A.A. Larionov, A.V. Davydov *Calculus of Positively Constructed Formulas, its Features, Strategies and Implementation* // MIPRO 2013/CIS (36-th international convention on information and communication technology, electronics and microelectronics), 20-24 May 2013, Chroatia, Opatija, 2013. - pp. 1289–1295.

[5] N. Bourbaki *Theory of Sets*. Paris: Hermann, 1968.

[6] G. Sutcliffe *The TPTP Problem Library and Associated Infrastructure. The FOF and CNF Parts, v3.5.0.* Journal of Automated Reasoning, V.43, No.4, pp.337–362,2009.

Monte-Carlo Randomized Algorithm: Empirical Analysis on Real-World Information Systems

Robert Kudelić*,1, Dijana Oreški* and Mario Konecki*,1
* Faculty of organization and informatics, Varaždin, Republic of Croatia
[1] Artificial Intelligence Laboratory
robert.kudelic@foi.hr, dijana.oreski@foi.hr, mario.konecki@foi.hr

Abstract - Determination of development priority of information system subsystems is a problem that warrants resolution during information system development. It has been proven, previously, that this problem of information system development order is in fact NP-complete, NP-hard, and APX-hard. To solve this problem on a general case we have previously developed Monte-Carlo randomized algorithm, calculated complexity of this algorithm, and so on. After previous research we were able to come into possession of digraphs that represent real-world information systems. Therefore, in this paper we will empirically analyze Monte-Carlo algorithm to determine how the algorithm works on real-world examples. Also, we will critically review the results and give some possible areas of future research as well.

I. Introduction and Related Work

When one develops information system the question of development order of its subsystems naturally comes up. To illustrate this let us look at Figure 1, where nodes are representing information system subsystems and arcs are representing data classes. For example, if we were to determine development order randomly, one order could be $O = [4,5,3,2,1]$. Nevertheless, if the criteria for determining development order is the sum of weights of feedback arcs, as defined in [1], then obviously previously determined order $O = [4,5,3,2,1]$ would be problematic. Namely, number of data classes going from node 2 to node 3 greatly exceeds number of nodes that are going in the opposite direction. Therefore it is very questionable whether order of information system development should stay as was determined in the first instance.

To solve this problem we have previously devised and tested Monte-Carlo randomized algorithm which solves this problem of information system development order (ISSDO) in polynomial time with arbitrary probability [2]. It should also be noted that there are many other algorithms that have also been devised for the purpose of solving MFAS (Minimum Feedback Arc Set) which is according to [2] essentially the same as ISSDO. For example, in [3] we find linear programming approximation algorithms for feedback set problems, in bipartite tournaments. In this paper authors have given, in addition to 2-approximation algorithm for the feedback vertex set problem, an alternative 4-approximation algorithm, with regards to 4-approximation algorithm proposed by Gupta that turned out to be incorrect, for the feedback arc set problem based on an algorithm for feedback arc set in (regular) tournaments [2]-[3]. Also, in [3] we find two algorithms for feedback arc set problem, both algorithms work on rotator graphs [2]-[3]. Authors have proved that both algorithms derive optimal solution, algorithm complexity for both algorithms is $O(A \times N!)$ [2]-[3]. Furthermore, in [4] we find parametrized algorithms for feedback set problems and their duals in tournaments. In this paper authors have found algorithm for weighted feedback arc set problem in tournaments whose running time is $O(2{,}415^k n^\omega)$, where ω represents the exponent of the best matrix multiplication algorithm, and an $O(4^k k + m)$ algorithm for the dual of feedback arc set problem (maximum arc induced acyclic graph) in general directed graphs [2] [4]. There are also many other algorithms which we will not be mentioning here.

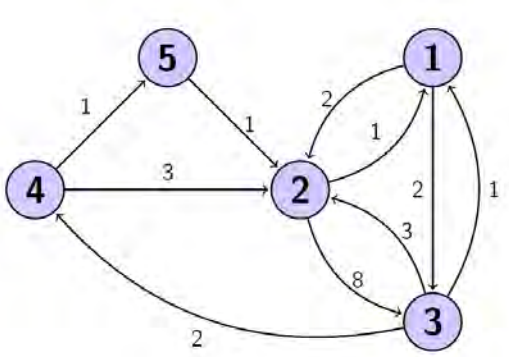


Figure 1: Example of an Information System digraph [1].

Nevertheless, all algorithms have been tested on artificial cases. And while this is not necessarily a bad thing it would be good though to make empirical testing on real-world cases - so it can be seen how difficult these cases are and how well algorithms cope with them. Therefore, this will be the main contribution of our paper.

II. Empirical Analysis

First and foremost, the algorithm that we are primarily interested in is Monte-Carlo randomization that we have developed. This algorithm is finding optimal solution, for the problem of information system subsystems development order, with arbitrary probability in polynomial time [2]. Algorithm input is multigraph, definition of which can be seen in Definition 2.1.

2.1. DEFINITION [2]. A graph which: allows multiple arcs between every pair of nodes, has no loops, and has no arc weights; is multi-graph.

Probability with which Monte-Carlo algorithm is finding solutions can be seen in Lemma 2.1.

2.1. LEMMA [2]. If we transform a graph that represents input for ISSDO into multi-graph, probability of breaking all cycles, between pairs of nodes connected with minimal number of arcs, in a graph by constantly breaking arcs uniformly is inversely proportional to the probability of choosing an arc between pairs of nodes connected with minimal number of arcs.

$$\vdots$$

$$P(opt) = 1 - \left(1 - \left(\frac{n-2}{n}\right)^{\frac{n}{2}}\right)^{pi} \qquad (1)$$

$$Q.E.D.$$

Algorithm complexity is $O(k|V|^3)$. Pseudocode can be seen in [2]. Empirical running time, on randomly generated digraphs, can be seen on Figure 2.

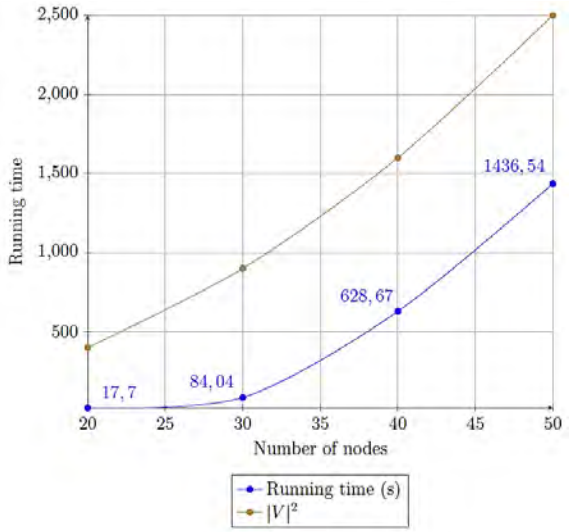


Figure 2: Empirical running times on randomly generated digraphs, for Monte-Carlo randomization [2].

Concerning real-world information system digraphs that we were able to acquire, on which we will do scientific analysis of our Monte-Carlo randomized algorithm and interpretation of the data. During their studies students are conducting, at the Faculty of organization and informatics, analysis of real-world information systems – with professor's help. From all analyses that were done by students ten projects were selected. From these ten projects we have, from p/dc matrixes, created information system digraphs which can then be sent as an input to Monte-Carlo randomized algorithm. Information system subsystems in those matrixes were determined according to business functions. These digraphs had at most six nodes, that is subsystems; this value was also mode. Number of arc was generally below $\binom{v}{2}$, where *v* represents number of nodes. Although, there are some instances where it was above. Minimal arc weight was of course one while maximal was six. After we have done empirical analysis on these digraphs, which were given as input to Monte-Carlo randomization, following results have been acquired.

In Table 1 we see the data that have been acquired by conducting empirical analysis on real-world IS digraphs. For every digraph Monte-Carlo randomized algorithm was run *1000* times.

Table 1: Empirical analysis on real-world IS digraphs.

Graph	Number of nodes	Number of iterations	Running time	Probability
1	6	2	0,00136	0,24603
2	4	2	0,00115	0,23438
3	5	1	0,00079	0,21600
4	6	7	0,01405	0,24316
5	6	87	0,06027	0,96865
6	4	4	0,00090	0,41382
7	6	11	0,00562	0,91114
8	5	28	0,02065	0,73758
9	6	39	0,01701	0,97222
10	4	11	0,00206	0,76981

After *1000* iterations the best solution was chosen. With having this in mind, algorithm obviously needed only a few iterations to find a solution upon which it couldn't make any more improvements. In seven cases number of iterations was ≤ 11, and in all of the cases number of iterations was within ≤ 87. Since after 87 iterations algorithm didn't find any solution that would be better for any digraph and since the remaining number of iterations was 913, solutions found by the algorithm are probably optimal or very close to optimum – as it can be seen from [2]. Corresponding directly to number of iterations are running times. As a direct result of small number of iterations running times are also quite short, which is of course expected with a situation like this. A more detailed comparison of running times will be done later in the paper. Lastly, for every digraph the algorithm has outputted probability that we have reached optimal solution. As it can be seen algorithm can find solutions that are optimal or close to optimum with probabilities that are ranging from $[21.6, 99.99 \ldots]$. This is of course an excellent characteristic since it requires smaller number of iterations. This was not surprising and it confirms our findings from [2] – where we were able to find solutions that are optimal or close to optimum even with smaller probabilities and where we have conducted comparison with known optimal solutions. Also, take note that even for smaller probabilities algorithm has found the best solution quickly and then didn't find another solution that would be better in a much larger number of iterations. That means that solution is very likely close to optimum or optimal since digraphs were not that difficult to solve because information system subsystems were determined according to business functions. Probabilities are usually small on digraphs that have small number of arcs, since the greater the number of arc the greater the probability that the smallest set of arcs will be broken, by breaking set pf arcs uniformly [2].

After previous analysis we have aggregated the data, for every node, so it can be determined how the algorithm fares generally for every number of nodes. These results can be observed in tables 2 - 3. Number of nodes of digraphs on which we have done empirical analysis was ranging from [4, 6]. An average number of iterations was for every number of nodes far below total number of

iterations which was *1000*. Standard deviation of number of iterations was actually high since there are some extreme values, nevertheless this is again far below total number of iterations which was *1000*.

Table 2: Aggregate data analysis for every number of nodes.

Number of nodes	Number of iterations ($\bar{x}$)	Number of iterations (σ)	Running time ($\bar{x}$)	Running time (σ)
4	5,66666	4,72581	0,00137	0,00061
5	14,50000	19,09188	0,01071	0,01404
6	29,20000	35,35816	0,01966	0,02355

Table 3: Aggregate data analysis for every number of nodes, continuation from Table 2.

Probability ($\bar{x}$)	Probability (σ)
0,47266	0,27252
0,47679	0,36881
0,66824	0,38749

If we look at running time, both the average and standard deviation, we see that algorithm runs quite fast. That is, when we take into account that for every instance Monte-Carlo randomization was run *1000* times we can conclude that the algorithm is quite quick, on these digraphs, since running times are very short. Lastly, when we look at the probability we see that, on average, they are ranging from [47, 67] and that on these probabilities algorithm starts to find solutions that probably can't be improved. The standard deviation is high but if we look at Table 1 we see that the frequency of probabilities that are ≤ 0.5 is *5*. Also, the frequency of probabilities that are $\geq 0{,}9$ is 3 and the frequency of probabilities that are $\geq 0{,}7$ *and* $\leq 0{,}8$ is 2. Therefore, we could conclude that Monte-Carlo randomized algorithm finds solutions that are optimal or close to optimum, with probabilities that are ≤ 0.5, in *50%* of the cases. We could also conclude that the algorithm needs high probabilities that are $\geq 0{,}9$ in *30%* of the cases. Globally, that is of course a good thing since it means that we will need smaller number of iterations and consequently shorter running times. All of this of course concerns only real-world information systems that is digraphs on which we have tested.

Lastly, let's consider running times of the algorithm on real-world information systems. That is, let us compare running times of the algorithm on real-world cases and on randomly generated digraphs that can be seen on Figure 2. Empirical running times on real-world digraphs can be seen on Figure 3. On Figure 3 we see three lines. The one in blue represents empirical running times on real-world information systems, the one in brown represents theoretical quadratic growth, which is deliberately uniformly reduced since empirical running times previously mentioned are quite small, and the one in magenta represents simple extrapolation from the data that represents empirical running time. Empirical running times obviously represents linear growth, therefore to make things more clearer we have calculated expected value and standard deviation and after that we have randomly generated values according to normal distribution and made simple extrapolation from the data. Growth of empirical running times is linear, which is obviously bellow running times that can be observed on Figure 2, representing running times on randomly generated digraphs that represent polynomial growth – approximately $O(|V|^2)$. Therefore we can conclude that digraphs which can be found in practical situations are much easier to solve, of course if the subsystems of an information system are determined according to business functions. If that is not the case then order of growth would probably be higher since there would be more information system subsystems, this is however an assumption, that requires additional research, and can't be presently determined.

There are some research constraints which need to be mentioned here. The first constraint is that the sample size is ten and it is possible that there are information systems which would not represent homogenous group with the present sample. The second constraint is that it is quite probable that there are information systems that have number of subsystems greater than six – which was the greatest number of subsystems in our sample. Nevertheless, these two constraints probably don't represent something that raises serious concern since according to [6] there are seven rudimentary business functions. Which means that regardless of the graphs that could be found Monte-Carlo randomized algorithm would find good solutions that are in accordance with the results published in [2]. According to our own practical knowledge it happens sometimes that business functions are additionally divided nevertheless this additional division is never enormous.

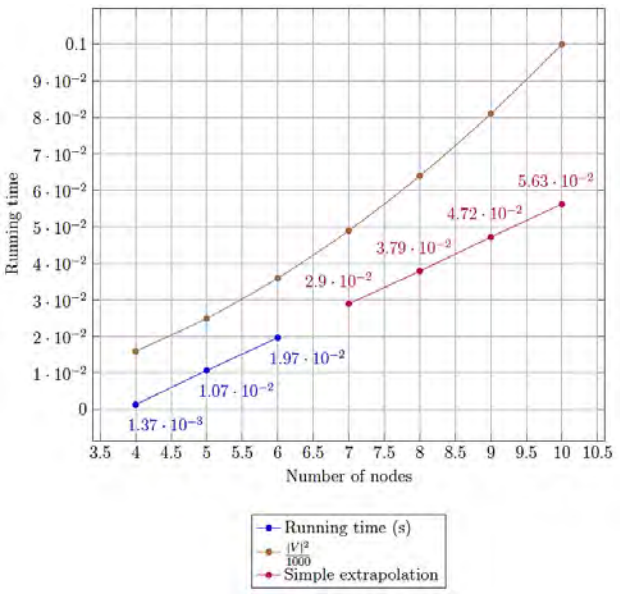


Figure 3: Empirical running time of Monte-Carlo randomized algorithm on real-world digraphs.

Third and last constraint is the fact that subsystems are determined, in the current sample of ten, according to business functions and not according to individual applications on which we would then look for optimal sequence of nodes according to the sum of weights of feedback arcs. This constraint is more serious than the first two but it can probably be resolved even without real-world digraphs in which subsystems are actually individual applications. Namely, by developing an algorithm that would generate larger digraphs that are of a type described in this paper. In this way we could probably make a good estimate of the situation since arc distribution would probably be similar in both cases. The reason of course why this further research will probably be done in this way is because it is actually quite difficult to obtain real-world information system digraphs of this type.

III. CONCLUSIONS

The goal of this research was to obtain digraphs that represents real-world information systems and test how well does the Monte-Carlo randomized fares on these kind of digraphs since until know this wasn't known.

After scientific analysis and interpretation of the results we came to the conclusion that real-world information systems are much easier to solve than expected, with having in mind of course that subsystems are determined according to business functions. Namely, empirical running time on real-world information systems was $O(n)$ versus running time that was previously obtained on randomly generated digraphs which was $O(n^2)$.

This means that if information system subsystems are determined according to business functions Branch and Bound algorithm which will always find optimal solution is the best choice. If on the other hand this isn't the case and IS subsystems are determined according to development of individual applications and number of subsystems is considerably larger than some kind of approximation algorithm is probably a better option – if we have efficiency in mind.

We will focus our further research on devising an algorithm which will be able to generate graphs of the type in our sample so we can determine some reasonable upper bound – with the assumption that arc distribution would probably be similar in both cases since determination of subsystems has no effect on arcs.

LITERATURE

[1] Robert Kudelić, Alen Lovrenčić, and Mladen Konecki, "Automatic Determination of Information System Subsystems Execution and Development Order", International Review on Computers and Software (I.RE.CO.S.), Vol. 6, N. 4, pp. 455 – 461, July 2011.

[2] Robert Kudelić, "Monte-Carlo Randomized Algorithm for Minimum Feedback Arc Set", Applied Soft Computing, Vol. 41, pp. 235–246, 2016, doi:10.1016/j.asoc.2015.12.018.

[3] A. van Zuylen, "Linear Programming Based Approximation Algorithms for Feedback Set Problems in Bipartite Tournaments", J. Chen, S. B. Cooper (Eds.), Theory and Applications of Models of Computation, Vol. 5532, pp. 370, 2009, doi:10.1007/978-3-642-02017-9_39.

[4] C.-J. Kuo, C.-C. Hsu, H.-R. Lin, D.-R. Chen, "Minimum Feedback Arc Sets in Rotator Graphs", http://par.cse.nsysu.edu.tw/algo/paper/paper09/C1-3.pdf, loaded 2.9.2015., 2009?.

[5] V. Raman, S. Saurabh, "Parameterized algorithms for feedback set problems and their duals in tournaments", Theoretical Computer Science 351 (3), pp. 446-458, 2006, doi:10.1016/j.tcs.2005.10.010.

[6] Žugaj, M., J. Šehanovic, and M. Cingula, "Organizacija", 2. Edition, TIVA Tiskara Varaždin, page 219, 2004.

Control Flow Graph Visualization in Compiled Software Engineering

Andrey Mikhailov*, Aleksey Hmelnov*, Evgeny Cherkashin* ** and Igor Bychkov*
*Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of Russian Academy of Sciences,
**National Research Irkutsk State Technical University,
Irkutsk, Russian Federation
{mikhailov,alex,eugeneai,bychkov}@icc.ru

***Abstract*—Engineering of the compiled software is an important task in software engineering. One of the stages of the engineering technologies is a construction and analysis of a control flow graphs, which reflect a general structure of algorithms. The paper presents a technique for analyzing and visualizing the control flow graph of a compiled software. The analysis is based on semantically equivalent transformations of the original graph resulting in an abstract node that contains a hierarchy of isolated regions, which then associated with predefined templates resulting a graphical representation. The templates and additional signature data allow one to recognize the high-level programming structures and user statements to construct a flow-chart notation of the original program.**

I. Introduction

Reverse engineering of compiled software (programs after processing its source code by a compiler) is an important task in software engineering as it allows one to solve a wide variety of software quality assessment problems, such as productivity of compiled code, quality of compiler and system libraries, peculiarities of hardware architectures the program running on. Reverse engineering is also used for regain the access to legacy code, whose source counterpart is already lost, and for investigation of virus and malicious program behavior. One of the helpful stages of reverse engineering techniques is a construction and analysis of so-called control flow graph that reflects a general structure of corresponding algorithms. Artificial intelligence techniques such as pattern recognition and rewrite rule engines have extensive use in the analysis.

The control flow graph is a natural representation that can be automatically figured out from the source code or its compiled binaries. The graph is used by compiler as an intermediate structure for optimizing transformations. The source code text contains all the necessary information of program behavior. Its analysis usually is a complex task even with technological support of nowadays integrated development environments, which make the analysis significantly easier.

Binary code analysis is a more complex task, and it requires a complex IT knowledge. Special software such as disassemblers and decompilers are used to partially automatize control flow analysis and visualization. The decompiling a binary runtime is in principle an impracticable task: the reconstructed high-level code is more difficult in understanding as compared to its assembler representation. The visual analysis of the control flow graph is an alternate way. There are a number of visual representation and processing software, which are relatively universal, such as uDraw (daVinci) [1], VCG [2], Graphlet [3], GraVis [4], Graph Drawing Server [5], graphViz [6], VisualGraph [7]. These systems have a rich general toolkit but visualization and analysis of control flow graph requires a special processing routines. Algorithms of graph visualization are developed since 60th [10]. The results of P. Eades [11], T. Kamada and S. Kawai [12] considered to be classical in this research field, they deal with universal approaches to graph visualization problems.

The control flow graph is directed and usually visualized by means of layers [13]. The process consists of four stages:

Distribution of graph nodes between layers. Each node is assigned a rank. All directed edges can connect nodes from a lower rank to a higher one. Obviously, nodes of the same rank cannot be connected. Rank distribution of the nodes is performed with various techniques, the simplest one is based on path length calculation in depth-first graph traversal procedure.

Defining order on the nodes in a layer. The nodes of a layer are ordered according to principle of minimization of intersections of edges. Method of median [14] is most widely used as a technique for solving the problem.

Figuring out of the node coordinates in a layer. Each node of each layer is assigned a coordinate so as the graph will correspond to predefined aesthetic criteria.

Edge drawing. On this stage the edges are drawn according to rules of visualization, for example, edges can be indirect lines.

The paper describes a technique and software tools for analyzing and visualizing the control flow graph of a compiled binaries. The analysis is based on semantically equivalent transformations of the original graph resulting in a singular abstract node that contains a hierarchy of isolated regions (subgraphs). The regions are associated with predefined templates to form resulting graphical representation. The templates and additional signature data for library modules allow one to recognize the regions as high-level programming language structures and statements to construct, e.g., flow-chart notation of the original program.

II. Theoretical basis of control flow graph visualization

An directed graph $G(V, E)$ is a *control flow graph* if the following restrictions hold:

1) graph G does not contain multiple edges between the same nodes;
2) node $start \in V$ is the entrance to the graph (*start node*);
3) node $end \in V$ is the exit from the graph (*terminal node*);
4) each node $v \in V$ is accessible from $start$;
5) node end is accessible from each node $v \in V$.

If a control flow graph has more than one terminal node, the terminal nodes are connected to one new fictitious terminal node.

We will use additional notions taken from [18] and used further in the paper. A node x is a *dominator* of y ($x \quad dom \quad y$) in a directed graph, if any path from $start$ to y includes x. A node x is a *postdominator* of y ($x \quad pdom \quad y$), if any path from y to end includes x. By convention, a node dominates and postdominates itself. A node x is an *immediate dominator* of y ($x \quad idom \quad y$), if $x \quad dom \quad y$ and there are no such p that $x \quad dom \quad p$ and $p \quad dom \quad y$. A node x is an *immediate postdominator* of y ($x \quad pidom \quad y$), if $x \quad pdom \quad y$ and there are no such p that $x \quad pdom \quad p$ and $p \quad pdom \quad y$.

Let a and b be a pair of distinct edges in a control flow graph. The ordered pair of a and b will be denoted as $\langle a, b\rangle$. If 1) a dominates b, 2) b postdominates a, and c) every cycle containing a also contains b (and vice versa), then the pair $\langle a, b\rangle$ is referred to as a *single entry single exit region* (SESE region). In the SESE region, a is the entry edge and b is the exit edge.

The first condition ensures that every path from $start$ (of the whole graph) into the region passes through the entry edge a. The second condition ensures that every path from inside the region to end passes through the exit edge b. The third condition implements two constraints: every path from inside the region to a point above a passes through the exit b, and every path from a point below b to a point inside the region passes through the entrance a, i.e., these path go around the SESE region.

Two SESE regions in a control flow graph must be either folded or do not overlap (contain no common nodes) [18]. Structural analysis on the base of recognition of SESE regions commonly used for efficient construction of an intermediate image in a so called *static single assignment* (SSA) form, as well as in data flow analysis. SESE regions are also organized in *program structure trees* (PST), which represent the control flow graph as hierarchy of folded regions (SESE regions and individual nodes folding tree). Each of the PST tree nodes denote a procedure consisting of other procedures and individual operators.

A control flow graph subgraph having one entry and one exit nodes is referred to as *two terminal region* (TT region). TT regions emphasize the points of control flow convergence and the following divergence (Fig. 1). The requirement condition of TT region recognition are weaker as compared to SESE regions, so all the SESE regions are TT regions but not vice versa.

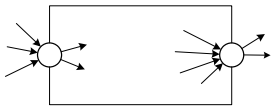


Fig. 1. A two terminal region

For the structural programming language, TT region corresponds to a branching operator or a cycle. Our approach to the control flow graph visualization is based on recognition of TT regions and classification them into finite set of variants concerned with various control flow operators of the original high-level programming language. TT regions of program graph of control flow replaced with abstract nodes (folded) form an analysis basis of visualization.

A. Quality criteria of the control flow graph visualization

A display of the nodes and the edges of a graph on a surface (or in a 3d-space) is referred to as a *graph layout*. A graph can be visualized using various layouts (Fig. 2), and different layouts are defined by sets of display rules and criteria of result assessment.

The main quality criterion of a control graph visualization is the correspondence of the constructed image to the nature of the visualized informational object. The quality assessment depends on various structural and semantic characteristics of the object. Three basic feature sets are emphasized specially among the criteria: *visual arrangement*, *aesthetics*, *restrictions*.

Visual arrangement is the main set of rules that a graph representation must obey to be acceptable as a desired result. For example, to visualize programs as a flowchart, layout engine must obey agreement rules, which represent all the graph nodes as flowchart symbols and the edges as polylines consisting only with vertical and horizontal sections. In a general case, an visual arrangement can be sufficiently complex and include a lot of details related to the target representation.

Aesthetics is a subset of the criteria that defines attributes of the constructed image, which must be obeyed as much as possible to improve visibility. A widely used aesthetic criteria are the following: minimizing a number of edge intersections, minimizing a number of foldings, minimizing the occupation area, maximizing the resolution,

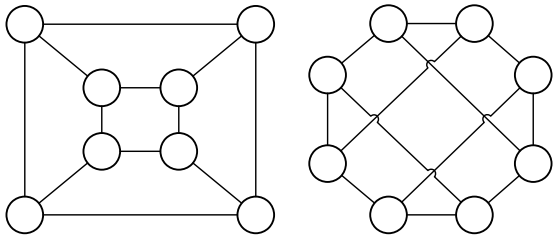


Fig. 2. Various layouts of the same graph

minimizing the total edge length, unification of edge lengths, minimizing the number of foldings on an edge, foldings unification, maximizing a symmetry of graph representation, etc.

Restrictions are a subset of the criteria that define layout rules for peculiar elements and subgraphs of the constructed image. The most common used criteria are *the center* that define whether the root node must be placed at the center of the image, *externality* that define the necessity to place a node outside a region, *cluster* defines that a number of nodes to be placed together near to each other, *sequence* defines the order of nodes placement during drawing paths (left to right, top-bottom).

A control flow graph reconstructed from a binaries consists of operator block nodes, starting node, and an non-empty set of finish nodes, as well as edges correspond to the transfer of control. A natural image of the control flow graph is a flowchart diagram. For the flow charts we define the following visual arrangement for its nodes (Fig. 3):

Operator shape represents a node where the control flow passed only one direction.

Decision shape corresponds to conditional operators in high-level programming languages; in a control flow graph it is a node, where flow control splits up.

Cycle edge shapes denote two graph nodes, one is for beginning of the cycle and one for its end, the cycle body is located between these shapes.

Starting and terminal shapes mark the entrance and the exit from a function or a program.

In this paper, we consider a layout algorithm that visualizes control flow according to criteria defined for flowcharts.

III. Control flow graph visualization

Nowadays programming languages in most cases contain a standard set of high-level control flow operators (`if-then`, `if-then-else`, `for`, `while`, etc.). These operators produce their specific subgraphs during compilation. For example, `if-then` operator produces a graph depicted in Fig. 4, where `then` part could correspond to another control subgraph.

One can describe a corresponding subgraphs for all flow control operators of all standard high-level languages. Each generated subgraph is described as a template. The control operators are revealed and reconstructed in whole flow control graph by means of recognition of the templates in the graph and folding the recognized structure into an abstract node. The edges terminating at and originating from the abstract node are reconnected accordingly. The reconstruction process is repeated until either the graph folds into one abstract node,

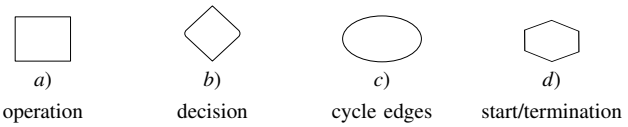


Fig. 3. Types of shapes used in flow charts

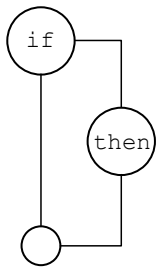


Fig. 4. Subgraph of `if-then` operator

and in this case the graph is referred to as *reducible*, or there is a subgraph that does not correspond to any of the templates. In the last case the graph is called *inreducible*, and the subgraph that was not recognized is named *undeterminate*.

Structural high-level programming language compilers produce only reducible flow control graphs. This will hold for most of other languages as long as `goto` operator is not used explicitly or implicitly. There are two approaches for overcoming the problem of folding impossibility:

1) Perform riddance operation over undeterminate subgraphs. The essential disadvantage of the approach is the necessity of addition and deletion of nodes and edges in the graph, this is not allowed in control flow graph model visualization.
2) Recognize the undeterminate subgraph and replace it with an abstract node.

A. Algorithm of control flow graph structuring

The developed algorithm is based on structural analysis technique proposed in [16] that suggests recognizing subgraph of a predefined form with templates and replacing them with an abstract node, then edge endpoints reassignation is performed accordingly. Template application is carried out according to depth-first order until the graph folds into one abstract node. A tree of dominators is constructed. The tree is used for edge classification and cycle recognition. The result quality strongly depends on template application order. The algorithm is as follows:

If a program has no predefined terminating nodes, the nodes is figured out by means of the following procedure. Firstly, the edges are categorized as *forward*, *backward*, and *cross* directed. Among nodes connected with forward edges, there will be at least one that has no source edge. If we found the only node, then the node is the termination. Otherwise all such nodes are connected with edges to an additional fictitious termination node. The obtained graph is correct as the termination node is accessed from any node of the graph.

Traversing of the dominator tree is done in a reverse breadth-first order, i.e. from leaves to the root, placing graph nodes in a `stack` structure. A region having the deepest folding level (TT region) is determined on each iteration of the algorithm according to the rule $p\ pidom\ v$: the postdominator of node v of S is the terminal node, where the control flow passing v converges. This node is not added to the recognized

Input parameters: G, D, P
Result: An abstract node containing a hierarchy of folded subgraphs
for each $v \in D$ *in a backward breadth-first order* **execute**
  for each $p \in Children(v)$ **execute**
    if p $pidom$ v **then**
      $S \leftarrow Children(v) \setminus p$
      if $Classify_Region(S) \neq$ *undeterminated* **then**
        $Apply_Template(S)$
      end of condition
      else
        $Hierarchical_Layout(S \cup p)$
        $Recognize_Undeterminanted_Region(S)$
      end of condition
      $Modify(G, D, P)$
    end of condition
  end of loop
end of loop

Algorithm 1: Algorithm of control flow graph structuring

region as it could be terminal node of other region of the graph.

As soon as a set of nodes of a TT region is recognized, the region is categorized with templates (Fig. 5), which are sequentially applied to it. If the region corresponds to a template it is considered to be *determined* otherwise *undetermined*.

For the determined regions a new abstract node is constructed. The node embodies its subgraphs, and bounding rectangle of the region calculated according to rules defined in the template. The Hierarchical layout engine is applied to the undetermined regions followed by figuring out the bounding rectangle with respect to the node coordinates. The undetermined region is replaced with new abstract node.

Finally, we obtain one abstract region, having no input and output edges. The region will contain all the hierarchy of folded subroutines. The bounding rectangle is figured out for the region. All graph nodes will be located inside this rectangle.

B. Layout engine

The layout process is carried out recursively from top region down to lower regions. For the top region initial coordinates is predefined. If a region has a template with visualization rules, the rules are applied. The layout example if `if-then-else` operator is given in Fig. 6.

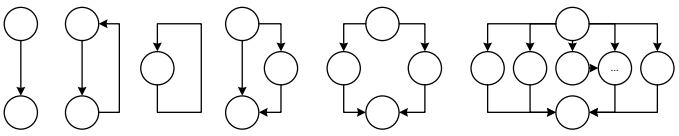


Fig. 5. Applicable templates

The coordinate computation rules for `if-then-else` operator is as follows:

```
if.x = x + abs(width - if.width) / 2;
then.x = x + wDist / 2, then.y = y +
if.height + hDist;
else.x = x - else.width - wDist / 2,
else.y = y + if.height + hDist;
```

where `wDist` is a vertical distance between nodes, `hDist` is a horizontal distance. These parameters is defined manually and reflect aesthetic criteria. The visualizing rules are defined analogously for each template. Thus, the layout engine consists of sequential applications of the visualizing rules to the folded regions.

IV. Implementation

The algorithm is implemented in the Java programming language. The main application is named "VisualGraph" and implemented using authors' library "JGraphX" containing the subroutines of the algorithm. The application uses hierarchical graph models for the control flow visualization using authors' and other known popular visualization algorithms. The "JGraphX" library allows one to make images of control flow graphs by defining shapes and colors for nodes, drawing edges between nodes and corner points automatically and manually.

At present the library and application support some visual arrangement criteria from section II-A. Categorization of nodes in branching regions of the control flow graph requires that the graph will be attributive and contain assembler code of the main program blocks (lineal parts of code). The control flow graph itself is obtained from assembler program semantics analysis by means of authors' FlexT based utilities [19].

V. A layout example

Let's consider a result of our algorithm in comparison to one of hierarchic layout engines (Fig. 7). In the figure, our algorithm (*b*) specially emphasizes the backward edge that corresponds to a loop. The edge style is visually distinguishable from other edges. The algorithm also marks the terminal nodes. Such visualization features allow to recognize visually the regions of interest and other kinds of analysis. In Fig. 7*a*,

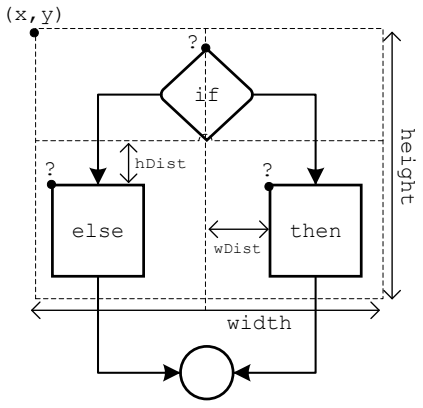


Fig. 6. Visualizing template of `if-then-else` operator

the edges are represented with straight plain lines (except edge connecting v9 and v5). This will ease of control transferring analysis as compared to the form in Fig. 7. Note, that the proposed algorithm accounts not all the variants of control flow patterns and can be improved for example to support undetermined edges removal in a region.

VI. Conclusion

We have proposed and implemented a software tools realizing an approach to surface graph layout on the base of techniques of structural analysis. The main difference of our approach as compared to the mentioned (in the paper) techniques is the usage of a set of matching rules, which are applied to the terminal nodes (a special subgraph) to recognize original control flow structures of high-level programming language.

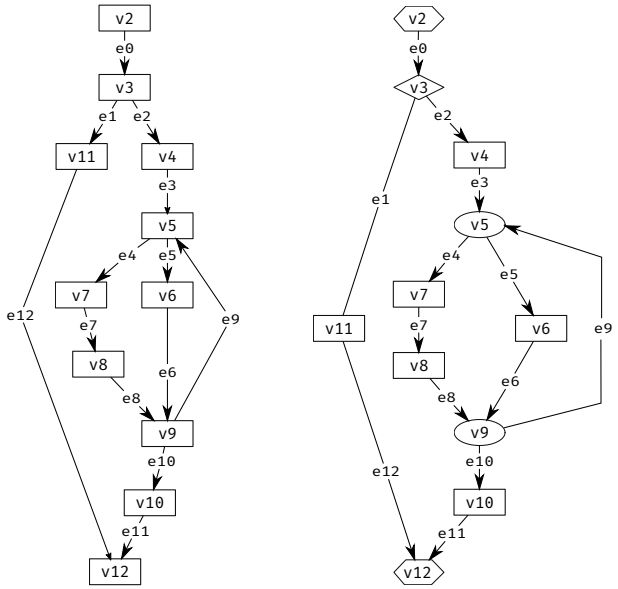
A probation of the developed algorithms was performed on CPU2000 test set[1]. The following results are obtained:

- 197.parser (Syntactic parser of a natural language);
- 252.eon (Ray tracing).

The automation of the flow control analysis allows an engineer to devote more efforts for description of the templates increasing both the personal productivity and detail level of the software description.

The results of the control flow graph analysis could be used in construction of Model driven engineering (MDE) [20] software development tools. The nowadays approaches to MDE is based on change propagation between various models of the software, e.g., source code generation from UML and BPMN2.0, and (in backward direction) refactoring the high-level models according to local source code modification. Comparing results of control flow graph analysis could determine initial change of the structure of the source code, then the change is propagated to the higher-level models. This is however a very complex fundamental problem containing both deduction and induction logical inferences.

[1]https://www.spec.org/cpu2000/ — Standard Performance Evaluation Corporation. SPEC CPU2000 is a test set for assessment of the performance of a central processing unit. The test are used for compiler quality evaluation.



a) hierarchical *b*) structural

Fig. 7. Layouts of a control flow graph

Acknowledgment

The development partially support by Russian Foundation of Fundamental Research, grant No 16-07-00411, "Technologies Development for Dynamic Planning of Service Compositions Execution in a Cloud Environment".

References

[1] M. Frohlich, M. Werner, *Demonstration of the Interactive Graph Visualization System daVinci*, LNCS 894, 1995, P. 266–269.
[2] G. Sander, *Graph loyout through the VCG tool*, LNCS 894, 1995, P. 194–205.
[3] M. Himsolt, *Graphlet System (System Demonstration)*, LNCS 1990, 1996. P. 233–240.
[4] H. Lauer, M. Ettrich, K. Soukup, *GraVis System Demonstration*, LNCS 1353, 1997, P. 344–349.
[5] S. Bridgeman, A. Garg, R. Tamassia, *A Graph Drawing and Translation Service on the WWW*, LNCS 1190, 1996, P. 45–52.
[6] E. R. Gasner, S. C. North, *An Open Graph Visualization System and its Applications to Software Engineering*, http://www.graphviz.org/Documentation/GN99.pdf
[7] T. A. Zolotukhin, *Graph Visualization with Programming System "Visual Graph"*, Computer Science in Science and Education, 2012, P. 135–148. (in Russian)
[8] A. A. Mikhailov, *Control Flow Graph Analysis in Decompilation of Object Files dcuil*, Vestnik of Novosibirsk State University, Computer Science Technologies Series, 2014, Vol. 12, Issue 2, P. 74–79. (in Russian)
[9] V. N. Kasyanov, *Visualizing Information on the Base of Graph Models*, Computational Technologies, 1999. Vol. 11, No. 11, P. 1123–1135. (in Russian)
[10] W. T. Tutte, *How to Draw a Graph*, Proc. of London Math. Society. 1963. 13(52), P. 743–768.
[11] P. A. Eades, *A Heuristic for Graph Drawing*, Congressus Numerantium, 1984, 42, P. 149–160.
[12] T. Kamada, S. Kawai, *An Algorithm for Drawing General Undirected Graphs*, Inform. Process. Lett, 1989, 31, P. 7–15.
[13] P. Eades, L. Xuemin, *How to Draw a Directed Graph*, Visual Languages, 1989, IEEE Workshop on, IEEE, 1989, P. 13–17.
[14] P. Eades, N. C. Wormald, *The Median Heuristic for Drawing 2–Layered Networks*, Technical Report 69, Dept. of Computer Science, University of Queensland, 1986.
[15] A. V. Aho, R. Sethi, J. D. Ullman, *Compilers: Principles, Techniques, And Tools*, Addison Wesley, 1986. MLA
[16] C. Cifuentes, *Structuring Decompiled Graphs. Compiler Construction*, Springer Berlin Heidelberg, 1996, P. 91–105.
[17] R. Johnson,D. Pearson, and K. Pingali, *Finding regions fast: Single entry single exit and control regions in linear time*. Tech. rep., Cornell University, Ithaca, NY, 1993
[18] R. Johnson, D. Pearson, K. Pingali, *The Program Structure Tree: Computing Control Regions in Linear Time*, In PLDI '94: Proceedings of the ACM SIGPLAN 1994 conference on Programming language design and implementation, New York, NY, USA, 1994. ACM. p. 171–185. http://iss.ices.utexas.edu/Publications/Papers/PLDI1994.pdf (access data: 12.02.2016)
[19] A. Hmelnov, S. Vassilyev, *Data description language FlexT: flexible types for description of static data*, Proceedings of CSCC'99 Multiconference. P. 1371-1376 URL:http://www.wseas.us/e-library/conferences/athens1999/Papers/137.pdf (access date: 12.01.2016)
[20] E. A. Cherkashin, V. V. Paramonov, et al, *Model Driven Architecture is a Complex System*. E-Society Journal Research and Applications 2 (2), 15–23.

Bottom-Left and Sequence Pair for Solving Packing Problems

T. Rolich*, D. Domović *, M. Golub **
* University of Zagreb, Faculty of Textile Technology, Zagreb, Croatia
** University of Zagreb, Faculty of Electrical Engineering and Computing, Zagreb, Croatia
*{tomislav.rolich, daniel.domovic}@ttf.hr, ** marin.golub@fer.hr

Abstract - Nesting problem is a combinatorial optimization problem with a goal of placing a set of polygons within container without overlapping. If polygons in shape of cutting patterns are used, and the container has rectangular shape with a fixed width, the problem is defined as a strip packing problem. In this paper two methodologies are described: Bottom-Left (BL) and Sequence Pair (SP). Bottom-left is order based strategy which can solve strip packing problem. It moves each polygon as far as possible to the bottom of the temporary layout and then as far as possible to the left until a stable position is found. In this paper BLD_i version was used with genetic algorithm. Sequence pair is an abstract representation method used for nesting problem that creates a relationship model between rectangular approximations of polygons instead of direct placement. The goal of the research is to obtain maximum packing density for described methodologies. Experiments have been conducted on benchmark datasets: Albano (24 parts), Dagli (30 parts), Mao (20 parts) and Marques (24 parts) in MATLAB environment, with GA-BLD_i performing 78.23, 70.15, 70.70 and 77.63 percent respectively (in average), and SP-BL_i performing 76.28, 63.10, 68.52 and 75.60 percent respectively (in average).

I. INTRODUCTION

Packing problems are combinatorial optimization problems where a set of items $I = \{i_1, i_2, \ldots, i_n\}$, $n = |I|$, need to be placed within the container C so that items i_i and i_j do not overlap ($i_i \cap i_j = 0, \forall i, j \in [1, n]$) and items do not exceed the boundaries of a container ($i_i \cap C = i_i, \forall i, j \in [1, n]$) [1]. If defined in a two-dimensional space, the container and items may be regular or arbitrary shaped polygons.

A packing (often referred to as a placement or a layout) is a set of positions where polygons are placed within a container. The 2D packing layout may be normalized or semi-normalized (Fig. 1). A layout is normalized if each polygon cannot be moved more down or to the left without causing an overlap. A layout is semi-normalized if each polygon that is added to the layout is placed as bottom-left as possible regarding the current packing contour. In this paper algorithms that create normalized layouts have been constructed.

The significance of packing problems lies in real world applications. In most manufacturing processes it is desirable to reduce the amount of waste, and hence maximize the utilization of material. In this paper, experiments on datasets from textile industry have been performed. In textile industry, a set of cutting patterns needs to be optimally allocated on a material in order to minimize waste, and utilize the material.

In computer science this problem is addressed to as the *strip packing problem* [2], but it is also known as the *marker making problem*. Since polygons are irregularly shaped the problem is often referred to as *nesting problem*, also.

The two-dimensional packing problems can be classified as: (1) *knapsack problems* with a goal of finding optimal subset of items in a container, (2) *bin-packing problems* with a goal of minimizing the number of containers used, and (3) *area minimization problems*, with a goal of optimally allocating polygons in a container to minimize the occupied container area.

In this paper two hybrid algorithms using metaheuristics and bottom-left strategies have been used. In the first approach, *a strip packing problem* is solved using genetic algorithm and BLD_i method. A set of irregular polygons are placed within the rectangular area with a fixed width and unlimited length. The goal is to minimize the length, i.e. minimize the waste area between polygons.

The second approach is an *area minimization problem*, where the size of both width and the length is arbitrary, and the goal is to obtain the enclosing rectangular of irregular polygons with a minimal area. Simulated annealing is combined with a sequence pair to perform this task.

One of the greatest challenges in solving packing problems is the overlapping detection methods. Many methods are used in the literature: raster method, no-fit polygon, phi-functions etc. In this paper an abstract representation method is used that prevents overlapping by determining relationships between polygons and decoding their position afterwards.

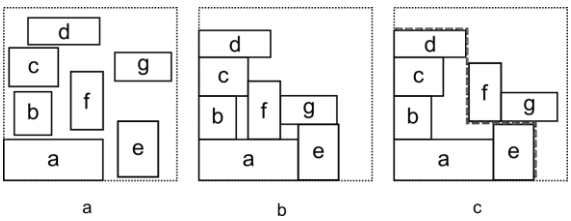


Figure 1. Polygons (a) in a normalized layout (b) and semi-normalized layout (c)

This paper is organized as follows. In Section II the methods used are described. In Section III the results are presented alongside discussion. In section IV a conclusion is given.

II. METHODS

A. Abstract representation methods

Polygon placement can be represented directly as a set of vertices with x and y coordinates of its reference point or by coordinates of all vertices in a container. The drawback of direct representation is that overlaps may occur, therefore complex methods for overlapping detection need to be used.

To resolve this drawback, abstract representation methods are used. Abstract representation methods eliminate overlapping, because they describe the relationship between polygons rather than dealing with their coordinates directly. Using abstract representation methods, determination whether polygon A is left of or right of polygon B, as well as vertical positioning can be determined [1].

Abstract representations methods need to be decoded for the placement of polygons to be visible. Sequence pair is an abstract representation method used in this paper.

B. Sequence pair

Sequence pair (SP) is an abstract representation method, mostly used for rectangle packing problem. It has been first applied on VLSI module design [3].

Sequence pair is a pair of permutation lists consisted of m indexes where $m = \{1, 2, ..., n\}$ and n is the number of rectangles. Each index m represents one rectangle.

The relationship between polygons is decoded using the following rules [8]:

$$(< ..m_i..m_j..>, < ..m_i..m_j.. >) \Rightarrow m_i \text{ is left of } m_j \quad (1)$$

$$(< ..m_j..m_i..>, < ..m_i..m_j.. >) \Rightarrow m_i \text{ is below of } m_j \quad (2)$$

Algorithms for rectangle packing problem have been created in O($n$loglog$n$) time using SP [4][5].

A decoding scheme needs to be constructed in order to determine the relationship between polygons in a packing. A Longest Common Subsequence (LCS) algorithm is used in this paper to perform the decoding.

C. Longest Common Subsequnce

Longest common subsequence (LCS) is an algorithm used to reconstruct the placement of polygons from a SP. Each rectangle i is represented as an element in SP lists (X, Y), with a reference point (x_i, y_i) (bottom-left corner of a rectangle), width (w_i) and length (l_i). The width and height of a container can be found by using LCS, as well as (x_i, y_i) coordinates of a reference point of each polygon.

Let (X, Y) be represented as (X_1 b X_2, Y_1 b Y_2) so that:

- b is the rectangle whose (x, y) position is being calculated

```
Initialize P and L to zeros
For each element i
  b ← X(i)
  p ← index of element b in sequence Y
  P(b) ← L(p)
  t ← P(b) + width of element b
  For j = p to n
    If t>L(j)
      L(j) = t
    Else
      Break
    End if
LCS width/length = L(n)
```

Figure 2. The Fast LCS algorithm

- (X_1, Y_1) are predecessor elements of b in (X, Y)
- (X_2, Y_2) are successor elements of b in (X, Y)

To obtain the length of a placement, LCS(X, Y) needs to be performed. The width of a placement is calculated as LCS(X^R, Y), where $X^R = (X_2^R\ b\ X_1^R)$ represents inverted X permutation.

X-coordinate of a reference point of a rectangle is found by calculating LCS(X_1, Y_1), while the Y-coordinate is calculated by LCS(X_2^R, Y_1^R).

In this paper a Fast LCS algorithm (Fig 2) is used to determine the length of the longest common subsequence with respect to the dimension (width, length) of polygons that indexes in (X, Y) represent [6]. In Fig 4, X and Y are permutations of $\{1\ldots n\}$. The block position array P is used to record the x or y coordinate of rectangle b. The length array L is used to record the length of candidates of the longest common subsequence.

D. Bottom-left

Bottom-left (BL) is a heuristic placement routine that places polygons in a container in a series of downwards and leftwards sliding movements. It was first applied on a rectangle packing problem [7]. Starting from the top-right corner, a rectangle is moved vertically down until it touches the partial layout, and then horizontally to the left. A series of described BL moves is performed until a polygon cannot be moved more bottom and left, i.e. polygon is touching the current placement contour from its left and bottom side.

Improved BL methods such as BLLT and BLD were also developed for the rectangle packing problem [8]. They defer in a position in which the sliding direction changes down to left and vice-versa. A BLD inspired method is used in this paper. In BLD, the initial position of a rectangle is different from BL - rectangle starts sliding from the top-right corner of the previously placed rectangle. If this initial position is infeasible, the top-left position of a container is chosen as in BL procedure.

When dealing with packing of irregular items, sliding algorithms based on rectangles (BL_i, $BLLT_i$, BLD_i) and sliding algorithms based on polygons are examined in the literature (BLP_i, $BLDP_i$, BLF_i) [8]. Sliding algorithms based on rectangles are examined in this paper.

For each *irregular polygon*
Approximate polygon with a enclosing rectangle (ER)
Place the BB to the top-right corner of previously placed ER
If position is not valid
Start the bottom-left (BLi) routine

Repeat
Move ER vertically down to the partial layout
Move ER horizontally left to the partial layout
Until vertical AND horizontal movement not possible
Repeat
Move polygon vertically down to the partial layout
Move polygon horizontally left to the partial layout
Until a stable position is found

Until all items are placed

Figure 3. The BLD_i algorithm

Algorithms that handle packing of irregular polygons based on rectangles use a two-step packing procedure. In the first step, a bottom-left method is performed on a layout with enclosing rectangles using bottom left methods for rectangle packing problem (BL, BLLT or BLD).

Once stable positions for all enclosing rectangles have been found, a second step of the routine is performed using only irregular polygons. Each polygon is then nested using iterative horizontal and vertical moves, starting from the current position in a polygonal layout until a stable placement is found [8].

In this paper the influence of BL_i and BLD_i method on a layout density is examined (Fig. 3). These methods defer in the initial position from which a rectangle begins the sliding routine. In BL_i the rectangle starts sliding from a top-right corner of a container, while in BLD_i rectangle starts sliding from a top-right corner of a previously placed rectangle. If this position is infeasible, a top-right corner of a container is chosen as initial position.

E. Simulated annealing

Simulated annealing (SA) is a local search technique that searches a neighborhood of solutions randomly. A neighborhood is a set of solutions in proximity of current solution, and can be defined based on problem definition. The method is inspired by process of cooling material in a heat bath, a process known as annealing.

Local search approaches often end up in local optimal solutions, since they only move to better solutions in the neighborhood. SA allows some downhill movements to achieve global optimality, by moving to a worse solution depending on a control parameter (the temperature) [9]. Downhill movement is used as a mechanism to better explore the solution space, and escape local optima.

If a solution X^{i+1} has a lower fitness value than X^i, a worse solution is accepted if:

$$r \le e^{\left(f\left(X^{i+1}\right)-f\left(X^{i}\right)\right)/T}, \qquad (3)$$

where $r \in [0,1]$. Temperature T is decreasing in each iteration according to a cooling schedule. The probability of a downward move is large in the beginning, but decreases as it gets closer to optimum later in the algorithm [10].

F. Genetic algorithm

Genetic algorithm (GA) is a heuristic algorithm based on evolutionary computation. GA mimics nature's evolution principle that describes the survival of the fittest. GA author, John H. Holland, presented the algorithm in the 1970s motivated by Darwin's theory of evolution. GA uses evolutionary methods to search the solution space in order to find a solution to the optimization problem [11].

Simulated annealing starts from an initial feasible solution from which a sequence of feasible solutions is obtained by applying a heuristic based on a neighborhood search technique. In the other hand, genetic algorithm starts with an initial population of feasible solutions. Then solutions from the initial population are mated to produce children (recombination).

Recombination operation specifies how children are produced. Usually, two feasible solutions are taken from the population, and are used to produce two children which inherit properties of the two parents. A simple recombination operation is called crossover.

When children are obtained, some type of mutation is allowed to occur, which corresponds to the neighborhood search.

This process produces the next generation of the population. The process can be iterated for as many generations as desired [10].

G. Proposed methodology

1) GA-BLD_i algorithm

Initially, a random population of 50 permutation sequences (i.e. individuals) in genetic algorithm is created, where each permutation determines the order in which the polygons are packed.

Irregular polygons are first circumscribed with a rectangle. These approximations are often called bounding box (BB). BLD_i heuristic packs rectangles first. Rectangles create a partial layout. Each new rectangle starts sliding from the top-right corner of a previously placed rectangle. If this position is infeasible, a BL procedure for rectangles is performed.

When a stable position for a rectangle is found, the second stage begins. Starting from a stable position of its corresponding rectangle, the irregular polygon continues to perform a bottom-left movement to nest better into the partial layout. Fitness function is calculated according to (6), (7) and (8).

Selection is performed by *stochastic universal sampling* (SUS). Each individual is mapped on a contiguous line segment based on their fitness values or selection probability obtained according to their fitness values. If N_i is the number of individuals to be selected, N_i pointers are distributed evenly over the line with distance $1/N_i$. N_i individuals to which the pointers point are then selected for mating [12]. In Fig 4, an example is given

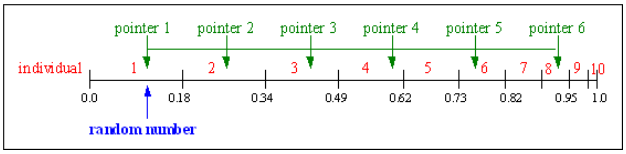


Figure 4. Stochastic universal sampling [12]

with 10 individuals in a population mapped on a line with size 1, based on their selections probability, and 6 evenly distributed pointer representing six individuals to be selected for mating process.

Recombination is performed using *partially matched crossover* (PMX) on consecutive pairs of individuals *P* and *Q*. Two randomly chosen crossover points split the individuals to three parts: $P_1P_2P_3$ and $Q_1Q_2Q_3$. The middle labels are exchanged to create new offspring: $P_1Q_2P_3$ and $Q_1P_2Q_3$. If P_2 consists of $P_{21}P_{22}\ldots P_{2n}$, as well as Q_2 consists of $Q_{21}Q_{22}\ldots Q_{2n}$, the duplicate labels problem is resolved by mapping of labels Q_{21} into P_{21}, Q_{22} into P_{22}, …, Q_{2n} into P_{2n} in the whole individual [13].

Insert mutation (Fig. 5) is performed on the offspring population by selecting two alleles at random, moving the second one next to the first one and shifting the rest of the alleles to accommodate. Insert mutation preserves most of the order [14].

2) SA-BL$_i$ algorithm

First, the SA parameters are initialized, and bounding boxes for irregular polygons are created. Rectangle packing is obtained using sequence pair.

An initial SP is created randomly. Neighborhood of a current SP is found by modifying at least one SP sequence by permuting the order of two elements in a sequence. Fitness function is calculated according to (6), (7) and (8).

Sometimes even a worse solution may be accepted in order to better explore the search space and escape local optima, regarding a certain probability:

$$r < e^{-D/T} \tag{3}$$

$$T = \frac{1}{t_0 + t_s \cdot a} \tag{4}$$

$$D = \frac{f(X_1, Y_1) - f(X, Y)}{f(X, Y)} \tag{5}$$

where $r \in [0, 1]$ is a random number, T is a temperature factor, t_0 is the initial temperature, t_s is temperature step, a is the current iteration, and D is the fitness ratio factor.

After rectangle packing has been obtained, a second phase begins, where a stable positions for polygons are found. First, the order of polygons is read from a rectangle layout. Starting from the bottom of the container, reference points of rectangles are read from left to right. If two polygons have the same *y*-coordinate of their reference points, polygon with lower *x*-coordinate has the priority while obtaining the order. Each polygon is then nested in the obtained order as described in section II.D.

3) Fitness function

Each solution is evaluated according to:

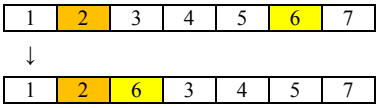


Figure 5. Insert mutation

$$\mathrm{f}(X, Y) = 1 - polyArea / bbArea \tag{6}$$

$$polyArea = \sum_{i=1}^{n} area(p_i) \tag{7}$$

$$bbArea = \mathrm{LCS}(X, Y) \cdot \mathrm{LCS}(X^R, Y) \tag{8}$$

where (X, Y) represents a sequence pair, *polyArea* is the sum of areas of all polygons, and *bbArea* is the area of a rectangle that circumscribes the layout, i.e. rectangles obtained by SP.

III. RESULTS

A. Environment

Experiments have been carried out on a personal computer with the following configuration: Intel Core i5, 4 GB DDR3, 500 GB HDD. Application have been implemented in MATLAB environment. GEATbx toolbox has been used for GA implementation [12].

B. Parameters

1) SA parameters

Parameters for simulated annealing have been chosen empirically and are as follows: initial temperature $t_0 = 0.1$, and temperature step $t_s = 0.01$.

2) GA parameters

One population consisted of 50 individuals is used in reproduction process of genetic algorithm in 100 generations. Best individual found upon this termination criterion is the considered as the best solution. Individuals are represented as permutations. Stochastic universal sampling (SUS), with selection pressure 1.6, and generation gap of 0.98 were used for the selection process. The value of selection pressure determines the fitness assignment and is used by the ranking algorithm, since fitness assignment is always done by ranking. Generation gap determines the fraction of the population to be reproduced in every generation [12]. In this paper 49 new individuals will be produced (0.98×50=49). Partially matched crossover (PMX) was used as recombination operator. Insert mutation was used as the mutation operator.

C. Benchmark datasets

A series of experiments have been conducted on 4 benchmark datasets available on EURO Special Interest Group on Cutting and Packing (ESICUP) website [15]: ALBANO (ALB, 24 cutting parts), DAGLI (DAG, 30 cutting parts), MAO (MAO, 20 cutting parts), MARQUES (MAR, 24 cutting parts), and all of which are instances taken from the textile industry. Experiments with rotation have not been included in this research.

D. Results

In this research two algorithms have been implemented, both of which pack polygons in two steps. In the first step, a layout using rectangular approximations of polygons is found, while in the second step a nesting procedure is applied on irregular polygons in order to tighten the layout. The goal of this research has been to obtain a dense polygon layout.

The results are shown in Table I. In the first column the datasets described in section III.C are defined. In the second column the results for GA-BLD$_i$ algorithm are presented. In the third column the results for SA-BL$_i$ algorithm are shown.

For each algorithm ten run of experiments have been performed. In column AVG/MAX the average/maximum of ten measurements is presented, while in column SD the standard deviation of 10 measurements is shown. The best results (MAX) for GA-BLD$_i$ are shown in Fig 6.

In GA-BLD$_i$, 4901 fitness function evaluations (Number of individuals + (Number of generations-1) × Number of individuals × Generation gap) have been performed, while in SA-BL$_i$ 10000 fitness function evaluations have been performed.

E. Discussion

1) Two steps approach reasoning

The second step is important in order to perform a dense packing with polygons touching. The reason for that lies in the lacking quality of rectangular approximation of polygons, i.e. large waste areas may occur between the polygon shape and its rectangular approximation. That may cause that polygons may have large waste space between each other, or not even touch. It can be said that the first step merely determines the initial position of the polygons while assuring they are all placed within the container without overlapping, while the second step tightens the layout.

2) Difference between the two methods

When SA-BL$_i$ is performed, the goal is to find a rectangle with minimal area that encloses rectangular items that are being packed, i.e. width and length of the container may be arbitrary. That is the reason why restriction on the width of the enclosing rectangle is not set in this research.

In GA-BLD$_i$ the width restriction is set for each dataset as it is defined in set definitions [15]: 4900 for ALBANO, 104 for MARQUES, 60 for DAGLI and 2550 for MAO. The goal is to minimize the area of the enclosing rectangle by minimizing its length.

The results indicate that, even though the SA-BL is more flexible, without restrictions on width nor length, GA-BLD$_i$ outperforms SP in average.

3) Results comparison

In implemented algorithms two versions of bottom-left strategies have been used. The reason why bottom-left strategies have been chosen for polygon nesting procedure is because polygons could be nested starting from the positions of their rectangle approximations, while maintaining similar position in terms of relationship between polygons. For instance, it is impossible for two polygons to swap positions using bottom-left.

Although the results suggest that GA-BLD$_i$ outperformed SA-BL$_i$ (in average), this hypothesis has to be viewed within the metaheuristic contexts, i.e. BL$_i$ has been used with SP and SA, while BLD$_i$ has been used with GA. The results may suggest that GA might be more appropriate than SA for solving packing problem with bottom-left strategies.

Previous research has indicated that SA-BL$_i$ solutions improve with the number of iterations, but the fact that GA-BLD$_i$ provided better results while performing less than half fitness function evaluations may lead to conclusion BLD$_i$ performs better than BL [16]. Additionally, better targeted experiments need to be run in order to verify both of these suspicions.

An overview table with best obtained results from the literature can be found in [17]. The best obtained result for ALBANO is 88.16%, for DAGLI 87.99%, for MAO 85.15%, and for MARQUES 89.82%. Even though our results are not as competitive as the ones in the literature it has to be noted that most of the algorithms include polygon rotation which allows them to produce higher density results.

IV. CONCLUSION

In this research a dense packing layout on four datasets taken from the textile industry has been obtained.

Two algorithms have been created. The SP-BL$_i$ algorithm hybridized Sequence Pair abstract representation method and BL$_i$ method to obtain a dense packing. Also, the GA-BLD$_i$ algorithm hybridized genetic algorithm and a modified BLDi approach.

The GA-BLD$_i$ version performed the strip packing problem, with limited width of a container and performed better on all datasets in average results. The obtained layout density is 78.23% for ALBANO, 70.70% for MAO, 70.15% for DAGLI and 77.63% for MARQUES.

The SP-BL$_i$ method did not have limitations on width, nor height of a container but has still not been able to outperform the GA-BLD$_i$ solution. The obtained layout density is 76.28% for ALBANO, 68.52% for MAO, 63.10% for DAGLI and 75.60% for MARQUES.

The algorithm using BLD$_i$ performed better than the one using BL$_i$ as the results might suggest. The influence of GA and SA on the layout needs to be additionally

TABLE I. RESULTS

DATA	GA-BLD$_i$			SA-BL$_i$		
	AVG	*SD*	*MAX*	*AVG*	*SD*	*MAX*
ALB	**78.23**	1.44	**80.42**	76.28	1.65	78.72
DAG	**70.15**	0.71	**71.66**	63.10	1.51	65.82
MAO	**70.70**	0.94	**72.34**	68.52	2.04	71.45
MAR	**77.63**	1.19	79.50	75.60	2.98	**80.25**

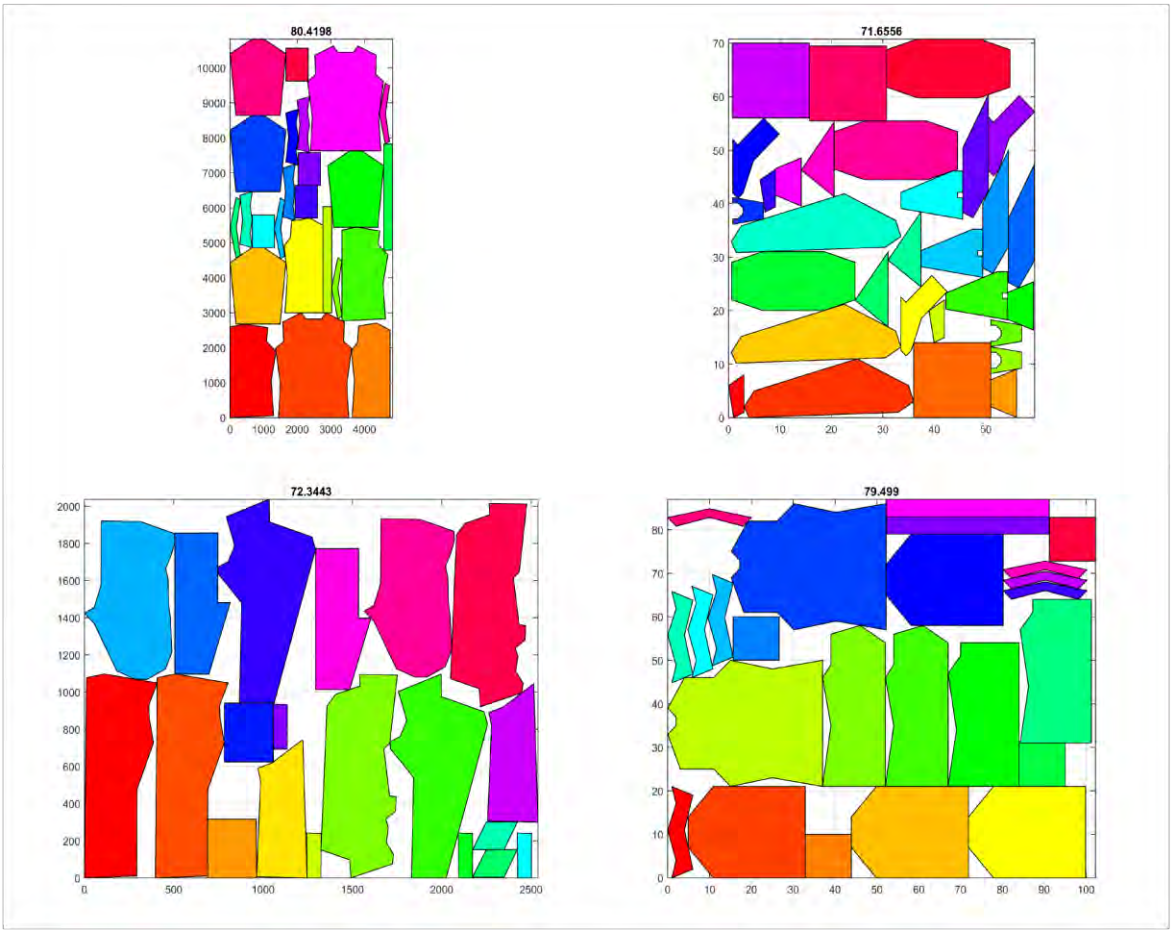


Figure 6. The best results from column MAX for GA-BLDP algorithm
(top-left: ALBANO, top-right: MAO, bottom-left: DAGLI, bottom-right: MARQUES)

determined in the better targeted experiments.

ACKNOWLEDGMENT

This work has been fully supported by Croatian Science Foundation under the project number 3011 Application of mathematical modeling and intelligent algorithms in clothing construction.

REFERENCES

[1] J. Egeblad, "Heuristics for Multidimensional Packing Problems," Københavns Universitet, Faculty of Science, Datalogisk Institut, Department of Computer Science, 2008.

[2] G. Wäscher, H. Haußner, and H. Schumann, "An improved typology of cutting and packing problems," *European Journal of Operational Research*, vol. 183, no. 3, pp. 1109–1130, Dec. 2007.

[3] H. Murata, K. Fujiyoshi, S. Nakatake, and Y. Kajitani, "VLSI module placement based on rectangle-packing by the sequence-pair," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 15, no. 12, pp. 1518–1524, 1996.

[4] Wong, D.F. and X. Tang, "FASTSP: A Fast Algorithm for Block Placement based on Sequence Pair," in *Design Automation Conference, 2001. Proceedings of the ASP-DAC 2001. Asia and South Pacific*, Yokohama, 2001, pp. 521–526.

[5] D. Pisinger, "Denser Packings Obtained in O (n log log n) Time," *INFORMS Journal on Computing*, vol. 19, no. 3, pp. 395–405, Aug. 2007.

[6] X. Tang, R. Tian, and D. F. Wong, "Fast Evaluation of Sequence Pair in Block Placement by Longest Common Subsequence Computation," in *Design, Automation and Test in Europe Conference and Exhibition 2000. Proceedings*, Paris, 2000, pp. 106–111.

[7] S. Jakobs, "On genetic algorithms for the packing of polygons," *European Journal of Operational Research*, vol. 88, no. 1, pp. 165–181, 1996.

[8] E. Hopper, "Two-dimensional packing utilising evolutionary algorithms and other meta-heuristic methods," Doctoral Thesis, University of Wales. Cardiff, 2000.

[9] A. M. Gomes and J. F. Oliveira, "Solving Irregular Strip Packing problems by hybridising simulated annealing and linear programming," *European Journal of Operational Research*, vol. 171, no. 3, pp. 811–829, Jun. 2006.

[10] D. L. Kreher and D. R. Stinson, "Combinatorial Algorithms: Generation, Enumeration, and Search," *SIGACT News*, vol. 30, no. 1, pp. 33–35, Mar. 1999.

[11] M. Golub, "Genetski algoritam - prvi dio," 2004. [Online]. Available: http://www.zemris.fer.hr/~golub/ga/ga_skripta1.pdf. [Accessed: 24-Feb-2014].

[12] "GEATbx - Genetic and Evolutionary Algorithms Toolbox in Matlab - Main Page." [Online]. Available: http://www.geatbx.com/. [Accessed: 21-Feb-2014].

[13] "GEATbx for Matlab (v. 3.8): Function recpm." [Online]. Available: http://www.geatbx.com/docu/recpm.html. [Accessed: 18-Feb-2014].

[14] A. E. Eiben and J. E. Smith, *Introduction to Evolutionary Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015.

[15] "ESICUP : ESICUP." [Online]. Available: http://paginas.fe.up.pt/~esicup/tiki-index.php. [Accessed: 29-Jan-2015].

[16] D. Domović and T. Rolich, "Hybridizing Sequence Pair and Bottom-Left Heuristic on Solving 2D-Irregular Strip Packing Problem," in *15th World Textile Conference AUTEX 2015*, 2015.

[17] M. A. Shalaby, "A Particle Swarm Optimization Algorithm for a 2-D Irregular Strip Packing Problem," *American Journal of Operations Research*, vol. 3, no. 2, pp. 268–278, 2013.

Automatic Image Annotation Refinement

M. Pobar, M. Ivašić-Kos
Department of Informatics, University of Rijeka, Rijeka, Croatia
mpobar@uniri.hr, marinai@uniri.hr

Abstract - Automatic image annotation methods automatically assign labels to images in order to facilitate tasks such as image retrieval, search, organizing and management. Incorrect labels may negatively influence the search results so image annotation should be as accurate as possible. Labels pertaining to objects or to whole scenes are commonly used for image annotation, and precision is especially important in case when scene labels are inferred from objects, as errors in the object labels may propagate to the scene level. One way to improve the annotation precision is by detecting and discarding the automatically assigned object labels that do not fit the context of other detected objects. This procedure is referred to as annotation refinement. Here, an approach to detection of likely incorrect labels based on the context of other labels and prior knowledge about mutual occurrence of various objects in images is tested.

I. INTRODUCTION

To facilitate image organization and search, automatic image annotation methods automatically assign labels from a predefined vocabulary to unlabeled images. Labels pertaining to objects present in images like *beach umbrella, sand, sea, buoy*, or to whole scenes like *beach in summer* are commonly used.

In general, a scene may be very complex and can be composed of many different objects, so it should also be annotated with many object labels. The image annotation task is closely related to the problem of classification or the problem of representing the correlations between images and labels, so the most of the automatic image annotation approaches proposed so far belong to the field of machine learning. The methods based on classification, like [1], classify images or image segments into predefined classes based on low-level features extracted from images. The automatic recognition of objects depends on the automatic annotation method used and on the features extracted from images.

In this case, the labels assigned to an image are usually treated as mutually independent. In reality, some objects typically occur together in scenes, while others appear together only very rarely or never, and are thus not actually independent. The knowledge about the context and about the frequency with which objects occur together in certain types of scenes should be thus exploited to improve the precision of automatic annotation. This could be done by detecting labels that are likely the result of misclassification and can be discarded.

For exploring the correlations between annotated labels graph-based image analysis algorithms have been recently proposed [2-8]. To detect and discard the irrelevant labels, a WordNet based semantic similarity is used in [8]. A graphical model is used in [5] for fusing visual content represented by a nearest spanning chain and label correlation by WordNet. A group of authors has examined several different approaches for annotation refinement: Markov chains in [4], conditional random fields (CRF) in [2], normalized Google distance (NGD) in [8] and re-ranking the annotations using the random walk with restarts algorithm in [3].

Incorporating knowledge into automatic image annotation procedure proved as a promising approach for improvement of annotation efficiency. Such an approach was proposed in [9] where a fuzzy knowledge base and inference algorithms are incorporated into automatic image annotation system for multi-level image annotation. In [10] a multi-label image annotation system was proposed with an inference based refinement procedure included as a step. In [11] image annotation on object level is refined by discarding or replacing object labels that do not fit the context. For that purpose, fuzzy inheritance and intersection algorithms are used.

In this paper, we propose an annotation refinement procedure that uses knowledge about co-occurrence frequency of objects on scenes, presented in Section 3. The procedure is not dependent on the algorithm that was used for image classification; it is a subsequent correction of obtained classification results. The procedure is tested on images from the outdoor image domain, with experiments using automatic annotation and simulated annotation sets for cases with one or more misclassified labels or when some of labels are replaced with wrong ones, as described in Section 4. In Section 5 the results of image annotation before and after refinement are presented. Concluding remarks are given in Section 6.

II. AUTOMATIC ANNOTATION PIPELINE WITH ANNOTATION REFINEMENT STEP

A pipeline for automatic image annotation system with the proposed refinement step is shown in Fig. 1. The main stages are feature extraction, multi label classification and annotation refinement. It is assumed that an unlabeled image is input to the automatic annotation system and the outputs are labels that are refined using the proposed algorithm.

First, low-level features such as color, texture, etc. are extracted from the image, which is then represented with a feature vector. The feature vector is then input into a trained classifier to obtain the image annotation. In the most general case, an image $e \in E$ could be annotated with many labels from the set of all labels L, and then a multi-label image classification approach should be used to obtain the annotation set $A(e) = \{a_1, a_3, .., a_p\}, a_i \in L, p \leq |L|$. Each label a in the annotation set $A(e)$ is associated with a weight $\gamma(a)$ so the fuzzy set

$(A, \gamma|e), \gamma: A \rightarrow [0,1]$ is denoted by $\{\gamma(a_1)/a_1, \gamma(a_3)/a_3, .., \gamma(a_{17})/a_p\}$. Weight can be determined in several ways, for example according to the confidence value of the classifier or the relative frequency of object a in the training set, or can be set to a value in the interval [0,1].

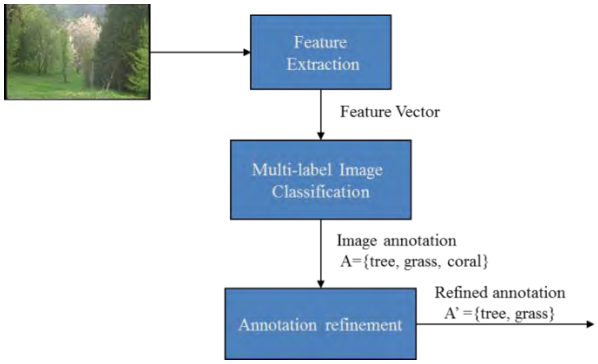


Figure 1. Automatic image annotation pipeline with annotation refinement

For example, classification result of Naive Bayes classifier suited for multi-class classification problems for an unlabeled image e in Fig. 2 makes the annotation set $A(e) = \{0.91/tree, 0.68/grass, 0.5/coral\}$. Note that the label *coral* is a result of misclassification because it is not present in the image.



Figure 2. Example of an unlabeled image

Most automatic annotation systems stop at this stage. However, to detect labels that are likely misclassified one can use the additional information that is not used in the classification step as the general knowledge about objects and scenes or some spatial or topological information that could be extracted from a set of images. This step is referred to as annotation refinement step.

The proposed procedure for annotation refinement is based on context of obtained labels and prior knowledge about mutual occurrence of objects in scenes. Object labels with low confidence values that do not commonly occur with other objects in the annotation set are discarded.

III. ANNOTATION REFINEMENT

The image annotation process relies on models that capture relations between features extracted from images and corresponding labels assigned to the image. Misclassification occurs commonly since different objects can have very similar values of image features and therefore the classification models cannot be learned well.

To reduce the influence of misclassification and to annotate an image as precisely as possible additional knowledge should be exploited in annotation refinement step. To detect and discard false and spurious labels the information about mutual occurrence of different objects in different contexts may be captured and used as will be explained hereinafter.

Knowledge about mutual appearance of objects in scenes is gained from the images in the training set $E' \subset E$. For each image $e \in E/E'$ in the test set, the image annotation set $A(e)$ is obtained after multi-label classification of image e. Then the consistency of labels in image annotation set $A(e)$ is checked using the proposed annotation refinement approach.

The annotation refinement procedure follows the general steps shown in Fig. 3.

Assumptions:
L – set of all labels;
$f: L \rightarrow \{1..|L|\}$; - a bijective function for enumerating L;
$\boldsymbol{R} = (R_{ij}) \in \mathbb{N}^{|L|\times|L|}$co-occurrence matrix, R_{ij} is the number of co-occurrence of labels $f^{-1}(i) \in L$ with $f^{-1}(j) \in L$;

Input:
Initial annotation set with corresponding weights $(A, \gamma|e)$

Steps:
define a power set $\mathcal{P}(A)$ for initial annotation set A
initialization $A' := \emptyset,\ r := 0, V' = \emptyset$
for each $S \in \mathcal{P}(A)$:
compute $V_j := sign(\prod_{a\in S} R[f(a), j]; j = 1..|L|)$
if $V_j = 1$ $V' = V' \cup f^{-1}(j)$
if $S \subseteq V'$ compute $RV(S) := \sum_{a\in S} \gamma(a)^2$
if $RV(S) > r$
set $r := RV(S)$
set $A' := S$
Output:
The set A' of all consistent labels from A

Figure 3. Annotation refinement procedure

The input to the algorithm are the classification labels paired with their weights. In this experiment, we used three different ways to set the weights:

- all labels are associated with the same weight,
- relative frequency of each label in training set,
- confidence values obtained with a classifier.

For each initial annotation set A the algorithm computes its power set $\mathcal{P}(A)$ (the set of all subsets of A, including the empty set and A itself) and according to the prior knowledge stored in $\boldsymbol{R} = (R_{ij})$ determines which of the subsets has the highest likelihood to appear as image annotation set for a given image. The largest subset with the highest likelihood will be selected as refined image annotation set A'.

If all the labels in a set A are mutually consistent, the whole set A will be assigned to the refined annotation set A'.

If a set A contains a label that is inconsistent with other labels in the set, the power set will contain subsets that

have never appeared as annotation sets for images in the test set. These subsets will have low reliability and will not be selected for annotation.

If all the labels in a set A are mutually inconsistent, only the one label that has the highest weight will be selected and assigned to refined annotation set A'.

We have implemented the proposed annotation refinement algorithm in Python programming language.

IV. EXPERIMENTS

To test the proposed annotation refinement procedure, we have compared the results of image annotation with and without refinement.

Results of image annotation were determined in two ways. In the first case, reference annotation set corresponds to the classification results obtained by Naive Bayes classifier on low-level features extracted from images (referred to as *NB scenario*). In the second case, the reference annotation set is formed so that the ground-truth set is modified to simulate various annotation scenarios that could be obtained with some other classification methods (referred to as *simulated scenarios*).

A. Data Set

The annotation experiments were performed on a part of the Corel image database [12] related to outdoor scenes. Each image in the Corel dataset was described with more than one label from a vocabulary of 54 keywords related to natural and artificial objects such as 'airplane', 'bird', 'lion', 'train' etc. and background objects like 'ground', 'sky', 'water' etc. Some labels were too rare to effectively train the classifiers, so the images that correspond to those labels were excluded from data. The resulting data consisted of 397 images labelled with labels from the set L of 22 labels, $|L| = 22$.

B. NB scenario

We have used the Naïve Bayes (NB) classifier with oversimplified independence assumption for the task of image annotation in the NB scenario. Since multiple labels should be assigned to each image, the problem we were dealing with is a multi-label classification problem.

Multi-label classification approaches mainly fall into two categories: problem transformation and algorithm adaptation [12]. One of the method of former approach is binary relevance, also known as one-versus-all or one-versus-rest strategy. This strategy was used in this scenario. It is a standard technique used to solve multi-label problems with binary classifiers like NB. Its main idea is to train K different binary classifiers on the multi-label training set with $K = |L|$ possible classes, each one trained to distinguish one of K classes from the remaining K-1 classes. Each trained classifier then decides whether an image belongs to its class or not. In that manner, the NB is suited for multi-class classification problems and the overall classification result for an image contains all class labels assigned to that image.

The NB was trained using a maximum likelihood parameter estimation from data in the training set. To capture the variety of perceptual and semantic information from images, we have used pixel-based and structure-based descriptors. All features were extracted from images that were sized 128 x 192 pixels or 192 x 128 pixels.

The pixel-based color descriptor is made up of global and local dominant colors, and of color histograms and color moments, similarly as in [10].

The color histogram was calculated for each of the RGB color channels of the whole image to determine the global dominant color. The histogram bins with the highest values for each channel were selected and they correspond to global dominant colors in decreasing order. As in [11] we have used 12 dominant colors per channel in each image as features for our classification tasks. The local dominant colors are calculated in the same manner but for each tile of 3x1 grid placed across each image (Fig. 5) and separately for central part of the image and for surroundings. The overall size of global and local dominant color vector is 216.

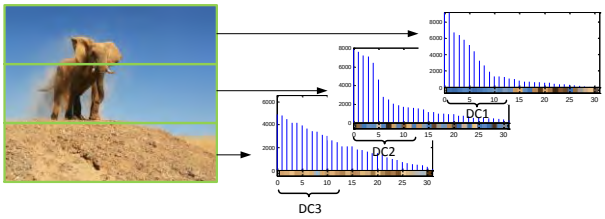


Figure 4. Computation of local dominant colors (DC1..DC3) from three image regions.

We have additionally computed the color moments (mean, standard deviation, skew and kurtosis) for each of the RGB channels. The size of the color moment feature vector is 12.

For representing dominant spatial structure of the scene, a structure-based image descriptor GIST [14] was used. The spatial properties are estimated using global features computed as a weighted combination of Gabor-like multi scale-oriented filters. In our case, we used 4x4 encoding samples within 8 orientations per 4 scales of image components, so the GIST feature vector has 512 components.

Because of high dimensionality of the feature set and low number of examples, the NB classifier could not be trained effectively without dimensionality reduction. We have thus used PCA to reduce the number of dimensions from 740 to 30, with the transformed features explaining 60% of the variance in the original feature set.

C. Simulated annotation scenarios

In the second case, annotation sets are formed by modifying the ground-truth set. The aim was to simulate various annotation scenarios, different from those that we got using a selected classifier. This is in order to examine the effects of refinement with different types of errors that classifier makes and to examine the robustness of the annotation refinement algorithm.

Two broad cases have been considered for defining various annotation scenarios, one in which new randomly selected labels are added to the existing annotation set (ground-truth) and second in which the existing label in

annotation set is replaced with new randomly selected label.

In the first group of annotation sets, marked SS1, one additional random incorrect label was added to each ground truth set, as in example:

Ground-truth $R(e)$	*Annotation set (SS1)* $R(e) \cup \{a\};\ a \in L/R(e)$
grass, sky, trees	grass, sky, trees, **rock**
Sea, dolphin	Sea, dolphin, **bear**

In the second group, marked SS2, two additional random incorrect labels were added to the ground truth set, for example:

Ground-truth $R(e)$	*Annotation set (SS2)* $R(e) \cup \{a_1, a_2\}; a_1, a_2 \in L/R(e)$
grass, sky, trees	grass, sky, trees, **rock, sand**
sea, dolphin	sea, dolphin, bear, **cloud, tiger**

Given that most of the images were annotated with three labels, we tested the cases of one or two added random labels to the ground-truth set of labels. If more labels were added that are incorrect, the majority of the labels in the annotation set would be incorrect and the algorithm could not determine which of the input labels are relevant for the image. The output would in that case be a variety of refined annotation sets that are consistent but not necessary related to a given image.

The last group of annotation sets, marked SSr, contained one random incorrect label in place of a correct label in the ground truth, as in example:

Ground-truth $R(e)$	*Annotation set (SSr)* $R(e) \cup \{a_1\}/\{a_2\}$, $a_1 \in L/R(e), a_2 \in R(e)$
grass, sky, trees	grass, **coral**, trees
sea, dolphin	sea, **flower**

As in case of adding labels, by replacing more than one label in the ground truth set the majority of labels would be incorrect for an image. Again, the output of algorithm would be refined annotation set that may not be related to the image, so this case was not tested.

Since in the simulated scenarios there were no actual confidence values to be used, the relative frequency of labels in the training set was used for label weights. In addition, the algorithm was tested with weights of all labels set to 1.

V. Experimental Results

We evaluate the difference in classification performance before and after refinement in terms of precision, recall and F1 score as instance- and label-based evaluation measures [13]. The instance-based evaluation measures are based on the average differences of the actual and the predicted sets of labels over all examples in the test dataset. The label-based or macro-averaged evaluation measures assess the predictive performance for each label separately and then average the performance over all labels [13]. These measures are used due to the fact that in case of multi-label classification an instance may not only be correctly or incorrectly annotated, but also partially correctly. For example, if an image *e* should be annotated with $R(e) = \{beach, sky, sea\}$ and is automatically annotated with $A(e) = \{cloud, sky, sea\}$, then the evaluation measure should reflect the insertion of wrong labels ($cloud \in A(e) \wedge cloud \notin R(e)$), missing labels ($beach \notin A(e) \wedge beach \in R(e)$) and correct labels ($sky, sea \in R(e) \cap A(e)$).

To define the evaluation measures, we assume that an instance $e_j \in E, j = 1..N$ should be classified into the set of true class labels $R(e_j) = R_j = \{C_l, C_m, \dots, C_r\}$, $R_j \subseteq C$ where E is a set of image feature vectors, C is a set of all class labels. For an example e_j, the set of labels that are predicted by a classifier is denoted as Z_j. In case of single-label classification $|R_j| = |Z_j| = 1$.

Instance based precision is defined as the average ratio of correctly assigned and all labels assigned to each example by the classifier:

$$Pr_{ins} = \frac{1}{N}\sum_{i=1}^{N}\frac{|R_i \cap Z_i|}{|Z_i|} \quad (2)$$

Instance based recall is defined as the average ratio of labels correctly assigned by the classifier and all labels of each example (ground truth):

$$Re_{ins} = \frac{1}{N}\sum_{i=1}^{N}\frac{|R_i \cap Z_i|}{|Y_i|} \quad (3)$$

Instance based F-Measure is the harmonic mean of Precision and Recall and can be interpreted as a weighted average of the precision and recall:

$$F1_{ins} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|R_i \cap Z_i|}{|Z_i| + |R_i|} \quad (4)$$

These measures reach their best value at 1 and the worst at 0.

Label based measures are computed firstly by computing the instance-based measure and then averaging over all labels.

If the majority of labels in the annotation set contains correctly classified labels, and the incorrect label is inconsistent with the others, the average precision of the image annotation can be increased, Fig. 7a, 7b. If the majority of labels are correct, and the incorrect labels are consistent with them, the incorrect label cannot be detected, fig. 7c. If the majority of labels are misclassified, a set of wrong but mutually consistent labels might be chosen. In that case the annotation refinement can discard the correct labels and both recall and precision falls, Fig. 7d. In the Fig. 7, incorrect labels are printed bold.

	(a)	(b)
Image e		
$R(e)$	*sky, wolf, trees, grass*	*airplane, sky, cloud*
$A(e)$	***coral**, sky, wolf, trees, grass*	*airplane, **dolphin**, sky, cloud*
	$Pr = \frac{4}{4+1} = 0.8, Re = \frac{4}{4} = 1$	$Pr = \frac{3}{3+1} = 0.75, Re = \frac{3}{3} = 1$
$A'(e)$	*sky, wolf, trees, grass*	*airplane, sky, cloud*
	$Pr = \frac{4}{4+0} = 1, Re = \frac{4}{4} = 1$	$Pr = \frac{3}{3+0} = 1, Re = \frac{3}{3} = 1$

Figure 5. Positive (a), (b) examples of annotation refinement.

	(c)	(d)
Image e		
$R(e)$	*sky, lion, trees, grass*	*shuttle, astronaut*
$A(e)$	*sky, lion, trees, grass, **rock***	*shuttle, **train, building***
	$Pr = \frac{4}{4+1} = 0.8, Re = \frac{4}{4} = 1$	$Pr = \frac{1}{1+2} = 0.33, Re = \frac{1}{2} = 0.5$
$A'(e)$	*sky, lion, trees, grass, rock*	***train, building***
	$Pr = \frac{4}{4+1} = 0.8, Re = \frac{4}{4} = 1$	$Pr = \frac{0}{0+2} = 0, Re = \frac{0}{2} = 0$

Figure 6. Neutral (c) and negative (d) example of annotation refinement

In the Table 1, the annotation results before and after refinement are presented for the NB scenario.

TABLE I. EVALUATION MEASURES FOR THE NB SCENARIO

Measure type		*No refinement*	*Refinement, weight = confidence*	*Refinement, weight = 1*
Label-based	*Precision*	0,361	**0,372**	**0,371**
	Recall	0,641	0,637	0,640
	F1	0,462	**0,470**	**0,470**
Instance-based	*Precision*	0,223	**0,234**	**0,232**
	Recall	**0,341**	0,338	0,336
	F1	0,270	**0,277**	**0,275**

The results after refinement show slight improvement of F1 and precision measures. The results with using confidence or weight set to 1 are very similar.

The results on the same image set published in [9], have shown better improvement of precision over non-refined set, however, the method used to obtain the non-refined set was different, so no direct comparison can be made. In the following tables with different simulated conditions, the proposed algorithm shows comparable or better improvements.

Next, in the in the Tables 2-4 the results of annotation refinement are presented for the simulated classifier scenarios. In the Table 2, the results of annotation refinement with the annotation sets SS1 are presented. In this case, the simulated classifier outputs for each image contain all the true labels (perfect recall) and an additional random incorrect label. Two cases were considered for annotation refinement, when the label weights are set to their relative frequency in the training set, and when they are all set to the same value of 1.

TABLE II. EVALUATION MEASURES FOR SIMULATED SCENARIO AND ANNOTATION SETS SS1

Measure type		*No refinement*	*Refinement, weight = rel. frequency*	*Refinement, weight = 1*
Label-based	*Precision*	0,612	0,727	**0,807**
	Recall	**1,000**	0,782	0,873
	F1	0,740	0,738	**0,817**
Instance-based	*Precision*	0,735	0,842	**0,857**
	Recall	**1,000**	0,924	0,950
	F1	0,847	0,881	**0,901**

The results show improvement in precision and F1 scores with a decline in recall score for both label-based and instance-based measures, as was expected. Better improvement has been achieved when setting the weight of each label to 1 than when using the frequency of labels in the training data.

In the Table 3, the results for the annotation sets SS2 are presented. These sets are such that the simulated classifier outputs for each image all the true labels (perfect recall) and two additional random incorrect labels.

TABLE III. EVALUATION MEASURES FOR SIMULATED SCENARIO AND ANNOTATION SETS SS2

Measure type		*No refinement*	*Refinement, weight = rel. frequency*	*Refinement, weight = 1*
Label-based	*Precision*	0,473	0,600	**0,604**
	Recall	**1,000**	0,677	0,747
	F1	0,615	0,602	**0,644**
Instance-based	*Precision*	0,588	0,731	**0,732**
	Recall	**1,000**	0,878	0,895
	F1	0,740	0,798	**0,805**

The results show the same behavior as in the first case, with improvements in precision and F1 scores and a decline in recall score. Again, using relative frequencies of labels in the training set as weights did not improve the refinement over using the same weight for all labels.

The Table 4 presents the results for the final tested case, where the simulated classifier outputs for each image

all true labels except for one that is replaced with a random incorrect label (annotation sets SSr).

TABLE IV. EVALUATION MEASURES FOR SIMULATED CLASSIFICATION SCENARIO AND ANNOTATION SETS SSR

Measure type		*No refinement*	*Refinement, weight = rel. frequency*	*Refinement, weight = 1*
Label-based	*Precision*	0,503	0,553	**0,537**
	Recall	**0,603**	0,476	0,473
	F1	**0,526**	0,497	0,482
Instance-based	*Precision*	0,639	**0,722**	0,709
	Recall	**0,639**	0,592	0,583
	F1	0,639	**0,650**	0,640

Here, when considering the label-based measures, the precision is slightly improved with the refinement procedure, however the F1 score is lower than without refinement. Looking at the instance-based measures, precision is increased without reducing the F1 score. In this case however, better results are obtained using relative frequencies as weights, as opposed to setting them all to 1.

VI. CONCLUSION

In this paper, a procedure based on observed co-occurrence of objects in the training set is proposed for refinement of labels obtained with automatic image annotation. Statistics about objects and their common appearance in images is exploited to find consistent sets of labels among the results of automatic annotation and to discard the inconsistent labels.

The proposed annotation refinement procedure was experimentally tested a part of the Corel image database related to outdoor scenes. The results of image annotation with and without refinement were compared using the annotation sets obtained with the Naïve Bayes classifier in the first scenario, and in the second scenario, on annotation sets formed by modifying the ground-truth set in different ways to simulate different classifier errors that could be expected. Three simulated annotation sets were used, two in which one and two new randomly selected labels are added to the existing annotation set (ground-truth) and one in which the existing label in annotation set is replaced with a new randomly selected label.

In the Naïve Bayes case, the automated annotation of images was modelled as a multi-label classification task, where a single image may be labelled with more than one class label. As the usual single-label classification algorithms can't directly be used to solve the multi-label problem, data transformation along with the Naïve Bayes were used.

In both NB and simulated classifier scenarios, the precision of annotation increased after refinement, with the cost of reduction in recall measure.

In the the simulated scenario SSr where a correct label label from the ground truth was replaced with an incorrect label, the annotation sets sometimes may not contain enough true positive examples for the algorithm to correctly choose the consistent and true set. In this case, the precision was increased however the overall label-based F1 score didn't improve.

As was expected, the annotation refinement proved most useful when the classifier makes false positive detections in addition to detecting true positives; in that case the annotation refinement algorithm has enough correct labels to determine a consistent and true set of labels and to discard the label that doesn't fit the context. This corresponds to simulated scenarios SS1 and SS2, where the greatest increases in both precision and F1 measures were achieved.

In all cases, the precision was improved, and in most cases also both label-based and instance based F1 measures were increased as well, so it can be said that the proposed algorithm improved results in all the tested cases.

REFERENCES

[1] J. Li and J. Z. Wang, ``Real-Time Computerized Annotation of Pictures," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, 2008, pp. 985-1002.

[2] Wang, Y., Gong, S. (2007). Refining image annotation using contextual relations between words. ACM CIVR 07', July 9–11, Nethelands, 425–432.

[3] Wang, C, Jing, F., Zhang, L., & Zhang, H.-J. (2006). Image annotation refinement using random walk with restarts. ACM MM 06', October 23–27, Santa Barbara, California, USA. 647–650

[4] Wang, C., Jing, F., Zhang, L., Zhang, H.-J. Content based image annotation refinement. CVPR'07.

[5] Liu, J., Li, M., Ma, W.-Y., Liu, Q., & Lu, H. (2006). An adaptive graph model for automatic image annotation. ACM MIR 06', Santa Barbara, California, USA, October 26-27, 61–70.

[6] Zhou, X., Wang, M., Zhang, Q., Zhang, J., Shi, B. (2007). Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. ACM CIVR 07', July 9–11, Amsterdam, Nethelands, 425–432.

[7] Yohan, J., Khan, L., Wang, L., & Awad, M. (2005) Image annotations by combining multiple evidence and WordNet. ACM conference on Multimedia (MM05'), Singapore, 706–715.

[8] Cilibrasi, R. L., & Vitanyi, P. M. (2007). The google similarity distance.Knowledge and Data Engineering, IEEE Transactions on, 19(3), 370-383.

[9] Ivasic-Kos, M., Ipsic, I., Ribaric, S. (2015). A knowledge-based multi-layered image annotation system. Expert Systems with Applications, 42(24), 9539-9553.

[10] Ivasic-Kos, M., Pobar, M., Ribaric, S. (2016). "Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme", Pattern recognition, 52, 287-305

[11] Ivašić-Kos, M.; Pobar, I.; Ribarić, S. "Automatic image annotation refinement using fuzzy inference algorithms", 16th World Congress of the International Fuzzy Systems Association (IFSA) 9th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT), Gijon, Asturias, Atlantic press. 13 (2015); doi:10.2991/ifsa-eusflat-15.2015.205

[12] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D.M. Blei, M.I. Jordan, "Matching words and pictures", J.Mach.Learn.Res.3(2003), 1107–1135

[13] G. Tsoumakas, I. Katakis, "Multi-label classification: an overview", Int.J.Data Warehous. Min. 3(3) (2007) 1–13

[14] Oliva, A., & Torralba, A. (2001). „Modeling the shape of the scene: A holistic representation of the spatial envelope", *International journal of computer vision*, *42*(3), 145-175.

Defining Ontology Combining Concepts of Massive Multi-Player Online Role Playing Games and Organization of Large-Scale Multi-Agent Systems

Bogdan Okreša Đurić, Markus Schatten
Artificial Intelligence Laboratory
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2, 42000 Varaždin, Croatia
Email: (markus.schatten, dokresa)@foi.hr

Abstract— **Large-scale multi-agent systems (LSMAS) development methods are becoming a necessity in a world of ever growing numbers of computer agents incorporated in various devices and services used by people in all areas of life. Organization in such large-scale systems, becomes therefore a very relevant concept in research on LSMAS, as the mentioned agents are expected to cooperate, compete, and share information, hence improving quality of peoples' life. An example application domain of LSMAS are massive multi-player online games (MMOG), and their specific genre – MMO role-playing games (MMORPG). MMORPGs provide us with the opportunity to study both large-scale virtual interaction of players (agents), and coordination of large-scale distributed artificial intelligence. In this paper, we will present an ontology of organizational concepts, suited for LSMAS development, merged with an ontology containing elements of a specific MMORPG. Combining these two ontologies with a special accent on identifying elements and aspects of organization, will hopefully present us with a basis for LSMAS structures which follow organizational constraints and are suitable for LSMAS development in the MMORPG context.**

Keywords*— *MMORPG, LSMAS, organization, ontology, software development

I. Introduction

Organization has always been present in the lives of people, yet only the last few years show an emergent need for organization in the digital world. The number of interconnected digital devices is increasing on a daily basis, and their capabilities are developing rapidly. Such a rapid development led to the Internet of Things (IoT) and increased activity in the digital (virtual) environment [1]. This increase in numbers of various smart devices led to a noticeable increase in the number of heterogeneous computer software agents co-existing the digital world. Interconnected using the Internet and other available networks, these computer agents have to be able to communicate with each other, cooperate, compete and interact in various ways both mutually and with their human owners. Since these software systems are distributed and smart, multi-agent systems (MAS) become a natural abstraction for the development of such complex socio-cybernetic systems. The idea of using organizational concepts in multiagent systems is not new [2], [3], but has only recently been proposed for the use in LSMAS [4], [5].

Organization of large-scale multi-agent systems (LSMAS) has only recently had an attempt in formalizing applicable organizational design methods, presented in [6]. The ultimate objective of the mentioned research is seen in a formalized organizational design ontology that is to act as a metamodel for a to be established LSMAS development framework.

One of the application domains of LSMAS are massive multi-player online role playing games (MMORPG). MMORPG is a mixture of role playing (RPG) games (genre of computer games in which players control actions of a protagonist in a virtual world), and of massive multi-player online (MMO) games (computer games that allow for a great number of players to play the same game simultaneously online, often causing, or demanding, interaction amongst them), as is stated in [7]. MMORPGs are therefore a valuable source of data on the matter of organization and interaction of agents (real in the role of players or virtual in the role of non-player character (NPCs)). Since MMORPGs include a great number of players, these players can be digitalized, or the player characters can be given autonomy and systems for reasoning and interaction with the virtual environment of their world, thus creating a dynamic LSMAS. Characters of a game enhanced in the said manner gain basic elements of an agent [8, p. 34].

The objective of this work in progress paper is to present initial developments in creating a joint comprehensive ontology covering concepts included in the domain of LSMAS organization, and those of a specific MMORPG ontology developed for the purposes of the *Large-scale Multi-Agent Modelling of Massively Multi-player On-line Role-Playing*

Games - ModelMMORPG[1] project. This domain ontology is particularly bound to The Mana World[2], which is an open-source MMORPG and subject to our study. The long-term objective of this research is a continuation on the work presented in [6], [9], [4], and aims to create a metamodel for organization-based LSMAS development framework. We hope that by combining the organizational design ontology with a domain specific ontology of MMORPGs, we will soon be able to develop LSMAS applications for MMORPGs.

The rest of the paper is structured as follows: Section II provides an overview of related work, Section III covers details about the two involved ontologies - one dealing with MMORPGs, and the other describing LSMAS organizational design; Section IV contains descriptions of the developed joint ontology. Section V contains conclusions and gives an outline of future research.

II. Related Work

Organization modeling *per se*, as well as its application to MAS development has been subject to a number of studies and can be considered a well established field of research (see [4] for a comprehensive state-of-the art overview). Organization of LSMAS has only recently been touched upon, since the specifics of large-scale systems haven't been in the focus of most organizational design related MAS development studies. Therefore, the present study builds on previously published research (see [6], [9], [10], [11], [7], [4]) lies in comprehension - greater inclusion of organizational modeling concepts, and their application to LSMAS elements.

A core ontology for organizational structures was published by World Wide Web Consortium (W3C) as a W3C recommendation in January 2014 [12]. This ontology allows for domain-specific extensions and can be used to describe organizational structures. Although this ontology is recommended, its applicability is constrained on the mentioned case of modeling structure only.

An overview of some of the most recent research on organizing MAS is given in [13], where the authors describe two levels of types of organization that can be used for organizing agents in a MAS: the individual level, and the organizational level. These levels of MAS organization are backed up by techniques for statically or dynamically organizing agents in MAS.

Organizational techniques for LSMAS are given in [9], covering organizational design methods for organizations consisting entirely of agents and agent systems. Work towards formalizing organizational design techniques aimed to provide a foundation for future studies of automated LSMAS development was presented therein.

Finally, an initial ontology about LSMAS organization is presented in [6]. The proposed ontology is extensive, since it includes organizational modeling concepts referring to organizational structure, organizational culture, strategy, processes, individual agents, organizational dynamics, as well as context and inter-organizational aspects.

Several papers contain records of gaming ontologies, specifically covering MMO, RPG, or MMORPG genre, on various levels of abstraction. The earliest found research with a specific result was [14] introducing the Game Ontology Project as a framework for description, analysis and study of games. Goal of representing an ontology including concepts related to a play event, rather than to the game as a piece itself is presented in [15]. In [16] it is aimed to better define a fantasy RPG using an ontology and creating a knowledge base. The Video Game Ontology, presented in [17], is aimed at modelling video game related information with the main goal of capturing knowledge about events that happen in video games and information about players. Basics for The Mana World ontology, as a work in progress, are laid out in [10].

III. Basic Ontologies

Two ontologies were used in order to create a joint ontology containing concepts related to organization of a specific LSMAS contained in The Mana World MMORPG. The organization-related ontology, based on various identified perspectives for modeling organization, is detailed in section III.A. The Mana World ontology contains all the identified most relevant concepts for describing the mentioned MMORPG [7], as is further detailed in section III.B.

A. Organizational Design of LSMAS Ontology

As mentioned above, this ontology is based on seven perspectives for complex organizational system modeling, as described in [6]:

- **Organizational structure** describes decision and information flows within an organization;
- **Organizational culture** defines relevant intangible concepts of an organization, including knowledge, norms, language, etc.;
- **Strategy** describes long-term objectives of the given organization along with action plans for their realization, and means of success measurement;
- **Processes** contain activities and procedures within an organization;
- **Individual agents** are the most important part of an organization, representing individuals who perform the work;
- **Organizational dynamics** represent organizational change, including reorganization of the earlier mentioned concepts;
- **Context and inter-organizational aspects** define organizational behaviour towards the environment of the organization, as well as relations between organizations.

The mentioned concepts are defined in the ontology using classes, as seen on Fig 1.

The most interesting aspects of definitions of the mentioned concepts, stated in [6], are recursive definitions which allow for working with individual agents, and various groups of agents, in the same manner. Furthermore, recursive definitions

[1]For more information visit http://ai.foi.hr/modelmmorpg

[2]See https://www.themanaworld.org/ for details.

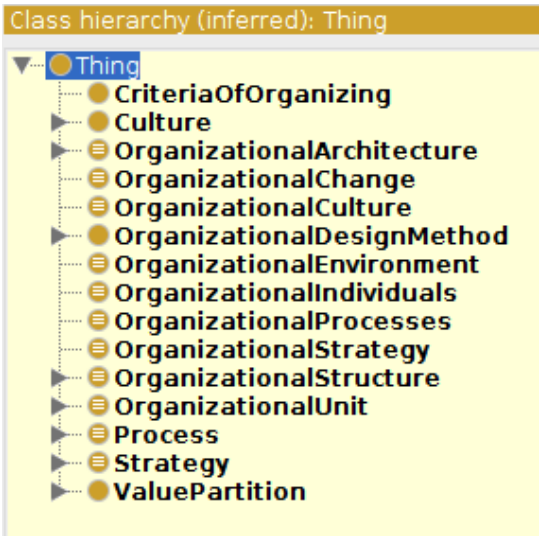


Fig. 1: Inferred class hierarchy of organizational concepts of the organizational ontology [6], as shown by Protégé

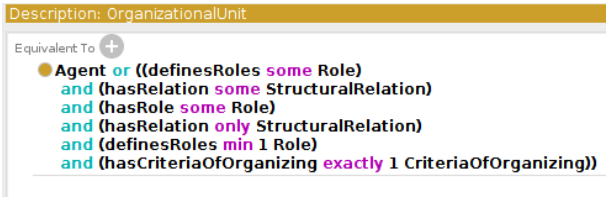


Fig. 2: Description of class `OrganizationalUnit`, as shown by Protégé

are used for organizational processes, organizational strategy and organizational knowledge artifacts as well. Such a nature of concept definitions enable us to work with the mentioned organizational concepts on various levels, i.e. from individual concepts representing individuals, to individual concepts representing groups of individuals. Furthermore, the important classes are defined using object properties, as depicted on Fig 2.

The ontology contains all the necessary concepts needed for description of organization of an LSMAS, thus proved an invaluable knowledge repository for the joint ontology.

B. The Mana World

The second ontology, which will be used to introduce basic MMORPG concepts to the organizational ontology mentioned earlier, is based on The Mana World MMORPG which contains most of the basic concepts which can be found in this genre of computer games. The Mana World ontology is interconnected with two more ontologies: Friend of a Friend (FOAF), and the Video Game Ontology (VGO). FOAF is a "project devoted to linking people and information using the Web."[3] The idea of FOAF is integration of three kinds of networks: social networks, representational networks, and information networks. On the other hand, VGO is "aimed at

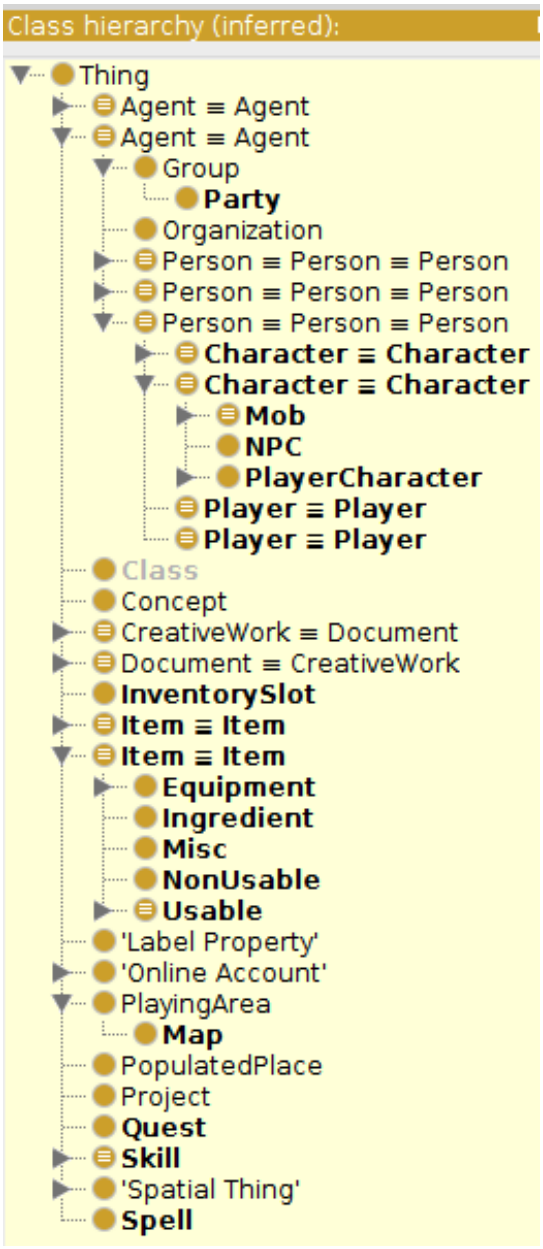


Fig. 3: Inferred class hierarchy of organizational concepts of The Mana World ontology [10], as shown by Protégé

modelling video game related information."[4] Furthermore, the main goal of VGO is "to capture knowledge about events that happen in video games and information about players." Both these ontologies are used in order to hone one of the main ideas of the Semantic Web, which is to allow an agent to "*start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing*" [18].

Hierarchy of the defined classes of concepts needed to describe The Mana World, and the included RPG elements, is shown in Fig. 3. Some similarities between the mentioned

[3]For more information see http://xmlns.com/foaf/spec/

[4]For more information visit http://vocab.linkeddata.es/vgo/

two ontologies (the organizational one and The Mana World one) are identifiable even by simple comparison of their class hierarchies, yet their further identified bonding places will be described in section IV. Several classes have been added to enrich the VGO ontology, e.g. specialization of Character class (representing all the characters that can be found in the game), Party class as a specialization of the FOAF Group class (containing all the groups of player characters, called parties), specialization of Item class, etc.

IV. Modeling Ontology

The ontology to be presented in this paper is created by combining both organizational ontology (section III. A) and The Mana World ontology (section III. B), in order to define concepts related to both MMORPG domain and the organizational domain, thus performing the first step towards an ontology encompassing organizational concepts needed for organization of LSMAS.

The organizational ontology already identified some of the most important concepts of organizational modeling, using an extensive literature review, as described in [6], [9]. On the other hand, The Mana World ontology identified most of the very important concepts of modeling this particular MMORPG, although it is a work in progress.

As mentioned earlier, the goal of joining these ontologies is in identifying elements of MMORPG as organizational elements, hence allowing for application of organizational modeling concepts on them. The joining points were identified as follows:

- `foaf#Agent` is equivalent to `org#Agent`
  - `foaf#Agent` contains all the character, character group and player classes of MMORPG ontology
- `mmorpg#Quest` is a `org#Objective`
- `mmorpg#CharacterClass` is a `org#Norm`
- `org#CriteriaOfOrganizing` contains the following:
  - `mmorpg#Item`
  - `mmorpg#Quest`
  - `mmorpg#Skill`

The mentioned joining points are described in further detail in the following subsections. Included are both specifications of the stated relations and applicable observations.

A. Organizational Unit

The most obvious class of the organization ontology that can be joined with the MMORPG ontology is `org#OrganizationalUnit` class, which represents agents as basic parts of an organization. `org#Agent` class of the organizational ontology, representing those agents which are organizational units, was identified to be equal to the `foaf#Agent` class already existing in the MMORPG ontology. Such a definition states that every agent contained by `Agent` class is considered an organizational unit, whether it is a single agent, or a group of agents, as defined by [6] in definition 6: *Any agent is an organizational unit. If*

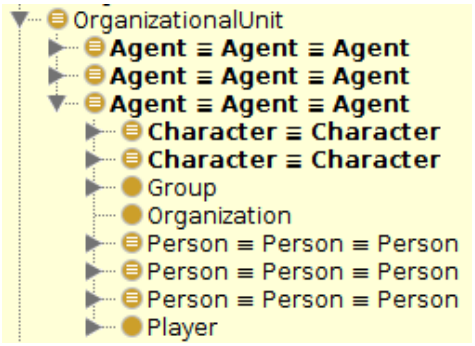


Fig. 4: Inferred class hierarchy of `OrganizationalUnit` class, as shown by Protégé

$OU = (O, R, C)$ is a labeled graph in which O is the set of organizational units (nodes), R is a labeled set of roles (relations, arcs) and C is a criteria of organizing then OU is an organizational unit.. The stated description is hierarchically represented in Fig. 4.

B. Organizational Objective

An important element of every organization are organizational goals, the driving force of an organization. The driving force of an RPG are quests which reward players with items or resources in exchange for a successfully completed quest. Each quest consists of a number of goals with specific criteria to be met in order for the quest to be completed. Therefore, quests can be categorized as a sort of objectives when speaking in the language of organization. Quests are one of the reasons why parties are created in a game and why players cooperate, since there are always quests that cannot be completed by players acting on their own, without cooperation with other players.

C. Organizational Norm

Normative system of an organization, as a concept combining all the norms applicable to an organization, and norms as acceptable behaviour in a group, which act as a behaviour template in the given group, have a very similar concept in the RPG world. Usually, player characters of MMORPGs are, in the course of the game, specialized in various classes. In some games this specialization occurs in the very beginning, and in some games this happens during the game, but the concept exists in most of RPG games. Character class defines the basic way a certain player, i.e. their character, is supposed to behave in the given world. For example, wizard class will definitely base on magic and magical powers, warrior class will build their strength skills and melee weapon usage thus specializing in hand-to-hand combat, whereas archer class will specialize in ranged attacks, agility skills and speed. Other classes may exist, usually as various combinations of those just mentioned. Based on everything stated, character class behaves as a certain norm, since it defines players and creates a set of behaviour expected of a certain class.

D. Role of an Agent

Although rather similar to a character class mentioned earlier, a role is argued to be a slightly different concept. Still being discussed, agent role may represent a temporary behaviour blueprint of an agent. If a role contains a set of permitted or suggested actions for an agent, then roles are temporary types of behaviour expected of an agent, containing processes, goals and actions which are supposed to lead the agent to fulfill their goals. For example, an agent with a goal of successfully solving a quest requiring them to gather 10 instances of an item available in the wild, may consider taking upon themselves a role of *Gatherer* which contains all the actions required for gathering resources, and strategy for successfully solving such goals. Other thoughts about roles in RPGs are contained in the view that character class may be a role that a player's character can take. This understanding of a role depends on the definition of a character class, which is, usually, fixed once chosen, and cannot be changed. Even though this constraint must be taken into account, character classes do pose certain normative restrictions on characters, e.g. a preferred type of battle, favoured character traits, skill, or basic stats, special kind of weapon, etc. Supporting this view of classes as roles, agents would have some constraints placed upon their behaviour and possible actions, as well as some customization of their goals, based on their chosen classes. For example, a Wizard class can use only weapons of type Staff, favours character stat Intelligence, prefers Ranged battle or Supportive battle, and their goals are concerned with upgrading said features.

E. Criteria of Organizing

The organizational modeling ontology describes the `org#CriteriaOfOrganizing` class as "*A particular criteria for organizing things like processes, organizational units, strategies or cultural artifacts.*" There are several concepts in the MMORPG world which may be used for organizing some of the stated concepts. The main criteria for creating a group of agents, i.e. an organizational unit consisting of several agents, may be the `mmorpg#Quest` concept, which consists of all the quests in the game. Agents currently trying to solve a specific quest may group in order to surpass requirements of the quest, e.g. a strong boss monsters. Class containing all the skills a character can have in the game, `mmorpg#Skill`, may be used for gathering a group of agents whose characters possess certain skills desired at the moment (whether needed by a specific quest, or another goal of an agent), e.g. specific school of magic, poison resistance, or close combat skill. Lastly, items, gathered in class `mmorpg#Item`, may be considered as criteria of organizing, since not all items are the same. Some monsters may be killed more efficiently with e.g. ranged weapons, some items are stronger than others hence being more suitable for group leaders, some items are required by a quest, etc. Described relations are shown in Fig. 5

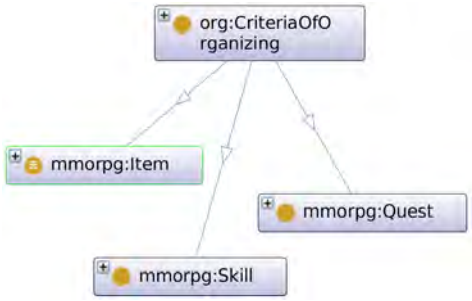


Fig. 5: Class hierarchy of `org#CriteriaOfOrganizing`, including classes `mmorpg#Quest`, `mmorpg#Skill`, and `mmorpg#Item`, as shown by Protégé

V. Conclusion

In this paper a work-in-progress integration between an organizational design ontology for the development of LSMAS as well as a specific domain ontology for MMORPGs has been presented. The main objective of the study is to develop an all purpose LSMAS ontology-based modeling framework that will be tested on MMORPGs as one of the obvious application domains. Thus, it was necessary to combine these two ontologies in order to allow for modeling of LSMAS in MMORPG scenarios, by using concepts from both domains.

As has been shown in this paper, the organizational design LSMAS ontology is applicable to the MMORPG domain and most important concepts can be mapped to meaningful concepts in the other ontology. In future research we hope to apply the organizational design ontology on other domains as well, especially the Internet of Things and Smart City domains.

In short term, we will try to develop a metamodel out of the organizational design ontology, that will be the main building block of a to be developed modeling tool for LSMAS. The modeling tool shall be tested in an MMORPG environment through the development of artificial player parties and their testing in special test-bed scenarios.

Acknowledgment

This work has been supported in full by the Croatian Science Foundation under the project number 8537.

References

[1] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[2] E. Argente, V. Julian, and V. Botti, "Multi-agent system development based on organizations," *Electron. Notes Theor. Comput. Sci.*, vol. 150, no. 3, pp. 55–71, May 2006. [Online]. Available: http://dx.doi.org/10.1016/j.entcs.2006.03.005

[3] B. Horling and V. Lesser, "A Survey of Multi-Agent Organizational Paradigms," *The Knowledge Engineering Review*, vol. 19, no. 4, pp. 281–316, 2005. [Online]. Available: http://mas.cs.umass.edu/paper/366

[4] M. Schatten, J. Ševa, and I. Tomičić, "A roadmap for scalable agent organizations in the internet of everything," *Journal of Systems and Software*, vol. in press, 2016.

[5] P. Chamoso, F. De la Prieta, F. De Paz, and J. M. Corchado, "Swarm agent-based architecture suitable for internet of things and smartcities," in *Distributed Computing and Artificial Intelligence, 12th International Conference*. Springer, 2015, pp. 21–29.

[6] M. Schatten, "Organizational architectures for large-scale multi-agent systems' development: An initial ontology," *Advances in Intelligent Systems and Computing*, vol. 290, pp. 261–268, 2014.
[7] M. Schatten, I. Tomičić, and B. Okreša Đurić, "Multi-agent modeling methods for massivley multi-player on-line role-playing games," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on*. IEEE, 2015, pp. 1256–1261.
[8] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed., ser. Prentice Hall series in artificial intelligence, S. Russell and P. Norvig, Eds. New Jersey, USA: Pearson Education, Inc., 2010.
[9] M. Schatten, P. Grd, M. Konecki, and R. Kudelić, "Towards a formal conceptualization of organizational design techniques for large scale multi agent systems," *Procedia Technology*, vol. 15, pp. 577–586, 2014.
[10] B. Okreša Đurić and M. Konecki, "Specific owl-based rpg ontology," in *Central European Conference on Information and Intelligent Systems 26th Inernational Conference*, 2015.
[11] M. Schatten and B. Okreša Đurić, "A social network analysis of a massively multi-player on-line role playing game," in *4th International Conference on Modeling and Simulation*, 2015.
[12] W3C, "The Organization Ontology," 2014. [Online]. Available: https://www.w3.org/TR/vocab-org/
[13] H. A. Abbas, S. I. Shaheen, and M. H. Amin, "Organization of Multi-Agent Systems: An Overview," *International Journal of Intelligent Information Systems*, vol. 4, no. 3, pp. 46–57, 2015.
[14] J. Zagal, "Towards an Ontological Language for Game Analysis," *Worlds in Play: International Perspectives on Digital Games Research*, pp. 21–35, 2005.
[15] J. Chan and W. Yuen, "Digital game ontology: Semantic web approach on enhancing game studies," *2008 9th International Conference on Computer-Aided Industrial Design and Conceptual Design*, 2008.
[16] M. Roman, I. Sandu, and S. C. Buraga, "OWL-Based Modeling of RPG Games," *Studia Universitatis Babes-Bolyai*, vol. 56, no. 3, p. 83, 2011.
[17] J. Parkkila, F. Radulovic, M. Poveda, and D. Garijo, "The Video Game Ontology," Tech. Rep., 2014. [Online]. Available: http://vocab.linkeddata.es/vgo/
[18] S. Hawke, I. Herman, P. Archer, and E. Prud'hommeaux, "W3C Semantic Web Activity," 2013. [Online]. Available: http://www.w3.org/2001/sw/

Comparison of solution representations for scheduling in the unrelated machines environment

Marko Đurasević, Domagoj Jakobović
Faculty of Electrical Engineering and Computing, Zagreb, Croatia
Email: marko.durasevic@fer.hr, domagoj.jakobovic@fer.hr

Abstract – Most scheduling problems belong to the class of NP hard problems. Because of that reason, search based approaches are often used in order to find solutions for scheduling problems. In this paper we compare several search-based approaches for finding solutions for the unrelated machines scheduling problem. These search based approaches use two different solution representations for the aforementioned problem and the representations are compared with each other. The first representation uses a permutation vector to encode the solution, while the second one uses a vector of floating point numbers. The results of the search based approaches are compared to several existing heuristics developed especially for solving the unrelated machines scheduling problem. We also perform a complexity analysis of the search-based approaches, in which we compute the time needed for them to outperform the problem specific heuristics.

I. INTRODUCTION

Scheduling can be defined as a decision-making process concerned with the allocation of scarce resources to tasks over a given time in order to optimize one or more objectives [1], [2]. A great number of scheduling problems belong to the class of NP hard problems, meaning that no efficient algorithms exist which could find the optimal solution in a reasonable amount of time. Therefore, such problems are often solved by using different heuristic algorithms. Based on the nature of how the solutions are generated, heuristic algorithms are divided into two groups: the first group, denoted as *search-based*, consists of metaheuristic approaches which search the whole solution space in order to find the optimal solution, while the second group consists of problem specific heuristics which iteratively construct the solution [3].

Search-based approaches have the advantage that they are usually able to find near optimal solutions. However, this advantage comes at a price, since search-based approaches are rather slow when compared to approaches which build the schedule iteratively. Additionally, search-based approaches can mostly be used in scheduling environments where all information about the environment is available beforehand, since they search the whole solution space. In such cases, when all the information about the scheduling environment is present beforehand, these approaches are mostly the methods of choice, since they outperform simple constructive scheduling heuristics.

One of the most important design choices in search-based algorithms is the solution representation of the given problem. The overall effectiveness of search-based approaches can, to a great extent, depend on the chosen solution representation [4]. Most commonly a solution representation which can best represent the given problem is chosen, but that may not always be the best choice, since some other solution representations could produce better results. Because of that, this paper will compare two solution representations for scheduling in the unrelated machines environment. The first representation is a permutation representation similar to the ones used in the literature. The second representation uses a vector of floating point numbers in order to represent the schedule. To the best of our knowledge, the second representation has not yet been used to represent solutions for the unrelated parallel machines environment. The results of those two approaches will be compared to several iterative scheduling heuristics. Additionally, this paper will examine the time complexity of genetic algorithms (using the aforementioned solution representations) when compared to some traditional heuristic methods for creating schedules for the unrelated parallel machines environment.

The paper is organized as follows: in the second section a short overview of the unrelated machines environment is given. The third section describes the two solution representations in more detail. The results are presented in the fourth section, while a discussion about the results is given in the fifth section. Finally, a short conclusion and future research directions are presented in the sixth section.

II. RELATED WORK

A lot of work has been invested in solving different scheduling problems by using various metaheuristic methods [5], [6], [7], [8], [9], [10]. Unfortunately, very little research has been done in the area of the unrelated machines environment. In [11] the authors propose a hybrid algorithm combining ant colony optimization, simulated annealing and variable neighborhood search in order to solve scheduling problems for parallel unrelated machine environment. The ant colony optimization algorithm was also used in [12], where the authors propose several new ideas in order to improve the performance of the algorithm. A Greedy Randomized

Adaptive Search Procedure (GRASP) metaheuristic was proposed in [13]. A Competitive Evolutionary Strategy Memetic Algorithm which for optimizing two objectives simultaneously was proposed in [14].

Genetic algorithms have also been used to solve the unrelated machines problem on several occasions. For example, in [15], the author adapts a genetic algorithm for the unrelated machines environment and shows promising results. In [16] a genetic algorithm using a permutation representation and several local search operators was proposed. It was shown that the proposed method achieved excellent performance, when compared to other similar methods. GARP, a genetic algorithm combined with variable neighborhood descent and path relinking was proposed in [17]. In [18] a hybrid genetic algorithm was proposed for the unrelated machines environment with worker allocations. This article additionally examines several solution representations for the given problem. Solving the unrelated parallel machines environment as a subset of the flexible flow shop problem was discussed in [19].

III. SCHEDULING IN THE UNRELATED MACHINES ENVIRONMENT

The unrelated machines environment consists of *n* jobs which compete in order to be processed on one of the *m* machines [2]. Each job is defined by several properties including the processing time p_{ij}, which defines the execution time of job with the index *j* on the machine with the index *i*; the release time r_j, which defines the time when the job arrives in the system; the due date d_j, which defines a point in time by which the job should finish with its execution, otherwise a certain cost will be caused by its delay; and a weight w_j, which denotes the importance of a job.

A. Scheduling Criteria

In order to assess the quality of a given schedule, certain scheduling criteria need to be defined. Since the number of the proposed scheduling criteria is numerous, four of the most prominent criteria from the literature are chosen. But before defining the criteria for the entire schedule, it is essential to first define certain criteria for individual jobs, which will then in turn be used to define the scheduling criteria.

For each job the following criteria are defined [1]:

- Finish time C_j – denotes the point in time in which the job ended with its execution
- Flowtime F_j – denotes the amount of time a certain job spent in the system before it finished with its execution, which is defined as: $F_j = C_j - r_j$,
- Tardiness T_j – denotes the amount of time that the job was executing after its due date. It is defined as: $T_j = \max(C_j - d_j, 0)$
- Is tardy U_j – denotes if a job is tardy or not. It is defined as: $U_j = \begin{cases} 1 : T_j > 0 \\ 0 : T_j = 0 \end{cases}$

In this paper the following four scheduling criteria will be optimized:

- Makespan $C_{\max}$ – represents the maximum finishing time of all jobs. It is defined as: $C_{\max} = \max(C_j)$
- Total flowtime Ft – represents the sum of all job flowtimes. It is defined as: $Ft = \sum_j F_j$
- Total weighted tardiness Twt – represents the weighted sum of the tardiness values of all jobs, and is defined as: $Twt = \sum_j w_j T_j$
- Weighted number of tardy jobs Uwt – represents the weighted sum of tardy jobs. It is defined as: $Uwt = \sum_j w_j U_j$

B. Scheduling Conditions

Scheduling problems can be classified into several classes depending on some of their properties. For example, based on the availability of system parameters (for example, the information about all jobs which will arrive to the system) scheduling problems can be divided into *offline scheduling*, in which all information is available from the start, and *online scheduling*, in which the information about a job becomes available only when that concrete job enters the system). On the other hand, scheduling problems can also be divided based on when the schedule is created. In *static scheduling*, the entire schedule is created before the system begins with its execution, while in *dynamic scheduling* the schedule is built simultaneously with the execution of the system.

Scheduling conditions can have a great influence on the algorithms which can be used for creating schedules. For example, search-based approaches can mostly be used in static offline scheduling; on the other hand, methods which build the schedule iteratively are mostly used in dynamic online scheduling, but can be also used in static and offline scheduling.

In this paper we presume the static scheduling conditions, where all the parameters are known before the jobs enter the system.

C. Constructive Scheduling Heuristics

Constructive scheduling heuristics are simple iterative procedures which build the schedule incrementally. The main concept of all such heuristics is the same: each time a machine becomes available and there are jobs to be scheduled, the heuristic selects one of the jobs and schedules it on the available machine. Specific heuristics then differ from each other based on the rule which is used to select the job. This rule can, for example, select jobs in the following ways: the job with the shortest processing time, the job with the highest weight, the job with the closest due date, etc. For the unrelated machines scheduling environment, there exists a wide variety of different scheduling heuristics, but for this article, we will limit ourselves to the following four popular scheduling heuristics: min-min [20], max-min [21], sufferage [22] and min-max [23].

IV. Solution Representations in Genetic Algorithms for Scheduling Problems

A. Genetic Algorithms

Genetic algorithms are a stochastic optimization technique which can be used in order to find high quality solutions for a wide variety of problems [24], including for different scheduling problems. The main idea of genetic algorithms is to simulate the process of natural evolution in order to find good solutions. The algorithm starts with a randomly generated population of solutions, where each solution is called an individual. In order to compare individuals with each other, a fitness function needs to be defined. The role of the fitness function is to determine how "good" a certain individual is. Based on that information the *selection* operator chooses "better" individuals to survive, while trying to eliminate the "worse" individuals. Other operators, like *crossover* and *mutation* are used to explore the search space of the problem. Crossover usually tries to combine two individuals into a single individual which contains features of both parents, and which will hopefully represent a better solution. Mutation, on the other hand, introduces random changes into an individual, in order to extend the search to less covered regions. The algorithm iteratively applies the aforementioned genetic operators on a population, until a given stopping criteria is met.

Since each solution representation uses a different set of genetic operators, the actual list of used genetic operators will be given in the following subsections with the descriptions of the representations.

B. Permutation representation

The permutation representation is commonly used in order to represent solutions to scheduling problems. Throughout the literature many different variations of permutation representations have been used [4]. Which variant will be used, greatly depends on the scheduling environment and scheduling conditions of the problem.

In this paper, a representation consisting of two vectors will be used. The first vector will represent a permutation of jobs, which determines the sequence in which the jobs will be executed. But with this information alone it is not possible to determine on which machine a certain job should be executed. Because of that, a second vector of integers is used to denote on which machine a certain job will be executed.

9	4	3	7	2	0	8	1	5	6
0	2	2	1	0	0	1	0	2	1

Figure 1 Solution representation in the permutation based GA

Figure 1 represents an example of a solution representation for a scheduling problem with three machines and ten jobs. This solution represents a schedule in which the jobs with the indices 9, 2, 0 and 1 will be executed (in that order) on the machine with the index 0, jobs with the indices 7, 8 and 6 will be executed on the machine with the index 1, while jobs with indices 4, 3 and 5 will be executed on the machine with index 2.

Since this representation uses two vectors, where the first one represents a permutation, while the second represents just a normal array of integers, two different sets of genetic operators were used for each vector. For the first vector, 13 different crossover operators have been used, including OBX, OX, PBX, PMX and CX [10]. The operators are combined in a way that each time a crossover is performed, a randomly selected operator is applied.

For the second vector, only the uniform crossover operator was used. This operator takes two parent integer vectors and combines them into a single child vector, where at each position a value from a randomly selected parent is copied into the child.

Regarding the mutation operators, three have been used for the permutation vector (insert, inverse and swap mutation) and one for the integer vector (simple mutation). The simple mutation changes a value at the randomly selected position in the vector.

C. Floating point representation

The floating point representation, which usually consists of one vector holding floating point numbers, is most commonly used for solving continuous problems. Since scheduling problems are not continuous but combinatorial, a decoding scheme which will transform the floating point vector into a schedule needs to be devised.

The presented representation uses a floating point vector with size equal to the number of jobs. Each element in the vector is a number in the interval [0, 1]. These numbers denote priorities for each job, where 0 represents the highest priority, while 1 represents the smallest priority. This information is still not enough in order to create a complete schedule, since the mapping between jobs and machines is still undefined. For that reason, the interval [0, 1] is divided into m subintervals and

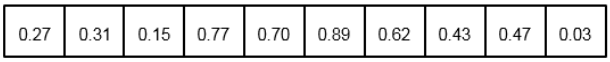


Figure 2 Solution representation in the floating point based GA

depending to which subinterval the priority value belongs, it will be mapped to the appropriate machine.

Figure 2 represents an example solution for a scheduling problem with three machines and ten jobs, encoded in the floating point representation. Since the scheduling problem contains three machines, the interval [0, 1] is divided into three subintervals: [0, 0.33> for the machine with index 0, [0.33, 0.66> for the machine with index 1 and finally [0.66, 1] for the machine with index 2. When a priority value belongs to a certain subinterval, it means that the job to which that priority value belongs will be mapped to the corresponding machine. In this example, the jobs with indices 0, 1, 2 and 9 will be mapped to the machine with index 0, jobs with indices 6, 7 and 8 will be mapped to the machine with index 1, and jobs with indices 3, 4 and 5 will be mapped to the machine with index 2. Since the mapping from each job to each machine has been resolved, only the sequences of those jobs on each machine need to be determined. For the first machine the job with index 9 has the smallest priority value, which means that this job will be scheduled first, following by jobs with indices 2, 0 and 1. The sequence for the other jobs is determined in the same way.

For this solution representation we used 15 different crossover operators, including BLX, arithmetic, one point and heuristic [25]. The crossover operators are combined in the same way as for the permutation encoding. As for the mutation, only the simple mutation operator is used, which changes the value of a randomly selected element of the vector.

D. Algorithms

Both the permutation and floating point encoding are used within a genetic algorithm that employs the same selection method. In this paper a steady state genetic algorithm is used with the tournament size of $k = 3$ individuals, which is given in Figure 3.

The floating point representation has an additional advantage that it can be used with existing continuous optimization algorithms other than GA. To illustrate that, in this paper we also experiment with a memetic genetic algorithm, which incorporates a local search method. The search method in this case is based on pattern search Hooke-Jeeves deterministic algorithm [26]. The goal of incorporating a local search method is to guide the algorithm towards the nearest local optimum, in the hope it will also represent a globally optimal solution. This usually comes at a cost of a higher number of evaluations needed to reach the same objective value. The analysis of efficiency of different local search methods is out of the scope of this paper, but we include an example of this approach to demonstrate its applicability.

```
Steady-state tournament selection
{  randomly select k individuals;
   remove the worst of k selected;
   child = crossover(best two of k selected);
   perform mutation on child, with given
   individual mutation probability;
   insert child into population;
}
```

Figure 3. Pseudo-code of the steady state tournament GA

V. RESULTS

In this section, results for both solution representations will be presented. The genetic algorithm with the permutation solution representation will be denoted as GA-PERM, while the genetic algorithm with the floating point representation will be denoted as GA-FP. An additional memetic algorithm which uses Hooke-Jeeves local search, coupled with floating point encoding only, is denoted with GA-HJ.

A. Benchmark setup

In order to test the different representations, a benchmark set consisting of 60 scheduling problem instances was prepared. In order to test the algorithms on scheduling problems with different characteristics, the problem instances were designed with different number of jobs and machines. Depending on the specific problem instance, the number of machines can be 3, 6 or 10 and the number of jobs can be 12, 25, 50 or 100. More details about the design of the problem instances can be found on the project website [27]. In order for the benchmarks to be statistically significant, the algorithms were executed 30 times for each problem instance. The complete result for each algorithm is calculated as the sum of the best achieved solutions over all problem instances.

B. Parameter tuning

Since it is known that parameter values can have a significant influence on the obtained results, it is of great importance to perform a thorough optimization of the algorithm parameters, not only to obtain better results, but also in order to make the comparison as fair as possible and to determine the sensitivity of algorithms to the parameter values. First off, the maximum number of function evaluations has been set as the common termination criterion to allow for a fair comparison between different algorithms. The maximum number of function evaluations has been set to one million in order to give all algorithms enough time to converge. Two other parameters, population size and mutation probability, have been optimized for each algorithm separately. Both parameters were tuned by optimizing the *Twt* criterion, and then used for the other three criteria.

Table 1 represents the results achieved for various population sizes by the three genetic algorithm variants, with the best result bolded for each algorithm. The quality of solutions found by GA-PERM steadily increases with the increase of the population size, achieving best results

Table 1 Weighted tardiness optimization with different population sizes

Algorithm	Population size			
	30	100	200	1000
GA-PERM	11.14	10.42	10.15	**9.816**
GA-FP	**9.577**	9.850	9.70	9.617
GA-HJ	**17.68**	22.20	25.69	41.99

for a population size of 1000 individuals. The other two approaches show best results for the smallest population size of 30 individuals. While the solution quality degrades with the increase of population size for the GA-HJ algorithm, for the GA-FP algorithm, the solution quality drops for the population size of 100 individuals, but then for greater population sizes the solution quality slowly increases.

Table 2 represents the results for various mutation probability values. From the results it is apparent that all algorithms prefer higher mutation values. Both GA-PERM and GA-HJ achieve the best results for the highest mutation probability of 0.7, while GA-FP achieves the best results for a mutation probability of 0.5.

C. *Result comparison*

Table 3 shows the results for all the tested criteria achieved by the search based approaches, and additionally by four constructive scheduling heuristics selected from the literature: min-min, max-min, sufferage and min-max. The best values achieved by either the group of scheduling heuristics and search based approaches have been bolded. As it can be seen from the table, there does not exist a single scheduling heuristic which achieves supreme results for all criteria. The sufferage heuristic and min-min heuristic each achieve the best results for two criteria. The max-min heuristic, on the other hand, achieved the overall worst results for all criteria.

From the genetic algorithm approaches, GA-PERM and GA-FP achieved results which were better than results from all the scheduling heuristics. GA-HJ, on the other hand, was able to surpass the results of scheduling heuristics in only a few cases, and was generally not very effective. When the genetic algorithm approaches are compared with each other, GA-FP found the best solutions for all criteria, followed closely by GA-PERM. GA-HJ

Table 2 Weighted tardiness optimization with different mutation probabilities

Algorithm	Mutation probability				
	0.05	0.1	0.3	0.5	0.7
GA-PERM	9.837	9.920	9.816	9.773	**9.725**
GA-FP	9.767	9.673	9.577	**9.533**	9.566
GA-HJ	17.45	18.38	17.68	17.50	**16.77**

Table 3 Complete results for all scheduling criteria

Algorithm	Criteria			
	Twt	Nwt	Ft	Cmax
Constructive				
Min-min	16.71	**7.143**	**157.2**	38.31
Max-min	22.06	8.138	195.8	38.83
Min-max	17.49	7.793	167.3	38.06
Sufferage	**16.65**	7.194	160.9	**37.92**
Search-based				
GA-PERM	9.725	5.615	151.2	37.14
GA-FP	**9.533**	**5.340**	**140.8**	**36.79**
GA-HJ	16.77	6.395	232.1	41.67

came last with the significantly worst results from all three search-based approaches.

D. *Execution time comparison*

Apart from the obtained solution quality, the execution time very often represents an important factor in the choice of algorithms for solving a certain problem. Because of that, it is important to be aware of the time complexities of all the different approaches.

In order to compare the execution time of genetic algorithms with the constructive heuristics, we examine the quality level of the GA during its execution. When a solution is found by the GA which is of equal or better quality than that of a corresponding heuristic, we record the time at which that solution was obtained. The same process is repeated for 30 independent runs, and the resulting average time is used for runtime comparison.

Note that the GA is in general not guaranteed to converge to a better solution for every problem instance; in that case, a maximum runtime needed for one million evaluations is recorded. In our experiments, both GA-PERM and GA-FP were always able to find a better solution. The GA-HJ variant, however, was consistently worse and is therefore excluded from the comparison.

Table 4 represents the average execution times of the scheduling heuristics, GA-PERM and GA-FP. The first column represents which scheduling heuristic is used as a reference, the second column represents the criterion which was optimized, the third column represents the execution times of the scheduling heuristics, while the rest of the columns represent results for GA-PERM and GA-FP. It can be seen that the scheduling heuristics can construct the schedule for all 60 problem instances in around 0.2 seconds (depending on the heuristic). On the other hand, the time needed by the GA-PERM and GA-HJ algorithms to reach solutions of the same quality ranged from around 3 minutes up until 22 minutes, depending on the criterion. Regarding the algorithm, it can be seen that the GA-FP was able to find solutions of the necessary quality in less time than GA-PERM. Depending on the criteria, GA-PERM and GA-FP need more time to find the necessary solutions for the *Cmax* and *Ft* criteria, than for the *Twt* and *Nwt* criteria, but that behavior is expected since these scheduling heuristics are more suited to the

former two criteria (which can also be seen from the results).

VI. DISCUSSION

Based on the results outlined in the previous section, several things can be concluded. First of all, it was shown that GA-FP achieved the best results for every scheduling criterion. GA-PERM came second, achieving results which are a few percent worse than those of GA-FP. This could come as a bit of a surprise, considering that the permutation representation is much more intuitive to represent solutions of scheduling problems. But on the other hand, the permutation representation needs two vectors to represent a complete solution, while the floating point representation encodes everything in a single vector. As a consequence, GA-PERM needs to perform genetic operators on both vectors, and it seems that this proves to be more challenging than performing genetic operators on only a single vector.

GA-HJ on the other hand achieved the worst results among the three GA approaches. This behavior was certainly not expected, since the Hooke-Jeeves local search was introduced to improve the results. In order to detect the cause of such behavior, results for each test instance of GA-HJ were compared to results of GA-PERM and it was noticed that both approaches achieve similar or the same results on smaller scheduling problems with 12 or 25 jobs. On the other hand, GA-HJ performs significantly worse on the larger problem instances with 50 and 100 jobs. This is probably the consequence of the local search algorithm, which in each iteration probes at least one additional point in every dimension and evaluates it, thus significantly increasing the function evaluation count. It could be possible that this algorithm would need much more evaluations in order to reach solutions of the same quality, but this is not an acceptable solution, since the maximum evaluation count is already large enough.

When compared to scheduling heuristics, GA-PERM and GA-FP have obtained better results, which is expected since they search the whole solution space in order to find the best schedule. The differences are more prominent for the *Twt* and *Nwt* criteria, since the constructive heuristics do not consider due dates of the jobs.

The execution time of the constructive heuristics is almost negligible, since they can create schedules for 60 problem instances in only 0.2 seconds. This characteristic makes these approaches applicable in dynamic and real world environments, where scheduling decisions need to be made very fast, and the algorithm needs to be able to quickly react to changes in the environment (such as the arrival of new jobs).

In order to find solutions of the same quality as those found by the scheduling heuristics, GA-PERM and GA-FP needed much more time, which ranged from 3 minutes up until 22 minutes, depending on the algorithm and optimization criterion. In all cases GA-PERM needed more time in order to find those solutions than the GA-FP algorithm. This behavior could also be the consequence of GA-PERM operating on two vectors, thus needing more time in order to reach good solutions. Additionally, since genetic algorithms are stochastic, there is no guarantee that they will be able to locate good solutions on the first try. As a consequence, these algorithms need to be executed several times in order to ascertain that good solutions have really been found.

Table 4 Execution time comparison between scheduling heuristics and genetic algorithms

Scheduling heuristic (SH)	Criterion	SH execution time (s)	GA-PERM execution time (s)	GA-FP execution time (s)	Ratio GA-PERM/SH	Ratio GA-FP/SH	Ratio GA-PERM/GA-FP
Sufferage	Cmax	0.197	1188.2	820	6031	4162	1.45
	Ft		1184.8	609.6	6014	3094	1.94
	Nwt		525.1	219.6	2665	1114	2.39
	Twt		265.2	200	1346	1015	1.33
Min-min	Cmax	0.097	1139.2	806.3	11744	8312	1.41
	Ft		1296.4	644.4	13364	6643	2.01
	Nwt		496.5	209.7	5118	2161	2.37
	Twt		264.6	203.5	2727	2097	1.30
Min-max	Cmax	0.215	1182.8	848.3	5501	3945	1.39
	Ft		1127.9	562.2	5246	2614	2.01
	Nwt		479.8	228.5	2231	1062	2.10
	Twt		259.6	206.8	1207	961	1.25
Max-min	Cmax	0.213	1057.6	728	4965	3417	1.45
	Ft		1002.8	458.8	4707	2153	2.19
	Nwt		416.6	200.5	1955	941	2.08
	Twt		237.7	176.7	1115	829	1.35

VII. CONCLUSION

This paper compares two different solution representations for scheduling problems in the unrelated machines environment. Although both representations were able to obtain good results, the floating point representation achieved the best results for all criteria. Unfortunately, contrary to our initial expectations, by adding a local search operator for the floating point representation, we were not able to improve the results of the genetic algorithm approach. Additionally, the genetic algorithm approaches were compared to several constructive scheduling heuristics, and it was shown that they are able to find much better solutions than scheduling heuristics, but with substantially longer execution time. This just shows that there is a tradeoff between solution quality and the time needed to find those solutions. The choice of the appropriate approach will then depend on the scheduling environment and user requirements.

For future research it is planned to further investigate some other solution representations, and compare them to the representations analyzed in this paper. Additionally, other local search mechanisms will also be tried out for both solution representations.

REFERENCES

[1] M. P. Pinedo, "Scheduling: theory, algorithms, and systems," Springer Science \& Business Media, 2012.

[2] J. Y-T. Leung, "Handbook of Scheduling: Algorthms, Models, and Performance Analysis", Chapman & Hall/CRC, 2004.

[3] A. Jones, L. Rabelo, "Survey of job shop scheduling techniques," tech rep., NISTIR, National Institute of Standards and Technology.

[4] E. Hart, P. Ross, D. Corne, "Evolutionary Scheduling: A Review," Genetic Programming and Evolvable Machines, 6, pp. 191–220, 2005.

[5] H. Zhou, Y. Feng, L. Han, "The hybrid heuristic genetic algorithm for job shop scheduling," Computers and Industrial Engineering, pp. 191-200, 2001.

[6] L. Wang, H. Siegel, V. Roychowdhury, A. Maciejewsky, "Task Matching and Scheduling in Heterogeneous Computing Environments Using a Genetic-Algorithm-Based Approach," Journal of Parallel and Distributed Computing, pp. 8-22, 1997.

[7] H. Zhou, W. Cheung, L. Leung, "Minimizing weighted tardiness of job-shop scheduling using a hybrid genetic algorithm,", pp. 637-649, 2009.

[8] W. Cheung, H. Zhou, "Using genetic algorithms and heuristics for job shop scheduling with sequence-dependent setup times", Annals of Operations Research, pp. 65-81, 2001.

[9] S. Petrovic, E. Castro, "A Genetic Algorithm for Radiotherapy Pre-treatment Scheduling", Applications of Evolutionary Computation, pp. 462-471, 2010.

[10] F. Werner, "Genetic Algorithms for Shop Scheduling Problems: a Survey".

[11] J. Behnamian, M. Zandieh, S. M. T. Fatemi Ghomi, "Parallel-machine scheduling problems with sequence-dependent setup times using an ACO, SA and VNS hybrid algorithm", Expert Systems with Application, 9637-9644, 2009

[12] C. W. Lin, Y. K. Lin, H. T. Hsieh, "Ant colony optimization for unrelated parallel machine scheduling", International Journal of Advanced Manufacturing, 35-45, 2013.

[13] J. P. C. M. Nogueira, J. E. C. Arroyo, H. M. M. Villadiego, L. B. Goncalves, "Hybrid GRASP Heuristics to Solve an Unrelated Parallel Machine Scheduling Problem with Earliness and Tardiness Penalties", Electronic Notes in Theoretical Computer Science, 53-72, 2014.

[14] C. C. Chyu, W. S. Chang, "A Competitive Evolution Strategy Memetic Algorithm for Unrelated Parallel Machine Scheduling to Minimize Total Weighted Tardiness and Flow Time", Computers and Industrial Engineering, 1-6, 2010.

[15] S. Balin, "Non-identical parallel machine scheduling using genetic algorithm", Expert Systems with Applications, 6814-6821, 2011.

[16] E. Vallada, R. Ruiz, "A genetic algorithm for the unrelated parallel machine scheduling problem with sequence dependent setup times", 612-622, 2011.

[17] M. N. Haddad, I. M. Coelho, L. S. Ochi, M. J. F. Souza, H. G. Santos, "GARP: A New Genetic Algorithm for the Unrelated Parallel Machine Scheduling Problem with Setup Times", International Conference of the Chilean Computer Science Society, 2012.

[18] A. Costa, F. Cappadonna, S. Fichera, "A hybrid genetic algorithm for job sequencing and worker allocation in parallel unrelated machines with sequence-dependent setup times", International Journal of Advanced Manufacturing Technology, pp. 2799-2817, 2013.

[19] H. C. Chang, H. T. Tsai, T. K. Liu, "Application of Genetic Algorithm to Optimize Unrelated Parallel Machines of Flexible Job-Shop Scheduling Problem", International Conference on Control & Automation, 2014.

[20] T. D. Braun, H. J. Siegel, N. Beck, L. L. Bölöni, M. Maheswaran, A. I. Reuther, J. P. Robertson, M. D. Theys, B. Yao, D. Hensgen, et al., "A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems," Journal of Parallel and Distributed computing, pp. 810-837, 2001.

[21] E. Davis, J. M. Jaffe, "Algorithms for scheduling tasks on unrelated processors," Journal of the ACM, pp. 721-736, 1981.

[22] D. Hensgen, R. F. Freund, "Dynamic mapping of a class of independent tasks onto heterogeneous computing systems," Journal of Distributed Computing.

[23] H. Izakian, A. Abraham, V. Snasel, "Comparison of heuristics for scheduling independent tasks on heterogeneous distributed Environments," Computational Sciences and Optimization, 2009. CSO 2009. International Joint Conference on, pp. 8–12, 2009.

[24] Z. Michalewicz, "Genetic Algorithms + Data Structures = Evolutionary Programs", Springer-Verlag, Berlin, 1992.

[25] S. Picek, D. Jakobović, "From fitness landscape to crossover operator choice," Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation, pp. 815-822, 2014.

[26] R. Hooke, T. A. Jeeves, ""Direct search" Solution of Numerical an Statistical Problems," Journal of the ACM, pp. 212-229, 1961.

[27] D. Jakobović, "Test case design," http://gp.zemris.fer.hr/scheduling/Test_cases_design.pdf

ISS

International Conference on

INFORMATION SYSTEMS SECURITY

Steering Committee

Chairs: Stjepan Groš, University of Zagreb, Croatia
Tonimir Kišasondi, University of Zagreb, Varaždin, Croatia

PAPERS

TECHNICAL TRACK

Technical Recommendations for Improving Security of Email Communications

A. Malatras*, I. Coisel* and I. Sanchez*
* European Commission, Joint Research Centre (JRC), Ispra, Italy
{apostolos.malatras; iwen.coisel; ignacio.sanchez}@jrc.ec.europa.eu

Abstract - With billions of emails exchanged worldwide on a daily basis, email is nowadays considered to be one of the most widespread forms of digital communications. The massive deployment of this technology is certainly due to its ease of use and interoperability. This advantageous ubiquity of email communications comes nonetheless at the cost of security, which is often side-lined in favour of maintaining backwards compatibility with older versions of the protocols and considering security features as optional to ensure compatibility between providers. As a consequence, email communications often fall short of protecting the privacy and the authenticity of the information exchanged. Taking into account the fact that email is used to exchange private and personal information, the latter risks become extremely prominent. We review here the outstanding privacy and security risks in worldwide email communications and we describe a set of practical countermeasures, based on combinations of existing standards, which are capable of effectively mitigating the identified risks. Based on this analysis we provide a set of technical recommendations to be followed by email providers in order to enhance security, whilst preserving compatibility in the ecosystem.

Index Terms − secure email, security, privacy

I. INTRODUCTION

Email is nowadays considered to be the most widespread form of digital communication with more than 190 billion of emails being exchanged daily. Email not only accounts for most user communications in corporate environments, but it is also widely used in both residential and mobile environments to support citizens' personal communications. Notions of security and privacy therefore are highly important and have recently gained further attention due to related revelations concerning mass surveillance programs, but also due to the increasing number of exposed security threats and vulnerabilities.

Email systems are by and large based on the Simple Mail Transfer Protocol (SMTP), which was designed decades ago around a simple but powerful premise: interoperability. The goal was to promote a simple yet reliable and ubiquitous protocol for email communications. In the early days of the Internet, the email system quickly started to gain popularity giving birth to a rich and ever growing ecosystem of SMTP compatible email providers. However, the simplicity of SMTP came at the cost of reduced support for security, which originally was regarded as secondary. Unfortunately, Internet has become a much more dangerous place than what we could have ever imagined more than three decades ago. Accordingly, the set of standards used in email communications has been extended over time in an attempt to cope with the ever-changing privacy and security threat landscape.

Nevertheless, the effective mitigation of security risks provided by the current set of standards and implementations is still very limited. This comes as a result of a strong need to preserve the highest possible degree of compatibility between email providers to ensure the successful delivery of messages regardless the security mechanisms supported by the email servers involved in their transmission. Indeed, in order to achieve this goal, until the entire community has adopted a given security upgrade/mechanism, all systems must remain compatible with insecure ones, thus greatly limiting the effective mitigation of security risks. As a result, email communications are currently subject to serious privacy and security risks, which can lead to spoofing of users' identities, interception of confidential communications and also modification of their contents.

In this paper we analyse the current state of the art with regard to the protection of security and privacy of email communications. The paper is focused on the email infrastructure and does not cover end-to-end encryption measures (e.g. such as PGP or S/MIME), which constitute an additional layer of protection. In this article, we review the architecture of email systems and relevant communication protocols and reflect on existing security mechanisms that have been designed by considering the trade-off between security and compatibility with the ecosystem. Accordingly, we examine the limits of this trade-off in search of striking a balance and we thus propose a set of security features to be considered as minimum requirements for a future interoperable and secure email system.

The paper is structured as follows. In section 2 we briefly review the SMTP standard and we analyse the current email ecosystem with regard to the practical implementation of the security features foreseen in the set of email related standards. In section 3, we describe how the current situation could be improved by combining existing technologies in order to better mitigate the current risks whilst keeping a reasonable balance between security and compatibility between email providers. In section 4, we present our proposal for a more secure email system defining a minimum set of requirements to achieve such a goal. Finally, we present our conclusions in section 5.

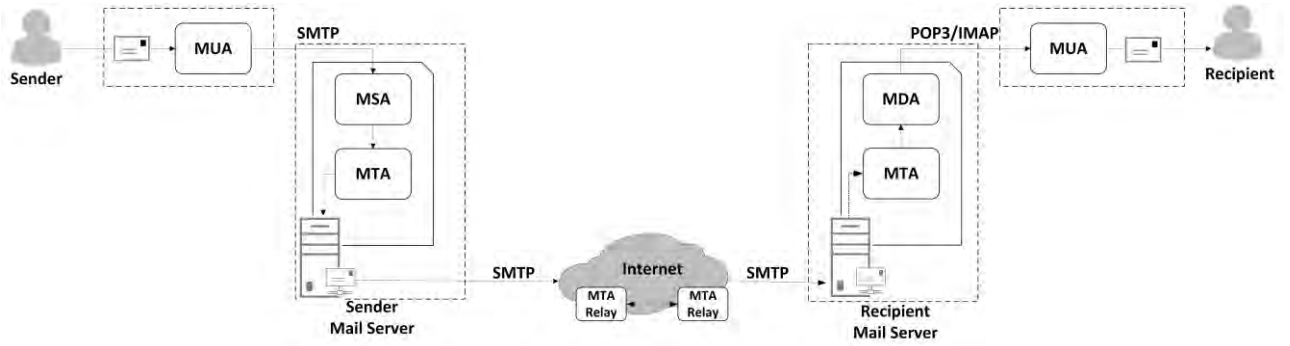


Figure 1 End-to-end architecture of email systems and their communications protocol

II. Analysis of Current Email Ecosystem

Email systems facilitate the exchange of information between interested parties whereby one sender can communicate a message to one or more recipients. The communication functionality offered by email systems involves a scalable and distributed architecture over the Internet, which is illustrated in Figure 1 and is discussed in what follows along with the relevant communication protocols.

A. Email systems architecture

The *sender* is the initiator of the email communication exchange. For senders to send an email to another user or a number of other users they need to utilise a dedicated software program, i.e. the *Mail User Agent* (MUA) that can be standalone programs or web-based ones, namely webmail. Conversely, the *recipient* of the email is the person or persons with whom the sender wishes to communicate, who also makes use of a MUA to receive emails. Mail User Agents have a *Message Submission Agent* (MSA) and *Mail Delivery Agent* (MDA) associated with them in order to handle respectively the transmission and reception of mails to/from *Mail Transfer Agents* (MTA).

The mail server is at the heart of the email system's architecture. Its functionality comprises two main activities, namely managing users' mailboxes and implementing the delivery of email messages between users. The latter process employs the functionality of an MTA, instances of which are located at the mail server of both the sender and the receiver. When the sender compiles an email to be sent, it is stored in a message queue in its mail server, which parses the queue and sends the email over the Internet to the mail server of the receiver, i.e. the receiver's MTA. Strictly speaking, the sender's MTA does not need to be running all the time, merely when a new email appears in the queue. On the other hand, the receiver's MTA needs to be online all the time, listening for new emails that might be delivered. When one such email arrives, it is stored in the local queue that is processed so that emails are stored in the intended receiver's mailbox. The recipient's MUA uses its MDA to collect emails that have been stored in his/her mailbox.

When the mail servers of the client and the recipient are not directly connected, MTA relays are utilised to store email messages and forward them towards the intended destination, whereas even in the case that the recipient's mail server is temporarily offline such a mechanism can be used to enforce robustness in mail delivery by making repeated attempts to deliver the mail on behalf of the server.

B. Email communication protocols

The core components of mail systems, i.e. the MTA clients and servers, communicate using standard protocols, the foremost of which is SMTP [15]. In addition, protocols such as POP3 (Post Office Protocol v3) [20] or IMAP (Internet Mail Access Protocol) [3] are used for the communication between MUAs and MDAs. When webmail systems are utilized, senders and recipients interact with them over HTTPS, whereas in the background SMTP and POP3/IMAP are employed as before. These interactions are highlighted in Figure 1.

SMTP is used for the communication between the sender and the mail server, as well as between the sender's and the recipient's mail servers. SMTP is a communication protocol that defines the exchange of messages between these entities using DNS MX (Mail Exchange) records, which hold the FQDN of the destination mail server [19] to route those messages. MX records can store information about more than one mail server to increase resilience, whereby priorities are assigned to the servers to establish which one will be contacted first. The SMTP protocol is based on the exchange of commands and responses between the involved parties. Originally being ASCII-based and thus having limited functionality, SMTP was subsequently updated to allow for binary content to be exchanged with the introduction of MIME (Multipurpose Internet Mail Extensions) [8], [9], as well as to support advanced features. This was achieved with the introduction of a flexible service extension model that led to Extended SMTP (ESMTP) [14], [16]. The extension model complements the originally limited functionality scope of SMTP. For example, authentication was introduced as a service extension [16] enabling the server to inform the client on the supported authentication mechanisms.

III. Email Security

Email communications were originally built on the principle of simplicity, however the need to enhance the provided functionality and to address emerging concerns such as those involved with security have both led to a plethora of add-ons and extensions. The fundamental

underlying premise of interoperability and compatibility between the email providers is the guideline that necessitates that this patchwork of standards, protocols and their extensions operate smoothly with each other and the existing systems. Nevertheless, this premise leaves the door open to many issues such as spamming, phishing, and more generally a lack of security, since compatibility between providers cannot always be guaranteed due to the great number of email systems that need to apply the new improvements. There exist standardised protocols and techniques capable of enhancing the security of email communications but they are not always used or they are not implemented properly in practice. In some cases those mechanisms are not sufficient to reach a reasonable level of security. We summarize here existing threats together with the countermeasures put in place to tackle them.

A. Communication channel

The SMTP protocol did not originally ensure the confidentiality of the communications. STARTTLS [10] was proposed as an extension to SMTP to encrypt the otherwise cleartext exchange of messages, whereas similar capabilities exist for IMAP and POP3 [21]. STARTTLS utilizes TLS to encrypt all SMTP traffic, thus additionally promoting authentication between the interacting mail servers and hindering potential Man-in-The-Middle (MiTM) attacks. Since STARTTLS complies with the extension model of ESMTP, there is no need for a dedicated port and instead uses the traditional SMTP port. Such design greatly enhances interoperability since it makes use of standard, well-known ports for communication and does not require modifications to existing configurations. The benefits of STARTTLS have spurred its wide adoption that is also strongly encouraged by major players in the field, e.g. Google, Microsoft, Yahoo, Facebook, etc. Figure 2 illustrates the increasing adoption rate of STARTTLS on Gmail.

To further promote compatibility between providers, SMTP servers are configured so as not to require the mandatory use of STARTTLS for their operation. This fact can be exploited by malicious users to break the security of SMTP by means of downgrade attacks [5].

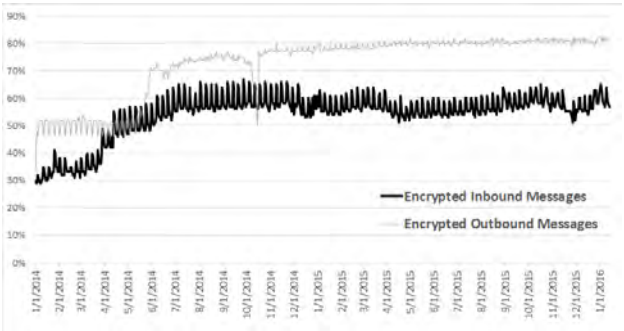


Figure 2 Evolution of STARTTLS adoption based on the raw data published by Google (https://www.google.com/transparencyreport/saferemail/)

STARTTLS is offered to the clients as a service from an SMTP server, among a list of supported services. Should the client decide to use this service, a TLS handshake is firstly carried out in order for sender and receiver to agree on the encryption scheme and common secret key to be used. Subject to the outcome of the TLS handshake, the SMTP session can proceed or be aborted. In case of successful authentication and key agreement, all subsequent communication for the particular SMTP session will be encrypted using the agreed key.

Despite the fact that STARTTLS reinforces privacy, it still nonetheless has certain vulnerabilities. STARTTLS is an opportunistic encryption scheme, meaning that it can be based on either TLS or SSL, with the protocol being decided upon negotiation between the two parties. Whereas TLS is generally considered secure, the same does not stand for SSL with its well-known vulnerabilities [6].

Moreover, another negative aspect of STARTTLS is that the negotiation on whether to use it or not takes place over a plaintext channel. A possible attack would be the interception of the listing of available services from the server to the client and its modification so as not to include the option of STARTTLS. The client would therefore incorrectly assume that STARTTLS is not supported by the server and will opt for an unencrypted SMTP communication. To alleviate this issue the parties could make the use of STARTTLS mandatory and therefore any server or client who does not support it would be sidestepped leading to a lack of compatibility between these MTAs. While such a measure does indeed enhance security, it however does not address another flaw of STARTTLS, namely the use of potentially misleading certificates to verify server and client identities.

Indeed, self-signed certificates are accepted in practice to maintain compatibility between providers leaving the door open to Man-In-The-Middle attacks. The certificates of mail servers can belong to a hierarchical Public Key Infrastructure (PKI) to alleviate such security risk. It roughly means that their certificates have been issued and signed by a trusted certification authority as described in the proposed standard RFC 6818 [24]. Server administrators can deliberately decide to impose that certificates are strongly verified (i.e. its validity as well as those of the issuing authorities) to further enhance security, albeit at the price of loss of compatibility with providers that do not use this type of certificates.

While securing the communication channel using STARTTLS is a significant improvement over traditional SMTP, this solution should not be considered as a panacea. Encryption of the channel can be subject to poor configuration, while additionally it should be noted that the channel is secured only on a per hop basis. Since email usually travels between a series of mail servers prior to reaching its destination, all of the traversed servers should support STARTTLS to ensure that encryption was enforced throughout the entire path. Moreover, when email is processed at the mail server it is not encrypted, therefore a compromised mail server could be exploited to gain illegitimate access to users' emails. End-to-end encryption approaches such as PGP [2] or S/MIME [12] could be applied by the users to protect the confidentiality of their email communications, but this would require efforts and expertise on behalf of the users. Furthermore,

lack of straightforward integration of such approaches in major email clients does not help to promote their wide adoption.

B. Identity spoofing and spam

Malicious users commonly try to masquerade themselves as legitimate ones by faking their source email addresses as part of phishing or spamming campaigns. Those attacks, known as spoofing attacks, are possible because of the security shortcomings related to transferring and routing of email messages between mail servers. In particular, in the delivery process of an email, the sending server uses the DNS protocol in order to find the SMTP server in charge of the recipient's email address domain. In order to ensure the delivery of the message, the receiving server then forwards the email to the intended user without verifying that the server that is sending it is authorised to do so on behalf of the corresponding email domain. As such, it is easy to spoof email identities and perform spamming, phishing or targeted attacks.

Reverse DNS lookup [19] has been used to protect against spammers and identity spoofing attacks. When a server receives an email message, it can use the information contained in the PTR record of the sender's IP address to check its consistency with the MX and A records associated to the domain of the email address. The resolution of a PTR record is the process of resolving an IP address to its associated hostname and it is essentially the inverse process of a DNS lookup. If inconsistencies are found the validation fails and the message is not accepted by the SMTP server, e.g. it can be safely considered as spam.

The Sender Policy Framework (SPF) DNS Resource Records was recently introduced to detect email identity spoofing [13] by defining for a particular domain the hostnames of its computers that are allowed to send emails. The receiver SMTP server can then quickly check against the listing to decide to accept an email or not, based on whether the sending host is in the SPF record or not. SPF information is stored in TXT or SPF type records.

To reinforce the previously introduced security measures, the header of an email can be cryptographically signed to secure its authenticity. As described in the standard RFC 6376 [4], DomainKeys Identified Mail (DKIM) allows signing many different fields of the header (e.g. "from", "to", "subject"...) and includes the resulting signature and technical details in a DKIM-signature field in the mail header. Upon reception of an email, the server obtains the sender's public key from the TXT record of the sender's DNS and uses it to validate (or reject) the signature.

Domain-based Message Authentication, Reporting and Conformance (DMARC) is an additional procedure built on top of both SPF and DKIM, allowing a sender's domain to advertise that the email is protected by these two mechanisms and specifying how to react in case of a failure. DMARC can be seen as a mechanism for policy distribution that furthermore provides feedback to the senders so that they can monitor how effective their authentication is. The DMARC policies are integrated in a TXT record of the DNS record of the sender's domain. More details can be found in the RFC 7489 [17].

C. Protecting DNS communications

The aforementioned security mechanisms increase the trust one can have in a sender mitigating the risks of email identity spoofing and spamming. DNS plays a central role in the email infrastructure acting as a trusted anchor that glues the several protocols together. As a matter of fact, it is crucial to secure the DNS communications and the DNS records to preserve the security brought by the other protocols.

Accordingly, the DNSSEC protocol, described in the proposed standard RFC 6840 [23], has been developed to address the lack of security in the DNS protocol that not only weakens the email ecosystem but more broadly the entire Internet. DNSSEC modifies DNS to add support for cryptographically signed responses making it infeasible for a third-party to modify a DNS record, or pretend it is empty or non-existent, without invalidating its signature. The public key of a domain and the signature of all the information contained in its zone are stored in additional records, namely the DNSKEY and the RRSIG records. The public keys of the child zones are also stored and signed in a dedicated DNS record. Thanks to this record, a chain of trust from the root (and trusted) DNS server to the leaves is built.

While it could increase in a significant manner the security of the DNS communication, the deployment of DNSSEC is surprisingly slow. As stated by OpenDNS in a press release [22], DNSSEC suffers from several drawbacks potentially responsible for this lack of adoption. The main flaws of the standard highlighted in this press release are the questionable security level and the loss of efficiency of the DNS protocols both in terms of computational and communications overhead. As a consequence, OpenDNS made the choice in 2010 to adopt another proposition to secure DNS, namely the DNSCurve solution introduced by Dan Bernstein [1]. In a nutshell, DNSCurve encrypts and authenticates the channel used to carry the query and its response using elliptic curve cryptography. Similarly to DNSSEC, such construction is based on the principle of a hierarchical trust that spreads all the way from the root to the leaves.

These two mechanisms are not exclusive and both help to significantly increase the security of the DNS as a whole, reinforcing the anti-spoofing measures previously described. Additionally, DNSSEC can also be used to facilitate the deployment of STARTTLS. To that end, the DNS-based Authentication of Named Entities (DANE) described in RFC 6698 [11] adds in a DNS record the TLS certificate of the corresponding domain containing the public key to be used to encrypt the communication channel. It avoids the potentially complex procedure to obtain a certificate from a trusted Certificate Authority (CA) while protecting the full system from a potential security breach of said CA.

D. Summary

Email systems have a dominant place in today's Internet having proliferated over the last years mainly due

to their simple yet extensible design. While people rely heavily on email for their communications, the relevant security and privacy risks threaten to diminish the effectiveness of such systems and their widespread adoption. Email messages are inherently tightly linked to private and personal information and thus the notions of confidentiality, integrity and authenticity naturally emerge as being of paramount importance. Considering that email messages have to traverse a series of entities over the Internet, e.g. mail servers and relays, to reach their destination, it becomes evident that they are subject to security risks such as intercepting and eavesdropping. Such risks were exacerbated by the fact that the protocols used for email communication were initially designed with limited consideration for security. As discussed, there have been a number of noteworthy enhancements to email systems and protocols to increase confidentiality and integrity (STARTTLS encryption) and authenticity (DNS-based techniques). Certainly, such enhancements do not constitute a panacea for email security, but they do serve one of the fundamental design goals of email systems, any email must be delivered. This functional requirement has hindered the adoption of more powerful security mechanisms, maintaining nonetheless the ubiquity and transparency of email systems. In what follows, we examine the limits of this design trade-off and propose an approach to strike a balance between both ends.

IV. Recommendations to Promote Secure Email Services

In this section we propose a series of recommendations driven by the following key elements that have emerged from the previous analysis.

- STARTTLS can be an effective means to secure the communication channel between SMTP servers contributing to enhance the protection of the confidentiality of email communications. The current way STARTTLS is deployed, as a type of opportunistic encryption, presents serious drawbacks. Therefore, it needs to be complemented by other security measures to mitigate more effectively the risk of downgrade and MiTM attacks.
- DNS plays a critical role in the protection of the security and privacy of email communications. Being the foundation upon which secure email protocols are built, DNS becomes the weakest link in the chain. Mature standard security protocols providing effective protection of DNS, such as DNSSEC, exist but are rarely used in practice.
- SPF and DKIM have proven to be effective in the fight against spam contributing to mitigate the risk of email identify spoofing.
- The scale between security and compatibility between email providers always tips towards the latter in a non-transparent manner to the end users who are thus unaware of the underlying risks.

Accordingly, we present in the following our recommendations promoting a more secure email ecosystem split in two categories. One considering the combination of existing security protocols and standards that fully preserves the compatibility between email providers, as well as a second one outlining a possible approach to break compatibility in a controlled manner by empowering both service providers and users.

In this respect, we first propose following a security-by-default approach where STARTTLS, SPF and DKIM are required to be fully supported and implemented in combination with DNSSEC by all SMTP servers in order to provide adequate protection against spoofing and secure effectively the SMTP to SMTP communication channel.

Given the critical role of DNS in the effective deployment of the above mentioned email security standards, we recommend that the usage of DNSSEC should be encouraged. Indeed, not only is DNS used to locate the SMTP servers in charge of receiving email for a given domain through the MX record, but also the SPF, DKIM and DMARC standards use it as a trusted anchor to keep the chain of trust solid and secure. DNSSEC is a mature standard capable of providing effective security protection to the DNS protocol and by extension to the entire email ecosystem. In combination with DNSSEC, SPF and DKIM can provide effective mitigation of the email spoofing risk and prevent spam very effectively.

STARTTLS shall be properly complemented with either the strict validation of the server certificates inside the hierarchical PKI (as done in HTTPS) or the usage of DNSSEC in combination with DANE, in order to provide effective protection against MiTM attacks. The extended practice to accept self-signed server certificates as valid should be avoided. As an example, “Let’s encrypt” [7] certificates can be used to secure email communications using STARTTLS and are a burden less secure alternative to the usage of insecure self-signed certificates. DNSSEC can also be used in combination with DANE as an agile method linking a server certificate to the identity of the server. SMTP servers should support validation of certificates protected by DNSSEC and DANE in order to promote this practice and provide secure interoperability.

We strongly believe that compatibility between providers is one of the success factors of email and any initiative to secure the ecosystem shall comply with this principle. Although it is tempting to put forward a proposal to enhance security by breaking compatibility between providers of the email ecosystem (such as by rejecting an email unless the sender uses DNSSEC), we believe that such a proposal wouldn’t be widely adopted. Instead, it is our opinion that we should enhance the awareness level of the email users in such a way that both SMTP servers and eventually email users can take an informed decision as to what extent a given email communication can be trusted. The DMARC standard can be seen as a first step towards this direction by empowering the sender to decide how non-compliant communications should be treated by the recipients. In the future, similar steps could be considered to empower the end user in taking an informed decision by reporting to him in a comprehensive way the security risks of a given email message based on the security measures actually applied during its transmission.

V. CONCLUSION

Email is one of the most common and widespread forms of digital communications, amassing an enormous user base that daily utilize it to exchange information. Protecting and securing this means of communication is therefore of paramount importance in order to promote user privacy and ensure the integrity and authenticity of the exchanged information. Largely based on SMTP, email was not originally designed with security as one of its main requirements, hence the various solutions that have arisen over the past few years to address its security-related shortcomings. The need to maintain full compatibility with the plethora of existing deployments and configurations has nonetheless inadvertently limited the effectiveness of such security enhancements.

In this paper, we reviewed the available solutions and technologies to improve the security of email communications, such as STARTTLS, DNSSEC, DANE, SPF, DKIM, etc. The fragmentation that we observed is attributed to the fact that security of email has been handled a posteriori to the design of email protocols and therefore as an add-on service it needs to retain interoperability with said protocols ensuring compatibility between email providers. Conversely, the necessity for advanced security solutions might require in some cases going against this compatibility in order to enforce more strict security policies. There is a balance to be stricken in regard to this trade-off between security and full compatibility between email providers. In this respect, we presented a set of guidelines to be followed in order to gain the best out of both standpoints. Effectiveness of the proposed guidelines is evidently subject to email providers adopting and incorporating into their deployments the protocols and techniques that we propose.

It is our firm belief that email communications security is in need of a paradigm shift. Whereas there already exist a series of approaches to enhance security of email, low adoption rates by providers and lack of user awareness have limited the impact of these approaches. Raising user awareness is in our view the foundation of promoting email security, following the HTTP to HTTPS transition that involved the users by informing them about the security risks of the former. This level of user awareness spurred the quicker and wider adoption of HTTPS by the majority of providers. Accordingly, if users were made aware of the security and privacy risks that their email communications are suffering from in an informed and interactive manner, this would undoubtedly raise the bar for email providers to integrate security guidelines such as the ones proposed in this paper into their deployments and infrastructures. A recent proposal [18] published after the acceptance of this paper, the SMTP Strict Transport Security, goes in this direction by enforcing STARTTLS with strict validation of the corresponding certificate, preventing the transmission of the email and informing the user in case of a failure. Raising user awareness over email security through openness and transparency is the focus of our future work.

REFERENCES

[1] D. Bernstein, "DNSCurve: usable security for DNS", dnscurve.org, 2009.

[2] J. Callas, L. Donnerhacke, H. Finney, and R. Thayer, "OpenPGP message format", RFC 2440, RFC Editor, November 1998. http://www.rfc-editor.org/rfc/rfc2440.txt.

[3] M. Crispin, "Internet Message Access Protocol -version 4rev1", RFC 3501, RFC Editor, March 2003. http://www.rfc-editor.org/rfc/rfc3501. txt.

[4] D. Crocker, T. Hansen, and M. Kucherawy, "Domainkeys Identified Mail (DKIM) signatures", STD 76, RFC Editor, September 2011. http://www. rfc-editor.org/rfc/rfc6376.txt.

[5] Z. Durumeric, D. Adrian, A. Mirian, J. Kasten, E. Bursztein, N. Lidzborski, K. Thomas, V. Eranti, M. Bailey, and J. A. Halderman, "Neither snow nor rain nor MITM...: an empirical analysis of email delivery security", in proceedings of the 2015 ACM Conference on Internet Measurement Conference, IMC '15, pages 27–39, New York, NY, USA, 2015.

[6] W. El-Hajj, "The most recent SSL security attacks: origins, implementation, evaluation, and suggested countermeasures", Security and Communication Networks, pp. 113–124, 2012.

[7] Let's Encrypt, Linux Foundation collaborative projects. 2016.

[8] N. Freed and N. S. Borenstein, "Multipurpose Internet Mail Extensions (MIME) part one: format of internet message bodies", RFC 2045, RFC Editor, November 1996. http://www.rfc-editor.org/rfc/ rfc2045.txt.

[9] N. Freed and N. S. Borenstein "Multipurpose Internet Mail Extensions (MIME) part two: media types", RFC 2046, RFC Editor, November 1996. http://www.rfc-editor.org/rfc/rfc2046.txt.

[10] P. Hoffman, "SMTP service extension for secure SMTP over Transport Layer Security", RFC 3207, RFC Editor, February 2002. http://www. rfc-editor.org/rfc/rfc3207.txt.

[11] P. Hoffman and J. Schlyter, "The DNS-based authentication of named entities (DANE) Transport Layer Security (TLS) protocol: TLSA", RFC 6698, RFC Editor, August 2012. http://www.rfc-editor.org/rfc/rfc6698. txt.

[12] B. K aliski, "PKCS #7: Cryptographic message syntax version 1.5", RFC 2315, RFC Editor, March 1998. http://www.rfc-editor.org/rfc/rfc2315. txt.

[13] S. Kitterman, "Sender Policy Framework (SPF) for authorizing use of domains in email, version 1", RFC 7208, RFC Editor, April 2014. http: //www.rfc-editor.org/rfc/rfc7208.txt.

[14] J. Klensin, "Simple Mail Transfer Protocol", RFC 2821, RFC Editor, April 2001. http://www.rfc-editor.org/rfc/rfc2821.txt.

[15] J. Klensin, "Simple Mail Transfer Protocol", RFC 5321, RFC Editor, October 2008. http://www.rfc-editor.org/rfc/rfc5321.txt.

[16] J. Klensin, N. Freed, M. T. Rose, E. A. Stefferud, and D. Crocker, "SMTP service extensions", STD 10, RFC Editor, November 1995. http://www.rfc-editor.org/rfc/rfc1869.txt.

[17] M. Kucherawy, and E. Zwicky, "Domain-based Message Authentication, Reporting, and Conformance (DMARC)", RFC 7489, RFC Editor, March 2015.

[18] D. Margolis et al, "SMTP Strict Transport Security", Internet Draft, March 2016. https://tools.ietf.org/html/draft-margolis-smtp-sts-00

[19] P. Mockapetris, "Domain names -implementation and specification", STD 13, RFC Editor, November 1987. http://www.rfc-editor.org/rfc/ rfc1035.txt.

[20] J. G. Myers, and Marshall T. Rose, "Post Office Protocol -version 3", STD 53, RFC Editor, May 1996. http://www.rfc-editor.org/rfc/rfc1939. txt.

[21] C. Newman, "Using TLS with IMAP, POP3 and ACAP", RFC 2595, RFC Editor, June 1999. http://www.rfc-editor.org/rfc/rfc2595.txt.

[22] OpenDNS, "OpenDNS adopts DNSCurve", technical report February 23, 2010. https://blog.opendns.com/2010/02/23/opendns-dnscurve.

[23] S. Weiler and D. Blacka, "Clarifications and implementation notes for DNS security (DNSSEC)", RFC 6840, RFC Editor, February 2013. http: //www.rfc-editor.org/rfc/rfc6840.txt.

[24] P. Yee, "Updates to the internet X.509 public key infrastructure certificate and Certificate Revocation List (CRL) profile", RFC 6818, RFC Editor, January 2013. http://www.rfc-editor.org/rfc/rfc6818.txt

Performance analysis of two open source intrusion detection systems

Boštjan Brumen* and Jernej Legvart*
* University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, Si-2000 Maribor, Slovenia
bostjan.brumen@uni-mb.si

Abstract - Several studies have been conducted where authors compared the performance of open source Intrusion detection systems, namely Snort and Suricata. However, most studies were limited to either security indicators or performance measurements under the same operating system. The objective of this study is to give a comprehensive analysis of both products in terms of several security related and performance related indicators. In addition, we tested the products under two different operating systems. Several experiments were run to evaluate the effects of open source intrusion detection and prevention systems Snort and Suricata, operating systems Windows, Linux and various attack types on system resource usage, dropped packets rate and ability to detect intrusions. The results show that Suricata has a higher CPU and RAM utilization than Snort in all cases on both operating systems, but lower percentage of dropped packets when evaluated during five of six simulated attacks. Both products had the same number of correctly identified intrusions. The results show that Linux-based solutions consume more system resources, but Windows-based systems had a higher rate of dropped packets. This indicates that these two intrusion detection and prevention systems should be run on Linux. However, both systems are inappropriate for high volumes of traffic in single-server setting.

I. INTRODUCTION

With the growth of use of Internet services, the need for their protection against intrusions is arising as well. The security measures introduced by organizations fall into two main categories, namely measures aimed at the detection of intrusions and measures intended to prevent intrusions. Among technologies designed to detect intrusion are intrusion detection systems, or briefly IDS [1]. The task of an IDS solution is to monitor and analyze network traffic and, if necessary, trigger an alarm, drawing attention to a particular type of activity. They can also monitor all system and user activities, check the integrity of files, and perform statistical analysis of unrecognized patterns of activities [2].

There are several IDS systems available on the market. In the open source area, there are two predominant products, Snort with dominant market share and its closest rival, Suricata [3]. They are both working on Unix-like operating systems and Windows, with unixes having almost two thirds of market share in the server area and windows with one third [4].

The aim of the paper is to make a comprehensive comparison of Snort and Suricata on different platforms including Windows and Linux. The related work in the area has shown that such a comprehensive analysis has not yet been done.

II. RELATED WORK

The authors of study [5] have performed an experiment to analyze the functioning of the system Snort 2.8.3 on two operating systems, namely Windows XP SP 2 and Linux 2.6. Identical tests were replicated on virtual Windows and Linux platforms. Snort system performance was, as expected, worse on virtual platforms.

A similar study that compared the operation of the solution Snort 2.8.1 on Linux operating systems 2.6 and Windows Server 2003, was made in a study [6].

Both studies conclude that the performance of the Snort on Linux is much better, although the authors of second study point to the fact that these results were obtained by optimizing WRITE (New API) parameter, which is part of the Linux kernel. The default value is not appropriate because Snort exhibited a lower throughput of network packets in comparison with optimized values.

A detailed analysis of the capacity of solutions was made in the study [7] in which the authors have studied the influence of various characteristics of network traffic, such as the number of network packets, protocol type, traffic speed and size of the network packets of data on the operation of the system Snort. By increasing the number of network packets the percentage of discarded packets increased. The authors made the same conclusion when determining the impact of network packet size and transmission speed. With the same number and network speed the situation with largest packages (1KB) resulted in more discarded packets of larger than smaller size.

The study [8] focused on the operation and use of system resources with Snort 2.9.5.4 under DDOS attack on a host with different hardware settings. The results of the experiment showed that the Snort's capacity of processing packets can be improved by upgrading the hardware configuration, but this does not apply when upgrading a processor with multiple cores. The reason for this is that the Snort does not take advantage of such architecture.

Many researchers have also analyzed the detection of attacks. The authors of the survey [9] carried out the experiment on the dataset DARPA 1998 and for the metrics they used the number of detected threats for each category of attacks and the number of false alarms. Poor results of detection were attributed to the fact that DARPA

test sets consists of attacks that are hard to detect using the signature system.

False alarms problem was presented in the studies [10] and [11]. The authors of the original study used the DARPA 1999 test set and then repeated the experiment with the private data set covering forty days of traffic of a web server of the University of Plymouth. Both studies have highlighted the very serious problem of false alarms. The study [11], which used the real traffic on the network, indicated that the true extent of the problem appears only with the introduction of the system in a real environment.

The study [12] discussed and compared the detection of "0-day" attacks, for which Snort has no signatures. The results of this experiment suggest that Snort can possibly detect certain "0-day" attacks if they are sufficiently similar to already known attacks.

In several studies, the authors also dealt with a comparison between Snort and Suricata. Comparison of IDPS systems in studies [13] and [14] showed a greater burden on system resources when using Suricata. The authors of the study [14] in which they used Snort version 2.9.0.5 and Suricata 1.1 beta 2 found that the CPU and memory load of Suricata is double, which they attribute to the use of multiple threads by Suricata. Snort uses less system resources, but one process instance has difficulties when processing the network traffic in faster networks (more than 300 Mbps).

With optimization of parameters in Suricata it is possible to achieve better performance, but in changing the parameters one needs to be careful not to get the opposite effect. The detection of malicious activity was analyzed and the results show that the system is prune to errors. The authors of the study [14] have wondered whether this is due to the implementation of the module for detecting or processing of certain Snort signatures.

The results of the survey [15] in some points contradict the above-mentioned findings. Suricata (1.0.2) detected all the designed attacks while Snort (2.8.5.2) failed at detecting attacks with specific malicious code (exploits) when CPU load was above 50 %. The authors also found that the percentage of dropped network packets with Suricata greatly increased by reducing the number of processor cores to less than one; this means that the implementation had no dedicated core and other processes were running concurrently. When using multiple processor cores Suricata had a smaller percentage of discarded packets than Snort.

A very detailed and a precise analysis of the performance of these two systems (Snort 2.9.2, Suricate 1.2) was made in the study [16]. It included the default configuration of both solutions and an optimized one where the authors used multiple instances of Snort running concurrently. In testing the default configuration Suricata had better results (more network packets processed per second) than Snort, also when the implementation was limited to a single core processor. By increasing the number of available cores, the capacity of Suricata increased, while the results of Snort did not improve. The results of the experiment showed that the default configuration of the solutions is not designed for large networks, because there are serious problems with scalability on high-performance computer systems. When using 24 cores, the capacity (number of processed packets / sec) of both systems is similar, but Suricata had better results. Optimized configuration used in the experiment is extensible and suitable for application on high-performance computers.

In the study [17], the authors evaluated the functioning of Snort (2.9.0.4) and Suricate (1.0.2) on three different platforms (virtual Linux, Linux 2.6 and FreeBSD). Experiments were performed separately for each protocol (TCP, UDP) with different sizes of network packets (1470, 1024, 512) and at different speeds. The authors concluded that the most appropriate platform for Suricata is Linux and FreeBSD most suitable for Snort (based on the criteria of number of dropped packets).

Some authors compared several software solutions. Study [18] and [19] compare three IDPS systems, namely Snort, Suricata and Bro. In the first study, the authors observed a large number of discarded packets with Snort. At 100Mbs 80% of all network packets were discarded, while the other two solutions exhibited less than 10% of the discarded packets. In the detection of XSS (Cross-Site Scripting) attack, which was repeated at different speeds, Snort was the worst. Contrary to other studies, the CPU load was observed lowest with Snort. The authors of the second study [19] have carried out a detailed analysis and compared the operation of three software solutions (Suricate 1.1 Snort 2.9.1, Bro 1.5) under different conditions, namely the implementation of a variety of attacks with different levels of traffic on the network, and a different number of activated signatures . An important conclusion of the researchers was that the individual software solutions in various attacks respond differently, which is reflected in the percentage of discards of network packets and the number of generated alerts. This may depend on the current traffic on the network and type of the simulated attack.

The overview of the related research has shown that no study has covered both aspects – security and performance – when observing two solutions on different operating system platforms.

A. Aims, scope, and organization of the paper

In this paper, our research goal is to answer the question of which software solution Snort or Suricata, performs better and on which operating system platform in terms of performance.

The paper is organized as follows: in the next section we present the methods and in the section IV the results are presented. The contribution is concluded in Section V with final remarks.

III. METHODS

To answer the basic research question we designed an experiment. The testbed consisted of two computer systems, namely the server and the client. The hardware configuration of the server and of the client are shown in Table I. We used Snort v. 2.9.7.0 and Suricata v. 2.0.7.

The client has generated and sent the network traffic for analysis to the server. The server was installed with

two operating systems, Microsoft Windows Server 2008 SP2 and CentOS 7 (Linux 3.10.0). The switch was done between them at the boot time. Client's primary operating system was Arch Linux. Attacks were generated using the Kali Linux virtual operating system aimed at testing the security and simulation of attacks. The connection between the client and the server was rated at 100 Mbs.

TABLE I. HARDWARE AND SOFTWARE CONFIGURATION OF CLIENT AND SERVER

	Configuration	*Software*
server	Intel Q6600 2.4 GHz, 4GB RAM	OS: Microsoft Windows Server 2008 SP 2, CentOS 7 (Linux 3.10.0) – dual boot IDS: Snort 2.9.7.0, Suricata 2.0.7
client	Intel i3 2.13 GHz, 4GB RAM	Arch Linux, Kali Linux (virtual OS)

In both operating systems we used the same configuration for each IDS (with differences only at specific operating system-related parameters). Snort and Suricata both worked in IDS (Intrusion Detection System) mode. For testing, we used the official free Sourcefire VRT (Vulnerability Research Team) signatures and complimentary signatures from Emerging Threats (ETOpen Rules).

When configuring the experiment, we relied on the methodology for testing the IDS systems, developed by the authors in the study [20]. It defines the general objectives, guidelines and strategies for the selection of test cases designed to measure the effectiveness of IDS software solutions. They are derived from the field of software testing and testing tailored to specific IDS solution. The methodology is intended for IDS systems' developers and administrators that can use it to identify the pros and cons of each solution, and thus choose the most suitable one for their own environment.

For each category of attacks (listed in Table II) we conducted 50 tests, each of which lasted one minute. For each attack, we first captured network traffic using the console application *tcpdump* and saved it in a packet capture (pcap) file. Within each test the identical network traffic was used and replied to avoid possible differences in packages. Each pcap file was sent by the client to the server using *tcpreplay* application with arguments "TopSpeed" and "loop = 0". The network traffic was transmitted at the maximum possible speed.

CPU and memory load was read on the server side after one minute of processing of the network traffic. At the same time only a single instance of IDS was running to prevent any interactions. Data on the number of discarded network packets were obtained by analysis of log files of each solution that are saved according to the interval specified in the configuration files. This functionality is not enabled by default, so it was necessary to explicitly turn it on. To obtain information on the burden on system resources for each individual process, we used the open source library *psutil*. It allows the capture of data on the utilization of the processor, memory, disk, network, and management and control of processes that are currently running [21]. A key reason for choosing this library was the support for both operating systems (Windows, Linux).

TABLE II. CATEGORIES OF USED ATTACKS

Category of Attack	*Description*
Bad Traffic	Module intended for sending network packets of data that are not in compliance with RFC standards
Fragmented Packets	This module sends to a server network packets that were intentionally fragmented into smaller units in order to test the behavior of the system during the assembly of such packages.
Brute Force	Implementation of Brute Force attacks to the FTP server by using the *ncrack* tool
Evasion Techniques	Testing the various techniques used to circumvent IDS system (e.g. encoding of URI addresses, nmap scanning with fragmentation of packages and test cases of circumvention tool Nikto - add a random set of data, use the tab within HTTP requests, etc)
Denial Of Service	This module implements various forms of DOS attacks. It includes three test cases using tools: scapy, Apache Benchmark (ab) and hping.
Test Rules	This module covers a multitude of diverse attacks aimed at stimulating IDS systems and the testing of their operation. It covers Local File Inclusion attacks, SQL inserting, different nmap scans, Nikto scans, etc.

In both operating systems, we have disabled certain services and applications, which could affect the results. On the Windows OS, we have disabled the automatic installation and downloading of updates (Windows Update) and firewall (Windows Firewall). On the CentOS, we have disabled the firewall (*iptables*) and the module that allows the user to monitor the activity of the system and define security policies for SELinux (Security-Enhanced Linux).

Analysis of the data of the experiment is divided into two parts. In the first part we compared the results of measurements of the same solution on different operating systems, while the second part covers the analysis of the two solutions on Windows and Linux platforms, respectively.

In both parts we compared the consumption of system resources and the percentage of discarded packets of data of Snort and Suricata system during the simulated attacks.

For the first part the hypotheses are as follows:

- $H_{0\text{-}1}$: There is no difference in usage of system resources between Snort on Windows and Linux.

- H_{0-2}: There is no difference in usage of system resources between Suricata on Windows and Linux.
- H_{0-3}: There is no difference in number of dropped packets between Snort on Windows and Linux.
- H_{0-4}: There is no difference in number of dropped packets between Suricata on Windows and Linux.

For the second part the hypotheses are as follows:

- H_{0-5}: There is no difference in usage of system resources between Snort and Suricata on Windows.
- H_{0-6}: There is no difference in usage of system resources between Snort and Suricata on Linux.
- H_{0-7}: There is no difference in number of dropped packets between Snort and Suricata on Windows.
- H_{0-8}: There is no difference in number of dropped packets between Snort and Suricata on Linux.

Data analysis was performed with IBM SPSS Statistics ver. 22.

IV. RESULTS

The data on CPU usage, RAM usage and number of dropped packets were measured in each of 50 tests and the average was taken. Normality of distribution of data was analyzed using the Shapiro-Wilk test, which is suitable for small samples. The statistical analyses were performed to check for the difference among different solutions and platforms. The related samples t-test was used where the data were normally distributed (t-value reported); otherwise, the nonparametric Wilcoxon Signed-Rank test was used (z-value reported).

The results of measurements are reported in Table III.

TABLE III. RESULTS OF MEASUREMENTS

		Windows			Linux		
Product	**Attack**	**CPU %**	**RAM Mb**	**Dropped packets %**	**CPU**	**RAM**	**Dropped packets %**
Snort	Bad Traffic	20,21	360,51	87,46	24,69	506,21	77,01
	Fragmented Packets	21,38	363,11	0,00	22,36	505,26	4,54
	Brute Force	20,29	360,41	87,32	24,46	506,08	77,99
	Evasion Techniques	21,31	365,25	95,86	24,74	510,22	91,82
	Denial Of Service	20,63	408,12	97,95	24,81	549,27	96,34
	Test Rules	20,41	452,10	97,91	24,81	607,46	96,06
Suricata	Bad Traffic	21,89	521,32	80,21	65,08	643,30	37,26
	Fragmented Packets	57,8	590,14	7,15	33,77	706,50	11,62
	Brute Force	22,45	521,30	79,22	71,45	643,10	29,51
	Evasion Techniques	58,98	523,21	73,98	82,78	646,50	4,28
	Denial Of Service	49,27	538,70	84,09	84,68	665,30	1,73
	Test Rules	43,33	569,47	89,07	79,46	705,10	1,80

In the first part, we analyze the differences of two systems on Linux and Windows platform (see Table IV).

TABLE IV. RESULTS OF DIFFERENCE TESTS: WINDOWS VS. LINUX

		Difference between Windows and Linux platform?		
		CPU	**RAM**	**Dropped packets**
Snort	**Attack**	**Statistics**	**Statistics**	**Statistics**
	Bad Traffic	t(49) = -17,625 **p < 0,001**	Z = -6,154 **p < 0,001**	t(49) = 35,223 **p < 0,001**
	Fragmented Packets	t(49) = -3,011 **p = 0,004**	Z = -6,154 **p < 0,001**	Z = -6,036 **p < 0,001**
	Brute Force	t(49) = -16,666 **p < 0,001**	Z = -6,154 **p < 0,001**	t(49) = 35,633 **p < 0,001**
	Evasion Techniques	t(49) = -20,746 **p < 0,001**	Z = -6,154 **p < 0,001**	t(49) = 56,024 **p < 0,001**
	Denial Of Service	t(49) = -21,428 **p < 0,001**	Z = -6,154 **p < 0,001**	t(49) = 42,824 **p < 0,001**
	Test Rules	t(49) = -17,957 **p < 0,001**	t(49) =-95,199 **p < 0,001**	t(49) = 35,378 **p < 0,001**
Suricata	**Attack**	**Statistics**	**Statistics**	**Statistics**
	Bad Traffic	Z = -6,154 **p < 0,001**	Z = -6,174 **p < 0,001**	Z = -6,154 **p < 0,001**
	Fragmented Packets	Z = -6,155 **p < 0,001**	Z = -6,154 **p < 0,001**	t(49) =-8,041 **p < 0,001**
	Brute Force	Z = -6,154 **p < 0,001**	Z = -6,195 **p < 0,001**	Z = -6,154 **p < 0,001**
	Evasion Techniques	t(49) =-35,444 **p < 0,001**	Z = -6,157 **p < 0,001**	t(49) = 396,627 **p < 0,001**
	Denial Of Service	t(49) =-61,699 **p < 0,001**	Z = -6,181 **p < 0,001**	t(49) = 448,929 **p < 0,001**
	Test Rules	t(49) =-81,889 **p < 0,001**	Z = -6,157 **p < 0,001**	t(49) = 1020,42 **p < 0,001**

The Table IV shows that there is a statistically significant difference on resource usage between Windows platform and Linux platform for both IDSs, Snort and Suricata for all six simulated attacks. Table V lists the situations in which either of the operating systems was superior.

TABLE V. SUPERIORITY OF OPERATING SYSTEMS

	CPU	**RAM**	**Dropped packets**
Suricata	Windows (except fragmented packets)	Windows	Linux (except fragmented packets)
Snort	Windows	Windows	Linux

The results show that both solutions consume fewer resources on Windows than on Linux, but the number of dropped packets is lower when using Linux; except for the fragmented packets attack where Windows platform is better.

Thus, all hypotheses H_{0-1} to H_{0-4} need to be rejected.

In the second part, we compared Snort and Suricata against each other on Windows and Linux platform, respectively. Results of comparisons are presented in Table VI.

The aim of this comparison is to determine whether Snort and Suricata are any different using the defined metrics, and in which of attacks these differences occur. These data will help us in setting the final verdict as to which solution is preferred for each operating system.

TABLE VI. RESULTS OF DIFFERENCE TESTS: SNORT VS. SURICATA

		Difference between Snort and Suricata system?		
Windows		**CPU**	**RAM**	**Dropped packets**
	Attack	**Statistics**	**Statistics**	**Statistics**
	Bad Traffic	t(49) =-4,396 **p < 0,001**	Z = -6,155 **p < 0,001**	t(49) = 32,606 **p < 0,001**
	Fragmented Packets	Z = -6,155 **p < 0,001**	Z = -6,154 **p < 0,001**	Z = -6,171 **p < 0,001**
	Brute Force	t(49) =-6,954 **p < 0,001**	Z = -6,155 **p < 0,001**	Z = -6,155 **p < 0,001**
	Evasion Techniques	t(49) =-54,084 **p < 0,001**	Z = -6,154 **p < 0,001**	t(49) = 227,523 p < 0,001
	Denial Of Service	t(49) = -47,135 **p < 0,001**	Z = -6,154 **p < 0,001**	Z = -6,155 **p < 0,001**
	Test Rules	t(49) = -55,930 **p < 0,001**	t(49) =-111,019 **p < 0,001**	t(49) = 191,892 **p < 0,001**
Linux	**Attack**	**Statistics**	**Statistics**	**Statistics**
	Bad Traffic	Z = -6,154 **p < 0,001**	Z = -6,156 **p < 0,001**	Z = -6,154 **p < 0,001**
	Fragmented Packets	Z = -6,156 **p < 0,001**	Z = -6,154 **p < 0,001**	t(49) = -7,682 **p < 0,001**
	Brute Force	Z = -6,154 **p < 0,001**	Z = -6,154 **p < 0,001**	Z = -6,154 **p < 0,001**
	Evasion Techniques	t(49) =-265,015 **p < 0,001**	Z = -6,154 **p < 0,001**	Z = -6,156 **p < 0,001**
	Denial Of Service	t(49) =-201,991 **p < 0,001**	Z = -6,154 **p < 0,001**	t(49) = 450,381 **p < 0,001**
	Test Rules	t(49) =-178,780 **p < 0,001**	Z = -6,154 **p < 0,001**	t(49) = 238,898 **p < 0,001**

The results show that Snort consumes statistically significant fewer resources than Suricata on both Windows and Linux platform. However, the number of dropped packets is lower with Suricata than with Snort, except for the fragmented packets attack where Snort is better.

Thus, all hypotheses H_{0-5} to H_{0-8} need to be rejected.

V. DISCUSSION AND CONCLUSION

Comparison of both solutions on Windows and Linux platforms showed that both Snort and Suricata use more resources on the Linux operating system. The CPU usage of each solution depends not only on the simulated attack, but also on the operating system on which the IDS is implemented.

When comparing Snort and Suricata, we found that Suricata used more system resources on both platforms, but was at the same time more efficient with fewer dropped packets (except for Fragmented packets attack).

The percentage of discarded packets is the most important metric as it tells what portion of the received data packets IDS is not capable of analyzing.

Snort on Linux recorded a smaller percentage of discarded packets compared to Windows in five out of six attacks - except with fragmented packets attack. Similar results were observed for Suricata. Differences in percentage of discarded network packets between the two platforms are much greater with Suricata. From this we can conclude that operating system affects the percentage of discarded packets of individual system, but the impact is greater with Suricata than with Snort. The latter had the highest average value of discarded packets on both platforms. Snort was better only in handling the fragmented packets attack, in which it had lower average value on both platforms. That management of fragmented packets is much more efficient with Snort.

At the beginning we have asked the following research question: Which software system is best suited for a particular operating system?

We have already mentioned that the key metric to be used for the selection of a IDS should be the number of discarded packets. Based on the results we can conclude that Suricata is superior on both operating systems in five out of the six simulated attacks. Suricata is, at the same time, performing substantially better on Linux platform considering the dropped packets metrics. The CPU usage and RAM usage is approximately 25 % higher on Linux, than on Windows, but the absolute values are still low (CPU: 70 % on Linux vs. 40 % on Windows; RAM: 540 Mb vs. 670 Mb).

The results show that Linux-based solutions consume more system resources, but Windows-based systems had a higher rate of dropped packets. This indicates that Windows-based solutions are inappropriate for these two intrusion detection systems.

Finally, based on the data from the experiment, one can conclude that Suricata is performing better than Snort, considering the experimental setting.

ACKNOWLEDGMENT

The work was partially supported by the Slovenian Research Agency under grant number P2-0057.

REFERENCES

1. Cavusoglu H, Mishra B and Raghunathan S. The value of intrusion detection systems in information technology security architecture. Information Systems Research 2005; 16(1): 28-46.
2. Vacca JR. Network and system security, 2nd ed. Waltham, MA: Syngress; 2014. ISBN: 0124166954
3. Pascucci M. Top five free enterprise network intrusion-detection tools. 2013; Available from: http://searchsecurity.techtarget.com/tip/Top-five-free-enterprise-network-intrusion-detection-tools. Accessed:

2016-02-23. (Archived by WebCite® at http://www.webcitation.org/6fVYOAsCU).
4. Q-Success. Usage Statistics and Market Share of Operating Systems for Websites, May 2015. 2015; Available from: http://w3techs.com/technologies/overview/operating_system/all. Accessed: 2016-02-23. (Archived by WebCite® at http://www.webcitation.org/6fVYyApEs).
5. Alserhani F, Akhlaq M, Awan IU, Mellor J, Cullen AJ and Mirchandani P. Evaluating intrusion detection systems in high speed networks. Information Assurance and Security, 2009. IAS'09. Fifth International Conference on; 2009. IEEE.
6. Salah K and Kahtani A. Performance evaluation comparison of Snort NIDS under Linux and Windows Server. Journal of Network and Computer Applications 2010; 33(1): 6-15.
7. Bulajoul W, James A and Pannu M. Network intrusion detection systems in high-speed traffic in computer networks. e-Business Engineering (ICEBE), 2013 IEEE 10th International Conference on; 2013. IEEE.
8. Saboor A, Akhlaq M and Aslam B. Experimental evaluation of Snort against DDoS attacks under different hardware configurations. Information Assurance (NCIA), 2013 2nd National Conference on; 2013. IEEE.
9. Brugger ST and Chow J. An assessment of the DARPA IDS Evaluation Dataset using Snort. UCDAVIS department of Computer Science 2007(2007): 22.
10. Tjhai GC, Papadaki M, Furnell SM and Clarke NL. The problem of false alarms: Evaluation with snort and DARPA 1999 dataset. Trust, privacy and Security in digital business. Springer; 2008. p. 139-150. ISBN: 3540857346
11. Tjhai GC, Papadaki M, Furnell S and Clarke NL. Investigating the problem of IDS false alarms: An experimental study using Snort. Proceedings of the IFIP TC 11 23rd International Information Security Conference; 2008. Springer.
12. Holm H. Signature based intrusion detection for zero-day attacks:(not) a closed chapter? System Sciences (HICSS), 2014 47th Hawaii International Conference on; 2014. Waikoloa, HI, USA: IEEE.
13. Kacha C and Shevade KA. Comparison of Different Intrusion Detection and Prevention Systems. International Journal of Emerging Technology and Advanced Engineering 2012; 2(12): 243–245.
14. Albin E and Rowe NC. A realistic experimental comparison of the Suricata and Snort intrusion-detection systems. Advanced Information Networking and Applications Workshops (WAINA), 2012 26th International Conference on; 2012. IEEE.
15. Day D and Burns B. A performance analysis of snort and suricata network intrusion detection and prevention engines. Proc. Fifth Int. Conf. Digital Society (ICDS); 2011. Guadeloupe, France.
16. White JS, Fitzsimmons T and Matthews JN. Quantitative analysis of intrusion detection systems: Snort and Suricata. SPIE Defense, Security, and Sensing; 2013. International Society for Optics and Photonics.
17. Alhomoud A, Munir R, Disso JP, Awan I and Al-Dhelaan A. Performance evaluation study of intrusion detection systems. Procedia Computer Science 2011; 5: 173-180.
18. Cusack B and Alqahtani M. Acquisition Of Evidence From Network Intrusion Detection Systems. Proceedings of the 11th Australian Digital Forensics Conference; 2013. Perth, Western Australia: Edith Cowan University.
19. Thongkanchorn K, Ngamsuriyaroj S and Visoottiviseth V. Evaluation studies of three intrusion detection systems under various attacks and rule sets. TENCON 2013-2013 IEEE Region 10 Conference (31194); 2013. Xi'an, China: IEEE.
20. Puketza NJ, Zhang K, Chung M, Mukherjee B and Olsson RA. A methodology for testing intrusion detection systems. IEEE Transactions on Software Engineering 1996; 22(10): 719-729.
21. GitHub. giampaolo/psutil · GitHub. 2015; Available from: https://github.com/giampaolo/psutil. Accessed: 2016-02-23. (Archived by WebCite® at http://www.webcitation.org/6fVfzoHuz).

Challenges of Mobile Device Use in Healthcare

S. L. R. Vrhovec
University of Maribor, Faculty of Criminal Justice and Security, Ljubljana, Slovenia
simon.vrhovec@fvv.uni-mb.si

Abstract - The use of mobile devices in healthcare offers various new possibilities and is spreading very fast due to this. However, the security aspects of mobile device use are often neglected. According to research, 44 percent of all data breaches happen in healthcare alone. Most of these breaches are directly related to mobile device use. In this paper, we conduct a literature review of mobile device security issues. We identified four pillars of information security in healthcare institutions: administrative measures, physical measures, data exchange with reliable institutions, and adapting of measures to changes. A key success factor in implementing information security is both awareness and training of healthcare workers, and commitment of healthcare institutions to ensure it.

I. INTRODUCTION

Mobile devices, such as smart phones, tablets and laptops, are being used more and more in healthcare as they offer new possibilities for its improvement and optimization of existing work processes [1]. Improved mobility, communication and coordination of healthcare workers, lesser redundancy of medical data, better accessibility of healthcare workers, and improved replying to crisis situations are just a few benefits of mobile device use that can significantly contribute to the accessibility and quality of healthcare services and lower costs of healthcare [1]–[5]. Individual healthcare institutions (i.e., hospitals and medical centers) are already introducing mobile devices to everyday work. However, these are not he beginnings of mobile device use as healthcare workers may already widely use their own mobile devices to access medical data (e.g., by accessing work email on their smart phones). Worldwide, this is actually encouraged and taken advantage of for lowering healthcare institutions' costs [6], [7].

However, mobile device use in healthcare has certain pitfalls in addition to the abovementioned benefits. Healthcare workers are dealing with sensitive medical data of patients therefore providing its security is a top priority. Nevertheless, research shows that 44 percent of all data breaches happen in healthcare alone [6]. This can be attributed to the attractiveness of medical data as it has a relatively high value on the black market.

A large share of data breaches in healthcare is a direct consequence of mobile device use, especially due to mobile device theft and loss [2], [6], [8]. An additional issue is the fact that healthcare institutions cannot follow the fast pace of mobile device adoption and they lag behind it with security measures for protecting medical data [7]. The key vulnerabilities of sensitive medical data security are the use of insecure mobile applications and devices, the lack of security management in healthcare institutions and inadequate education of healthcare workers on technical and organizational aspects of security management of medical data [2], [7], [9]–[11].

Mobile devices are already being used and are going to be used more and more in the future. An overview over formal and informal use of mobile devices in healthcare however does not exist. There is no information on how widespread mobile device use is or how secure it is leaving a big hole in this area of research and practice. In this paper, we present a literature review of use of mobile devices in healthcare institutions, the issues encountered when using mobile devices in healthcare, and foundations for adequate security management of sensitive medical data in healthcare institutions as a step to fill in this gap and to raise awareness of healthcare workers and institutions.

II. METHOD

In this paper, we presents the results of a literature review on the use of mobile devices in healthcare. We adopted the following method:

- Main research lines were first carefully explored. Bibliographic databases were used extensively.
- Web search facilities were used and articles concerning the topic were collected and analyzed.
- Papers were analyzed to determine their main message. Out of scope papers were eliminated.

The primary sources were the *Web of Science* and *Scopus* databases. After extracting from these databases, the papers were reviewed to identify the benefits and drawbacks of mobile device use in healthcare.

III. MOBILE DEVICES IN HEALTHCARE

In recent years, mobile devices (i.e., smartphones and tablets) became a part of our everyday lives. At the same time, they are being used in business more and more frequently and healthcare is no exception. In the United States, mobile devices are officially used in more than a quarter of hospital institutions [1], [2], [7]. Healthcare workers are allowed to use their mobile devices more often and they are also encouraged to use their own devices at work which is also known as *Bring Your Own Device* – BYOD. BYOD is convenient for healthcare workers on one side as they are already familiar with the mobile device that they use every day also at work while on the other side it lowers costs for healthcare institutions [6], [7]. Over 90 percent of healthcare workers use their

own mobile devices at work to access medical data in the United States [6], [7]. Mobile devices are therefore omnipresent in developed countries [9]. Also, healthcare workers are using their own mobile devices at work, e.g., to access work email accounts.

Mobile device use may have many benefits for healthcare. Currently, mobile devices are mostly used as an alternative to workstations (e.g., personal and laptop computers) for accessing medical data [1]. They can be used to access sensitive medical data, such as medical records, treatment plans and history, and update them regardless of where the healthcare workers are located, in the patient room, on a meeting, at patient's home or elsewhere [2]. Use of mobile devices also contributes to efficiency and quality in healthcare as it lowers data redundancy and improves communication and coordination between various stakeholders within a healthcare institution or among several healthcare institutions [1]–[3]. Due to its convenience, the use of mobile devices also positively influences the work satisfaction of healthcare workers [1]. Patients also increasingly tend to use electronic communication channels, such as emails and texts, for contacting healthcare workers and institutions [10], [12].

Mobile devices offer additional benefits that may significantly change healthcare in the near future by connecting them to various IoT (Internet of Things) devices [13]–[15]. Today one can already find tens of thousands mobile applications related to medicine, health, and healthcare in official application stores (e.g., Apple Store, Google Play, and Windows Store) [7]. Mobile devices can already be connected with wearable / IoT devices, such as heart rate monitors, glucose meters etc. that gather medical data and transfer it to the internet in real time [6], [16], [17]. Mobile device producers also produce mobile devices that have integrated the capabilities to capture and transfer data (e.g., Google Glasses) and recently supporting application frameworks have also been presented (e.g., Apple HealthKit, Google Fit, and Samsung S Health) [13].

Additionally, mobile devices can be used to support healthcare during and after crisis events ranging from broader crises arising from natural disasters or epidemics to isolated and limited cases. Mobile devices have the potential to include and connect a broader set of stakeholders. For example, civilian mobile device users can be connected with remotely located healthcare workers while dealing with a crisis locally [4].

IV. Challenges of Mobile Device Use in Healthcare

A. Security Challenges

Fast adoption rate of mobile device use in healthcare also increases the risk of sensitive patient medical data breaches [2]. Mobile device users are adopting new mobile technologies much faster than healthcare institutions can provide adequate conditions for securing sensitive medical data of patients [7]. Lawmakers face similar issues as they struggle to follow the development of new technologies [7], [18]. The fact that 44 percent of all data breaches occurs in healthcare [6] clearly shows the difficulty of securing sensitive medical data in healthcare in comparison to other disciplines. With such numbers, the patients cannot trust healthcare institutions on providing the needed level of privacy, security, and accuracy of medical data and may therefore be reluctant to provide healthcare workers with all their medical data [10]. In order to reestablish the patients' trust into the healthcare system and to sufficiently motivate healthcare institutions to provide adequate levels of security, information on data breaches may be disclosed to the public, such as on The Wall of Shame which can be accessed publicly on www.hhs.gov [6].

There are various reasons for breaching the medical data security. Cybercriminals may try to acquire direct financial gain, commit an electronic fraud, steal the medical identity, or extort the victims [2]. Data can be also breached without such bad intentions, e.g., for satisfying the own curiosity of healthcare workers. However, also in these cases when unauthorized healthcare workers look into the medical data of a patient the privacy of patients is violated. In some cases, for example, when medical data of higher state functionaries is breached, the country's national security can also threatened.

Medical identity theft is spreading fast worldwide. The cybercriminal (or the medical identity buyer) presents himself to the healthcare institution as the victim and benefits healthcare services in victim's name or in this way gains access to forbidden medical substances [6], [8]. Medical identity theft is very tempting for cybercriminals as it is not difficult to steal a mobile device from healthcare workers and the stolen identities can be sold for a considerably higher price than stolen credit cards on the black market [6]. In addition to the financial consequences, medical identity theft can also have severe medical consequences. By cybercriminals benefiting from healthcare services, the victim's medical record is contaminated with their medical data which may in turn affect victims future treatment, e.g., confusion in diagnosis stage, serious medical damage, or even death [6], [8].

Cybercriminals most often try to attack mobile devices through email and malware [2]. However, this is far from the biggest problem as the vast majority of data breaches in healthcare, over two thirds, is directly connected with mobile device loss and theft [2], [6], [8]. While dealing with an emergency situation, the healthcare workers might unintentionally leave the mobile device unattended for a short while which offers an opportunity for its unauthorized use or theft [2].

The use of common security measures, such as mobile device locking and using encryption on memory cards, would incapacitate the thieves from breaching the data stored on the mobile device [2]. However, it is worrying that most healthcare workers never even lock their mobile devices [9]. In addition, most mobile device users are not aware or do not know about the security and privacy aspects of mobile device use [2], [7], [9]. Also, it is very common in healthcare institutions to use simple passwords (e.g., abcd), use general passwords for the whole department (which is usually written on a piece of paper

in the nurse's room) or not to use a password at all [2]. Providing sensitive medical data security is a shared responsibility however healthcare institutions are not aware of this enough as only 38 percent of them defines a formal policy of mobile device use [2], [7], [10].

Mobile applications are being developed very fast. At the same time there are no standard methods for ensuring the development of secure mobile applications that mobile application developers could take advantage of [7]. The developed mobile applications thus do not provide adequate security [19]. Also, the security of mobile applications use by healthcare workers may be questionable when the mobile applications are not developed by healthcare institutions or their trusted partners. The sensitive medical data of patients may be particularly vulnerable and exposed [20]. Mobile applications are not the only ones that cast doubt on secure use of mobile devices. Mobile devices themselves should provide secure access to sensitive medical data [1]. However, even some of the latest mobile devices lack adequate security and cannot provide secure access at all marking the importance of the decision on appropriate mobile devices [13].

Mobile devices are already being formally introduced into some hospitals. Healthcare workers are using their own mobile devices at work, e.g., to access the work email account, meaning they may be accessing sensitive medical data of patients on their mobile devices without the knowledge of their employers. Currently, there is no comprehensive overview over mobile device use or the trends of mobile device adoption. Further, there is also no information on how well are healthcare institutions prepared to adopt mobile device use. Also, there is no data on how well are healthcare workers who are already using mobile devices at work aware of security and privacy aspects of mobile device use. Even the development strategies of healthcare do not include the development in the area of providing adequate security and privacy (e.g., see the Slovenian healthcare development strategy [21]).

TABLE I. BENEFITS AND DRAWBACKS OF MOBILE DEVICE USE IN HEALTHCARE

Benefits	Drawbacks
Healthcare workers' familiarity with device (BYOD)	Mixing personal and organizational data (BYOD)
Lower costs for healthcare institutions (BYOD)	Higher risk of data breaches
Alternative access to medical data	Higher risk of outdated security measures due to fast developing technology
Lesser location dependence	Higher risk of malware attacks
Lower data redundancy	Higher risk of device theft
Improved communication	Higher risk of device loss
Improved coordination	Low awareness of healthcare workers
Higher work satisfaction	Existing security measures not used
Connection with IoT devices	Unsecure mobile applications
Crisis support	Unsecure devices (wearable, IoT)

This lack of preparations for mobile device adoption in healthcare is problematic as trends from abroad show that mobile devices are already being used and they are going to spread much more in a very fast pace. Without a comprehensive overview of the status of mobile device use in healthcare, the preparation of adequate measures on the national levels is not possible. These measures are however necessary in order to benefit from mobile device use in healthcare. Otherwise, healthcare may face a significant drop in the security of sensitive medical data in healthcare institutions and related financial consequences and loss of trust of patients.

B. Managing Security

Table I summarizes the benefits and drawbacks of mobile device use in healthcare as described above. The main foundation in managing the security of sensitive medical data in working with mobile devices is the establishment of a system for managing cyber security in healthcare institutions. The first pillar of information security management are administrative measures that healthcare institutions need to define, which includes policies and procedures for preventing, determining and responding to data security breaches [2], [10]. Administrative measures should be based on a thorough analysis of security risks in which security risks are first identified and analyzed and then adequate measures for taming them are prepared and implemented [10]. Due to the importance of securing sensitive medical data of patients, new specialized arose, such as privacy officer, who have the authority and responsibility for providing security in dealing with sensitive medical data in healthcare institutions [8].

Implementation of administrative measures cannot be successful in practice if healthcare workers are not adequately trained about it [2]. Healthcare workers may start their training on safe mobile device use as early as during their studies. Usage of mobile devices among medical students is widely spread just as it is among any other student population. Furthermore, medical students work with sensitive medical data already during their studies [9]. It makes sense therefore to start training medical students at least on the basics of safe mobile device use [9].

In healthcare institutions, an overview over knowledge and awareness of individual healthcare workers on information and cyber security while using mobile devices is needed [8]. Training healthcare workers is however a challenging task for the management of healthcare institutions because healthcare workers are typically fully loaded with everyday work, i.e. treatment of patients. Healthcare workers consider of secondary importance all other work, the so called "details", such as information technology and cyber security. It is thus vital to raise the healthcare workers' awareness of the importance of privacy and security of sensitive medical information as well as the consequences of mobile device theft and loss. This has to be done through safe mobile device use training along with strong support from the management. Healthcare workers should be aware that already simple security measures, such as locking the mobile device, setting a password for unlocking and encrypting sensitive

medical data could prevent cybercriminals from stealing the medical identities even if they steal the mobile device [2], [8].

Trainings may be customized for different types of healthcare workers (physicians, nurses, surgeons etc.) and their everyday work. Only in this way the healthcare workers can get all and only the key information on safe mobile device use that they really need every day at work. Any additional training on information security could seriously affect their motivation to spend their precious time on it and compromise the training efforts.

The second pillar of information security management are physical measures that aim to protect the information system as a whole and places and devices that are related to it [10]. These measures have a big potential for lowering the alarmingly high breach record in healthcare as the vast majority of healthcare breaches is attributed to mobile device theft and loss [2], [6], [8]. Physical measures include technological solutions as well as security policies and policies for their use [10]. Similarly to administrative measures, physical measures also cannot be effective unless healthcare workers are adequately trained. Without the support and cooperation of all healthcare workers the physical measures cannot be effective.

The third pillar of information security management refers to organizations that exchange sensitive medical data with healthcare institutions [10]. Mobile devices enable connectivity with various organizations through a number of mobile applications for numerous purposes. It is essential that sensitive medical data is exchanged only between healthcare institutions and organizations that are legally bounded to ensure the security of sensitive medical data and are also capable of doing this which is also recommended to be regularly checked [8], [10].

Nevertheless, it is a little bit utopian to expect contractual partners to ensure the privacy and security of sensitive medical data themselves. For example, email providers can provide adequate security on their email servers however they have no control whatsoever over what is going on with the email once it leaves the servers (e.g., what happens with the received email on a physician's mobile device) [6].

The fourth pillar of information security management is about periodically adapting and maintaining all security measures according to changes in the environment of the healthcare institution and organizational changes that have influence the security of sensitive medical data in any way [8], [10].

V. CONLUSION

In this paper we presented the use of mobile devices in healthcare, the security challenges associated with it, and how to adequately manage information security when dealing with mobile devices. As the use of mobile devices in healthcare is spreading lightning fast as they bring several benefits for all – healthcare institutions, healthcare workers and patients. Healthcare institutions will be thus confronted with a big challenge of how to adequately prepare for mobile device use and ensure a satisfactory level of security and privacy of sensitive medical data of their patients.

REFERENCES

[1] HIMSS Analytics, "2014 Mobile Devices Study," 2014.

[2] L. Storbrauck, "Mobile Device Use: Increasing Privacy and Security Awareness for Nurse Practitioners." p. 32, 2015.

[3] Slovensko društvo za medicinsko informatiko, "Zaključki kongresa MI'2014." p. 3, 2014.

[4] C. H. Ren, W. K. Smith, and J. Christensen, "A Medical System for Supporting Civilian Crisis Response," J. Homel. Secur. Emerg. Manag., vol. 12, no. 2, pp. 299–318, Jan. 2015.

[5] B. Schooley, S. Walczak, N. Hikmet, and N. Patel, "Impacts of Mobile Tablet Computing on Provider Productivity, Communications, and the Process of Care," Int. J. Med. Inform., vol. 88, pp. 62–70, 2016.

[6] Bitglass, "The 2014 Bitglass Healthcare Breach Report." 2014.

[7] B. Martínez-Pérez, I. de la Torre-Díez, and M. López-Coronado, "Privacy and Security in Mobile Health Apps: A Review and Recommendations," J. Med. Syst., vol. 39, no. 1, p. 181: 1–8, 2015.

[8] J. P. McDavid, "HIPAA Risk Is Contagious: Practical Tips to Prevent Breach," J. Med. Pract. Manag., vol. 29, no. 1, pp. 53–55, 2013.

[9] E. C. Whipple, K. L. Allgood, and E. M. Larue, "Third-year medical students' knowledge of privacy and security issues concerning mobile devices," Med. Teach., vol. 34, no. 8, pp. e532–e548, 2012.

[10] The Office of the National Coordinator for Health Information Technology, "Guide to Privacy and Security of Electronic Health Information." p. 62, 2015.

[11] P. Wicks and E. Chiauzzi, "'Trust but verify' - five approaches to ensure safe medical apps," BMC Med., vol. 13, pp. 205:1–5, 2015.

[12] J. P. Attri, R. Khetarpal, V. Chatrath, and J. Kaur, "Concerns about usage of smartphones in operating room and critical care scenario," Saudi J. Anaesth., vol. 10, no. 1, pp. 87–94, 2016.

[13] S. Safavi and Z. Shukur, "Conceptual Privacy Framework for Health Information on Wearable Device," PLoS One, vol. 9, no. 12, p. e114306: 1–16, 2014.

[14] H. Lewy, "Wearable devices- from healthy lifestyle to active ageing," in 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2015, pp. 7748–7751.

[15] V. Gay and P. Leijdekkers, "Bringing Health and Fitness Data Together for Connected Health Care: Mobile Apps as Enablers of Interoperability," J. Med. Internet Res., vol. 17, no. 11, p. e260, 2015.

[16] S. Liu, S. Hu, J. Weng, S. Zhu, and Z. Chen, "A novel asymmetric three-party based authentication scheme in wearable devices environment," J. Netw. Comput. Appl., vol. 60, pp. 144–154, 2016.

[17] K. Zhang, K. Yang, X. Liang, Z. Su, X. Shen, and H. H. Luo, "Security and privacy for mobile healthcare networks: From a quality of protection perspective," IEEE Wirel. Commun., vol. 22, no. 4, pp. 104–112, 2015.

[18] S. K. Garg, C. R. Lyles, S. Ackerman, M. A. Handley, D. Schillinger, G. Gourley, V. Aulakh, and U. Sarkar, "Qualitative analysis of programmatic initiatives to text patients with mobile devices in resource-limited health systems," BMC Med. Inform. Decis. Mak., vol. 16, pp. 16:1–12, 2015.

[19] M. Anwar, J. Joshi, and J. Tan, "Anytime, anywhere access to secure, privacy-aware healthcare services: Issues, approaches and challenges," Heal. Policy Technol., vol. 4, no. 4, pp. 299–311, 2015.

[20] J. T. Kim, U. G. Kang, Y. H. Lee, and B. M. Lee, "Security of Personal Bio Data in Mobile Health Applications for the Elderly," Int. J. Secur. Its Appl., vol. 9, no. 10, pp. 59–70, 2015.

[21] Ministrstvo za zdravje, "Resolucija o nacionalnem planu zdravstvenega varstva 2015 - 2025 (ResNPZV 2015 - 2025)." p. 63, 2015.

Safe Use of Mobile Devices in the Cyberspace

S. L. R. Vrhovec
University of Maribor, Faculty of Criminal Justice and Security, Ljubljana, Slovenia
simon.vrhovec@fvv.uni-mb.si

Abstract – The number of mobile device users is rapidly growing and spreading to new populations, such as seniors and children. Users are increasingly more vulnerable in the cyberspace despite available security measures. This leaves many opportunities for potential attackers as mobile devices are connected to the cyberspace practically 24/7. Unawareness of mobile device users who do not know neither the challenges of safe mobile device use in the cyberspace nor the solutions to them is a major factor contributing to this. In this paper, we address this issue by providing a literature review of key security threats in the cyberspace and available security measures that can help users effectively address them. Even though technical solutions exist for most identified challenges, they cannot protect the users themselves. In addition to technical solutions, mobile device users need to use their own common sense in order to achieve adequate protection.

I. INTRODUCTION

Mobile technologies enable full-time connectivity of mobile device users to the cyberspace. The development of various mobile devices, such as smartphones and tablets, brought the cyberspace to a broad specter of users, from elementary school kids to seniors. Mobile devices carry the capability for adequate user security. However, mobile device manufacturers tend to promote the most basic security settings because it enables users of all specters to use the mobile device without prior learning or training. At the same time, this enables mobile device use without the needed basic understanding of the security aspect. In this way, the available security settings often stay disabled.

The use of security settings, such as using various types of passwords for unlocking the mobile device and encrypting data, is almost entirely left to the users who mostly do not have the needed knowledge. Many mobile devices are due to this unsecured and vulnerable to potential cyber-attacks. Mobile device users also do not pay much attention to the threats in the cyberspace because they are not aware of them and to what extent they can affect them [1]. Without the needed motivation, mobile device users never even bother with advanced security settings that remain alien to them even after years of mobile device use [2]. In the paper, we present a review of key security threats in the cyberspace and available security measures for addressing them.

II. CHALLENGES OF SAFE MOBILE DEVICE USE IN THE CYBERSPACE

In this section we present the key challenges of safe mobile device use in the cyberspace. The presented challenges are independent of the mobile device operating system.

A. Malicious Mobile Applications

Malicious mobile applications can be found even among mobile applications on official application stores, such as Google Play, Apple Store, and Windows Store [18], [11]. These applications want to do more with the mobile device than they are supposed to. They may want to gain access to data they should not need considering their functionality or unduly use mobile device services (e.g., a flashlight handling application accessing photos, contacts, notes, or GPS). Application stores offer various ways to check applications before and after publication [18]. Thus, the share of malicious mobile applications in official application stores is significantly lower than among mobile applications available elsewhere.

B. Malware

There is much less malware for mobile devices than personal computers as the number of mobile malware just passed 12 million in comparison to the number of overall malware closing in to 500 million [19]. Even though,

TABLE I. MALWARE TYPES

Type	Description
Adware	Displays various unwanted advertisements and gathers data about the mobile device user and his activity in the cyber space [3], [4]. In this paper, we do not refer to adware that renders advertisements in order to generate revenue for the author, such as advertisements in free legal games.
Spyware	Enables remote monitoring of mobile device users without their consent or awareness. It is used for espionage (industrial or state) [4]–[7].
Ransomware	Encrypts data (e.g., files) on the mobile device and demands ransom from to user to unencrypt it [8]–[10].
Scareware	Notifies the mobile device user about a nonexistent issue and offers a phony solution for it often requiring a payment to complete [9]–[11].
Backdoor	Enables the attacker to control the mobile device undetected [12], [13].
Botnet	Turns the mobile device into a bot performing various tasks in the background without the knowledge of the user and connects it to a broader net of bots [14]–[17].

malware remains a big issue on mobile devices. We know several types of malware that threaten the mobile device in various ways as presented in Table I.

Malware may delete data from the mobile device or deny access to it, e.g., by formatting the memory card, crashing the operating system or deleting installed applications. Often malware increases the use of device's resources (e.g., increases the CPU load, occupies RAM) that slows down the mobile device and shortens the battery life. Malware can also use mobile operator services, such as making expensive calls and texting, without the knowledge of the user. Also, bots can be controlled remotely and connected to various botnets used to perform attacks on web servers, sending spam email, spreading malware etc. [10], [20]–[22].

Users can easily infect their mobile devices without even noticing. A mobile device can get infected just by visiting the wrong web address that causes a *drive by download* which installs the malware in the background. Another common way of infecting the mobile device is by installing a compromised legal application (e.g., free versions of officially non-free games). As with personal computers, mobile devices can also get infected in more classical ways, such as email attachments or by physical installation of the attacker.

C. Social Engineering

Social engineering is an attack in which the attacker uses deceives the mobile device user to do something that he wants [11]. For example, the mobile device user can be tricked to give the attacker his personal data or install some malware on his mobile device that the attacker can use. The area of social engineering is vast and not necessarily related to mobile devices or computers at all. In Table II, we expose the types of social engineering most related to mobile devices.

TABLE II. THE MOST COMMON TYPES OF SOCIAL ENGINEERING

Type	Description
Phishing	Phishing is an attack done through the internet that exploits the inattention of mobile device users. Realizations of phishing vary from emails with fraudulent links that are similar to official ones (e.g., http://accounts.g00gle.com/) to fraudulent Wi-Fi access points with appealing names, such as "Opatija FREE Wi-Fi", and others.
Baiting	Attackers exploit human curiosity with baits. Baits can be as simple as a seemingly lost or forgotten memory cards that mobile devices users can find. Malware may be installed on the mobile device just by inserting such bait into it.
Smishing and vishing	Similar to phishing, attackers can try to exploit the inattention of mobile device users and create the sense of urgency through texts (SMS phishing = smishing) or phone calls (VOIP phishing = vishing).
Scareware	Scareware can be considered as social engineering malware that targets the mobile device user by raising fear and panic in them.
Shoulder surfing	Shoulder surfing can be considered one of the simplest social engineering techniques. The attacker monitors the mobile device display directly or through a video camera in order to get their password, PIN or other information about the user or his mobile device.

D. Physical Use of Mobile Device

Mobile device is most of the time used by its owner and he may also lend it to others who he trusts. The problematic situations however arise when unauthorized people get in touch with the mobile device [4], [23]. These situations may vary greatly. For example, while chatting in the restaurant mobile device owner's friends might borrow the mobile device unnoticeably and post a funny post on the social network. Such cases may actually be quite frequent and do not have serious consequences as they are not malicious.

However, in other cases, such as health institutions, the consequences can be much more severe. On one hand, with personnel mobile devices it is possible to access sensitive medical data of a large numbers of patients. On the other hand, the attacker can seriously influence the treatment of a patient which may even result in death if he changes the dosage of prescribed medicines for the patient.

E. Social Networks

Social networks are typically global web sites intended for virtual socializing of individuals. They enable mobile device users (and others) to create their own profile through which they communicate with friends, post their thoughts, publish their photos and video tapes etc.

The problem about social networks is that mobile device users intentionally or unintentionally publish a vast amount of personal data, such as contacts, information about user's current location, photos and video tapes from which it is possible to determine sensitive information etc. This information can be abused for creating fake user profiles, planning break-ins into the apartment, preparing social engineering attacks, extortion, and similar [1], [9], [10].

F. Bloatware

New mobile devices often have preinstalled mobile device manufacturer custom developed software in addition to the pre-installed operating system. In addition, the mobile operator can install additional software on the mobile device. We refer to all these pre-installed software as bloatware [13], [24].

Bloatware is meant to make the mobile device more user-friendly by customizing applications for sending texts, making phone calls, managing the photo gallery, playing music etc. It also improves interaction with various services available on the internet, such as cloud synchronization.

However, bloatware consumes additional mobile device resources (e.g., increases CPU usage, occupies RAM memory, and shortens battery life). At the same time it may gather data on the mobile device user and forward it to the mobile operator or the mobile device manufacturer. Due to this, mobile device users decide to remove bloatware in order to improve mobile device performance or limit data gathering. Bloatware removal has however also negative consequences, such as lesser operating systems stability and the emergence of new

security vulnerabilities. Finally, there may be also other consequences of bloatware removal, e.g., warranty void.

G. Spam Email

Spam email or also junk email is each email sent with the intent of imposing certain content. It is a globally omnipresent issue and present on practically all mobile devices. Spam email is problematic from several viewpoints. First, messages often try to deceive the recipients or offer low quality products. Next, senders advertise illegal products and products that pose a risk to the health (e.g., steroids and drugs with unknown or questionable origin). Finally, spam email is also a nuisance to recipients who are not interested in its contents [25].

H. Mobile Device Loss and Theft

Sometimes it happens that users forget the mobile device somewhere, lose it or somebody steals it from them. When this happens, the users do not only lose the mobile device itself. They also lose access to all data saved on it that does not have a backup somewhere (e.g., in the cloud) [2], [23], [26].

Also, the finder or thief of the mobile device gets access to all data stored on it including access to all saved accounts (e.g., Google, Facebook, Snapchat, Instagram, PayPal, and Twitter) unless the mobile device is adequately protected against theft and loss. In addition, the finder or thief can benefit from mobile device services.

I. Mobile Device Damage and Destruction

Mobile devices usually get damaged when users least expect it. A fall to the water or to the ground is just some of the more common reasons. Mobile device users may lose access to some or all of the data on the mobile device in addition to the loss of mobile device value.

J. Short Range Connections

Bluetooth and Near Field Communication – NFC are technologies for wireless connection of devices that enable data transfer between two or more devices on short ranges [12]. The abuse of Bluetooth and NFC is only possible when the connection is enabled. Even so, mobile device users often forget to disable the connection, have it enabled all time around because they use wireless devices (e.g., wireless headphones), do not know how to check if the connection is actually disabled after use, or do not know how to disable it at all.

An attacker can use the short range connections to access practically all data saved on the mobile device (e.g., contacts, messages, user accounts, and photos). There is a limitation however as the attacker needs to be close to the mobile device in the time of an attack. This makes mobile devices more vulnerable where people gather in smaller places, such as the classroom, the bus, or in the restaurant.

K. Wi-Fi Networks

Mobile device users often connect to the internet through Wi-Fi networks as it is often free and offers faster internet connection than mobile data. However, connecting to Wi-Fi networks enables relatively easy capturing and collecting of data related to mobile device user activities, such as his browsing history. Anyone who has sniffing software and is connected to an unprotected Wi-Fi access point can eavesdrop on all traffic on it. Actually, it is not much more difficult to do it as well on protected Wi-Fi networks as it is often relatively easy to get network keys. Only WPA2-Enterprise security is considered as secure nowadays. These Wi-Fi networks require that users have their own passwords which make breaking the security much tougher.

Even though eavesdropping is relatively easy on Wi-Fi networks, it does not mean that it is not possible to connect securely to the internet. Mobile device users can use encrypted connection to websites (i.e. HTTPS) to make content transmitted to a website private. This however leaves a trace of visited websites despite the fact that the content is not disclosed. To tackle this issue, mobile device users can use other means, such as connecting to the Tor network [8], [9], [18].

L. Compromised Operating System

The operating system is compromised if the mobile device user gains more rights than the manufacturer or mobile operator allow. Compromised operating system is closely related to the escalation of privileges. Different operating systems require various degrees of compromise in order to achieve it.

On Android, rooting enables the mobile device user to have full access to the operating system and consequently to the applications, other software and hardware. This is a case of classical escalation of privileges [13], [24], [27].

On iOS, the mobile device user needs to jailbreak the mobile device in order to bypass its hardware security. In this way, the mobile device user gains access to the root folder and system files which enables him to install applications, extensions and themes that are not available on the official Apple Store. The mobile device user violates the end user agreement with the jailbreak and thus voids the warranty. Jailbreak requires tackling the hardware of the mobile device and is therefore considered as more than just escalating the privileges [9], [11], [18], [28].

On Windows, the mobile device can be unlocked by changing the Windows Registry key. The mobile device user can install or run applications that are not available on Windows Store. There are three levels of unlocking, namely developer unlock, interop unlock / OEM developer unlock, and full unlock. Unlocking may be considered as a case of escalation of privileges. However, it cannot be considered as a classical one as the operating system natively supports it.

Rooting, jailbreak, and unlocking the mobile device enable bypassing the device and operating system security mechanisms. This enables mobile applications to work outside the expected limitations. Therefore, compromising the operating system lowers the mobile device security [24].

TABLE III. SOLUTIONS FOR ADDRESSING THE CHALLENGES OF SAFE MOBILE DEVICE USE IN THE CYBERSPACE

Challenge	Solutions
Malicious mobile applications	• Disable unknown sources • Report applications • Enable scanning of applications • Enable harmful app detection • Install antivirus software
Malware	• Update operating system • Enable automatic update of applications • Install antivirus software • Install secure web browser [28] • Double-check URL • Pre-check shortened URL • Check connection security
Social engineering	• Install antivirus software • Install secure web browser • Search parts of text on internet • Double-check URL • Pre-check shortened URL • Check connection security • Disable last character display in password fields • Report spam email
Physical use of mobile device	• Enable mobile device lock • Enable automatic mobile device lock • Enable store lock
Social networks	-
Bloatware	• Disable or remove bloatware • Restrict bloatware
Spam email	• Install antivirus software • Report spam email
Mobile device loss and theft	• Enable mobile device lock • Enable automatic mobile device lock • Enable store lock • Set mobile device administrators • Install remote control application • Encrypt data • Backup data
Mobile device damage and destruction	• Backup data
Short range connections	• Install antivirus software • Disable Bluetooth after use • Disable NFC after use
Wi-Fi networks	• Install antivirus software • Enable open Wi-Fi warning message • Disable shared internet connection • Use virtual private network – VPN
Compromised operating system	• Check if operating system is compromised

III. ADDRESSING THE CHALLENGES OF SAFE MOBILE DEVICE USE IN THE CYBERSPACE

Mobile devices offer various solutions for addressing the challenges of safe mobile device use in the cyberspace as presented in Table III. These solutions are mostly available on all major mobile platforms (Android, iOS, and Windows) except for few that are specific for some platforms or even versions (e.g., *enable harmful app detection* is Android specific). Also, some solutions may not be available on all devices (e.g., *disable or remove bloatware*). The described solutions require little expertise and knowledge, and can be used by average mobile device users that devote enough time to understand their mobile device and its operating system.

Several solutions varying on the operating system may be applicable for some challenges and the same solutions may address more challenges. In the case of social networks, a technical solution does not exist therefore it is up to the mobile device user to use his own common sense to deal appropriately with this challenge.

IV. CONCLUSION

In this paper we overviewed the threats that lurk on mobile device users in the cyberspace. Even though it is sometimes enough to know the threats and use common sense to stay safe from them, the mobile device users can use various technical solutions to adequately secure themselves and their mobile devices in the cyberspace.

REFERENCES

[1] B. Markelj and S. Zgaga, "Comprehension of cyber threats and their consequences in Slovenia," Comput. Law Secur. Rev., 2016.

[2] B. Markelj and I. Bernik, "Safe use of mobile devices arises from knowing the threats," J. Inf. Secur. Appl., vol. 20, pp. 84–89, 2015.

[3] D. Hrestak, S. Picek, and Ž. Rumenjak, "Improving the android smartphone security against various malware threats," in 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015, 2015, pp. 1550–1555.

[4] M. Yesilyurt and Y. Yalman, "Security Threats on Mobile Devices and their Effects : Estimations for the Future," Int. J. Secur. Its Appl., vol. 10, no. 2, pp. 13–26, 2016.

[5] T. Herath and H. R. Rao, "Protection motivation and deterrence: a framework for security policy compliance in organisations," Eur. J. Inf. Syst., vol. 18, no. 2, pp. 106–125, 2009.

[6] A. C. Johnston and M. Warkentin, "Fear Appeals and Information Security Behaviors: An Empirical Study," MIS Q., vol. 34, no. 3, pp. 549–566, 2010.

[7] H. Daojing, C. Sammy, and G. Mohsen, "Mobile Application Security: Malware Threats and Defenses," IEEE Wirel. Commuincations, vol. 22, no. 1, pp. 138–144, 2015.

[8] Kaspersky Lab Global Research and Analysis Team, "Kaspersky Security Bulletin 2013," 2015.

[9] ESET, "Trends 2016: (In)security everywhere," 2016.

[10] C. F. McGuire, "TIM Lecture Series The Expanding Cybersecurity Threat," Technol. Innov. Manag. Rev., pp. 46–48, 2015.

[11] A. Lacerda, R. de Queiroz, and M. Barbosa, "A Systematic Mapping on Security Threats in Mobile Devices," in 2015 Internet Technologies and Applications, ITA 2015 - Proceedings of the 6th International Conference, 2015, pp. 286–291.

[12] P. Faruki, A. Bharmal, V. Laxmi, V. Ganmoor, M. S. Gaur, M. Conti, and M. Rajarajan, "Android Security: A Survey of Issues, Malware Penetration, and Defenses," IEEE Commun. Surv. Tutorials, vol. 17, no. 2, pp. 998–1022, 2015.

[13] H. Zhang, D. She, and Z. Qian, "Android Root and Its Providers: A Double-Edged Sword," in Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS), 2015, pp. 1093–1104.

[14] K. Muthumanickam and E. Ilavarasan, "Demanding Requirement of Security for Wireless Mobile Devices : A Survey," Res. J. Appl. Sci. Eng. Technol., vol. 8, no. 24, pp. 2381–2387, 2014.

[15] B. Rashidi and C. Fung, "A Survey of Android Security Threats and Defenses," J. Wirel. Mob. Networks, Ubiquitous Comput. Dependable Appl., vol. 6, no. 3, pp. 3–35, 2015.

[16] H. M. Jung, I.-S. Hwang, J.-K. Moon, and H.-S. Park, “A security monitoring method for malicious P2P event detection,” Peer-to-Peer Netw. Appl., vol. 9, no. 3, pp. 498–507, 2015.

[17] K. Zhao, D. Zhang, X. Su, and W. Li, “Fest: A Feature Extraction and Selection Tool for Android Malware Detection,” in Proceedings - IEEE Symposium on Computers and Communications, 2016, pp. 714–720.

[18] F-Secure, “Threat Report 2015,” 2016.

[19] McAfee Labs, “McAfee Labs Threats Report - March 2016,” 2016.

[20] V. Hamon, “Android botnets for multi-targeted attacks,” J. Comput. Virol. Hacking Tech., vol. 11, no. 4, pp. 193–202, 2015.

[21] A. M. Memon and A. Anwar, “Colluding apps: Tomorrow’s mobile malware threat,” IEEE Secur. Priv., vol. 13, no. 6, pp. 77–81, 2015.

[22] S. Sheen, R. Anitha, and V. Natarajan, “Android based malware detection using a multifeature collaborative decision fusion approach,” Neurocomputing, vol. 151, no. P2, pp. 905–912, 2015.

[23] D. Dang-Pham and S. Pittayachawan, “Comparing intention to avoid malware across contexts in a BYOD-enabled Australian university: A Protection Motivation Theory approach,” Comput. Secur., vol. 48, pp. 281–297, 2015.

[24] Y. Shao, X. Luo, and C. Qian, “RootGuard: Protecting rooted android phones,” Computer (Long. Beach. Calif)., vol. 47, no. 6, pp. 32–40, 2014.

[25] T. Lee, H. Cho, H. Park, and J. Kwak, “Detection of Malware Propagation in Sensor Node and Botnet Group Clustering based on Email Spam Analysis,” Int. J. Distrib. Sens. Networks, p. #530250, 2015.

[26] A. Das and H. U. Khan, “Security behaviors of smartphone users,” Inf. Comput. Secur., vol. 24, no. 1, pp. 116–134, 2016.

[27] S.-T. Sun, A. Cuadros, and K. Beznosov, “Android Rooting: Methods, Detection, and Evasion,” in SPSM 2015 - Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices, co-located with: CCS 2015, 2015, pp. 3–14.

[28] N. Virvilis, A. Mylonas, N. Tsalis, and D. Gritzalis, “Security Busters: Web browser security vs. rogue sites,” Comput. Secur., vol. 52, pp. 90–105, 2015.

Securing web content and services in open source content management systems

*H. Jerković, *P. Vranešić, **S. Dadić
Zagreb School of Economics and Management,
*IT department, Zagreb, Croatia,
**MBA graduate student,
(hrvoje, pvranesi, stipe.dadic) @zsem.hr

Abstract - Content management systems (CMS) are information systems designed dominantly for managing different types of publicly available web content although they could be used for various other purposes. Today, CMS open source solutions are one of the most popular platforms for developing web sites, portals, web shops and other publicly available content and services. Main problem in this area is lack of proper understanding of security issues and procedures which are frequently leaving content and services vulnerable for various types of attacks. This work focuses on risk analysis of main CMS open source systems from the point of security of data and services. Usual "points of failure" are analyzed, compared to similar solutions and final measures for full protection are proposed.

I. INTRODUCTION

Most used CMS in world today are WordPress, Joomla and Drupal which are all open source software built on PHP with MySQL database support primarily. WordPress alone is used by more than 23.3% of the top 10 million websites as of January 2016 with around 60% of CMS market share and more than 60 million websites altogether [1]. Joomla and Drupal together take around 10% of CMS market share [2]. We'll analyze main causes of security problems in CMS in general and compare their security with enterprise CMS solutions. Since CMS security dominantly depends on security of underlying systems (server, database server and web server) we will briefly cover comparison between usual underlying technology of three most used CMS (Linux server, MySQL,MariaDB or PostgreSQL as database servers and Apache or Nginx as web servers) with underlying Microsoft technology for running CMS (Windows Server, MS SQL as database server and .NET based CMS running on IIS[1] as web server). We'll give analysis of most abused and most severe cases of vulnerabilities and try to identify most common ways in which they are being exploited. Chapter 9. analyses importance of community in preventing widespread abuse of newly discovered vulnerabilities. Chapter 10. analyses effectiveness of CDN[2] and their ability to eliminate standard threats. Final goal of paper is to define overall best practices for securing open source CMS.

II. CURRENT RESEARCH

Research from 2013 [3] compared security of three most used CMS (WordPress, Drupal and Joomla) and another from same year [4] explored security and performance of main CMS. Other researchers are usually focused on general accessibility, usability and security [5]. Responsible development of CMS with security in mind is topic of research for researchers in [6]. Quality of plugins and their effect on security of CMS is another popular topic of research [7],[8]. Researcher also explored about software vulnerability markets [9] and also general comparisons researches are done without special focus on security related to three main CMS [10]. Some researchers also focused on various tool for testing security of CMS [11],[12]. CMS are also frequently analyzed from point of application usefulness for various data integration purposes in LMS[3] and SIS[4] [13] or as comparison tool for support in learning environment where they could take role of Learning CMS (LCMS) [14]. We can conclude that there is no recent research in area of CMS security that covers influence of CDN networks and responses of developer community when vulnerability is discovered. This research also covers analysis of all other relevant major security factors with analysis of latest statistics that is of relevance to overall security of CMS.

III. CMS FEATURES AND GENERAL PROBLEMS

CMS usually supports modular and extensible framework for installation of various plugins so common CMS core functionalities could be easily extended with many additional features. WordPress.org directory contains 42.807 plugins which are downloaded 1.16 million times (as of January 2016) [15]. CMS core functionalities are usually secure if updated regularly. Most common problems with such a large plugin base is that majority of plugins are out of date, not properly working with current release of CMS, they are bloated with additional content and many plugins are inherently unsecure because of poor outdated or reused code. Attacker identification of website with outdated CMS and or outdated vulnerable plugins can lead to security problem. Unfortunately all major CMS could be successfully identified remotely: WordPress, Joomla and Drupal are all in that category. Experts agree that security by obscurity is not true security but still there is agreement

[1] Internet Information Services
[2] Content Delivery Networks
[3] Learning Management Systems
[4] Student Information Systems

that obscurity is extremely helpful in situation where users are running major CMS with number of publicly disclosed vulnerabilities. So general security problems with CMS are: easy remote identification of system, poor programming practices while creating plugins, lack of oversight while submitting plugins, poor inspection of potential security issues with plugins and lack of auto-update of CMS installations and plugins which are leaving system open for knows CMS and plugins vulnerabilities.

IV. SERVER VULNERABILITIES

Experts agree that securing application itself is futile without server-side validation measures[16]. So in this chapter we'll compare general security of Linux servers (which are most commonly used for running three most used CMS) with security of other servers (mainly Windows servers). This is important so we can get clear understanding of overall security of supporting CMS technologies. Table 1. shows statistics based on data from NVD[5] [17]. We can conclude that Linux servers have approximately the same number of high level

TABLE 1. OPERATING SYSTEMS VULNERABILITY THREATS

	Number of vulnerabilities			
OS	**Total**	**High**	**Medium**	**Low**
Apple Mac OS X	147	64	67	16
Linux Kernel	119	24	74	21
MS Win Server 2008	38	26	12	0
MS Win Server 2012	38	24	14	0

vulnerabilities as Windows Server 2012 (which is still one of the most used Windows Server OS). Linux Kernel has significantly larger number of medium size threats detected (74) in comparison with Windows Server 2012 and 2008. Since those are lower level threats, we can argue that with proper maintenance all main server types offer adequate level of security for hosted services and applications. Experts agree that although proper server maintenance is important, main cause of security problems are usually caused by outdated applications [18].

V. WEB SERVER VULNERABILITIES

Three major open source CMS have to be hosted on web servers that support PHP/MySQL. So security of application greatly depends on security of web server. This chapter brings comparison of general security of web servers commonly used to host three major CMS and other currently used web servers. Apache and Nginx are usual choice for hosting major open source CMS so comparing their vulnerabilities with other web servers could give us better understanding of overall security of major CMS. We've tried to find relevant statistics that will show us usage of current most used web servers. Netcraft Web Server Survey explores usage statistics of current most used web servers. Survey has run since August 1995, exploring the Internet to find new websites. Every month, an HTTP request is sent to each site, determining the Web server software used to support the site through careful inspection of the TCP/IP characteristics of the response [19, 20]. Statistics shows that Apache web server is still most used web server (51%) and if we exclude Google (8%) and Windows servers (11%) next in line suitable for PHP/MySQL open source applications is Nginx with 14% of market share and steadily growing in usage.

If we compare vulnerabilities of Apache HTTP server, Nginx and IIS number of reported CVE[6] [21] since 2010 is: Nginx – 15 , Apache HTTP - 97 , IIS– 27. If we analyze number of reported CVE just for 2015 results are: Apache – 9, IIS – 2 and Nginx – none. So we could conclude that Nginx could be more secure choice for hosting any popular CMS.

VI. DATABASE SERVER VULNERABILITIES

Final goal of attacker is usually access to database so securing database is probably most important issue in overall security of web application. This chapter brings comparison of general vulnerabilities of databases used by three most used CMS (MySQL, MariaDB and PostgresSQL) and MS SQL Database as usual database of choice for .NET based CMS. Most popular CMS usually are using MySQL databases but since Oracle acquired it many users are turning to MariaDB which is a community-developed fork of the MySQL and it intends to remain free under the GNU GPL [22]. Researchers in [23] made comparison of both DBMS in 2012. but no recent study was published. PostgreSQL could also be viewed as viable alternative since many CMS are supporting it. Microsoft SQL Server in this case is proprietary alternative mainly used as database server for .NET based CMS and other web applications. *TABLE 2.* shows total vulnerabilities for each DBMS in last three years. In 2015. MySQL accounted for growing 76

TABLE 2. TOTAL VULNERABILITIES PER YEAR FOR MAIN DATABASE VENDORS

Total vulnerabilities per year	**Oracle MySQL**	**PostgreSQL**	**MariaDB**	**MS SQL Server**
2013.	66	6	3	0
2014.	63	9	1	2
2015.	76	3	0	3

detected vulnerabilities but it's also worth noting that none of those vulnerabilities have score higher than 8/10. Number of total vulnerabilities for MariaDB is 4 and for MS SQL is 5 in last 3 years.

So we can conclude that MariaDB and PostgreSQL are more secure alternatives when it comes to choice of database support for CMS.

VII. CONTENT MANAGEMENT SYSTEMS VULNERABILITIES

Main cause of security problems are CMS application itself. We'll analyze security situation for main most used CMS: WordPress, Joomla and Drupal and compare that to DotNetNuke (.NET based CMS) and phpBB (PHP based CMS with market share equal to DotNetNuke). Table 3. shows number of detected vulnerabilities for compared CMS. Results vary from year to year but we can see that in 2015. situation is similar for all three main CMS. WordPress has much larger share of the CMS market than

[5] National Vulnerability Database

[6] Common Vulnerabilities and Exposures

TABLE 3. TOTAL VULNERABILITIES FOR LAST THREE YEARS FOR CMS COMPARISON

Total vulnerabilities	Word Press	Joomla	Drupal	DotNet Nuke	phpBB
2013.	20	11	14	2	0
2014.	29	10	35	3	0
2015.	11	13	10	1	2

Joomla or Drupal [2], nevertheless number of discovered vulnerabilities is aligned with those discovered in other popular CMS in year 2015. Figure 1. shows WordPress common types of exploits for period 2004-2015. Cross Site Scripting (XSS) with 73 CVE has leading number of vulnerabilities but after analysis of each vulnerability for 2015 we can see that CVSS[7] [21] score for all of them is between 3,5 and 4,3. Common example of such CVE is CVE-2015-5734 : Cross-site scripting (XSS) vulnerability in the legacy theme preview implementation in *wp-includes/theme.php* in WordPress before 4.2.4 allows remote attackers to inject arbitrary web script or HTML via a crafted string[24]. In regard with integrity impact of this threat and threats of similar level; we should add that modification of some system files or information is possible using this attack, but the attacker does not have control over what can be modified. Also scope of what the attacker can affect is limited and there is no impact to the availability of the system.



FIGURE 1. WORDPRESS VULNERABILITIES BY TYPE [25]

Most dangerous types of attacks are related to SQL injection vulnerability because these attacks have higher availability impact on the system. Usually there is reduced performance or interruptions in resource availability. In year 2015. only one CVE of such type is reported for WordPress with score 7.5 and situation is similar with Drupal. All other CVE discovered for WordPress are rated below 5.0. Joomla statistics for 2015 are showing 8 CVE registered with severity of 7.5, majority of them are SQL injection vulnerability which puts Joomla on the 1st place as most vulnerable CMS.

General hacking probability of CMS is defined in Figure 2. Hackers dominantly focus on CMS with greater market share, than on existing CMS exploits and server vulnerabilities.

[7] Common Vulnerability Scoring System

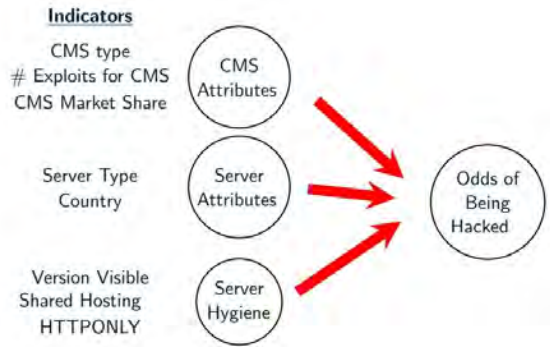


FIGURE 2. CALCULATION OF HACKING PROBABILITY

DotNetNuke CMS is most used .NET based CMS but it takes only 0,6% of market share and with only 6 vulnerabilities discovered in last 3 years. Intensity of usage greatly influences number of discovered vulnerabilities so we cannot easily conclude that DotNetNuke is safe choice. To prove that we have taken into comparison phpBB - PHP based CMS which has almost equally low market share as DotNetNuke (0.5%). *TABLE 3.* clearly shows that point is proven since total number of vulnerabilities in last 3 years is 6 for DotNetNuke and only 2 for phpBB.

We can conclude that overall security of main most used PHP based CSM applications is equal or better than of those .NET based.

VIII. ANALYSIS OF HIGH SCORE CMS VULNERABILITIES

In order to propose adequate security procedures we must analyze common vulnerabilities and common ways in which they are being exploited.

There are many ways in which CMS vulnerabilities could be exploited. Besides manual methods, recently many automated solutions appeared in forms of web kits for exploiting common vulnerabilities. Analysis of

TABLE 4. WORDPRESS VULNERABILITIES RANKING BY CVSS SCORE

CVSS Score	Number Of Vulnerabilities	Percentage
0-1		0.00
1-2		0.00
2-3	6	6.90
3-4	4	4.60
4-5	40	46.00
5-6	12	13.80
6-7	16	18.40
7-8	7	8.00
8-9		0.00
9-10	2	2.30
Total	**87**	

vulnerability severity for WordPress for period February 2012. to February 2016. shows 9 threats with score higher than 9 (see Table 4.). Severe threats that have CVSS score higher than 7 are falling in following types of attacks: SQL injection, code execution, denial of service (DoS) and Cross-site request forgery (CSRF).

SQL injection allows attacker injection of destructive SQL code in order to misuse database of CMS application [26]. This is a code injection technique that exploits a security vulnerability occurring in the database layer of an application and service. This is most often found within web pages with dynamic content [27]. In 2015 serious SQL injection vulnerability was discovered in WordPress. SQL injection vulnerability in the *wp_untrash_post_comments* function in *wp-includes/post.php* in WordPress before 4.2.4 allows remote attackers to execute arbitrary SQL commands via a comment that is mishandled after retrieval from the trash [28]. This vulnerability could for example allow attacker to change default email address of all users or only of administrator of web site to whatever he wants. But still attacker have to get database access in order to modify meta value of post metatag and execute injection code.

Second common scenarios include compromised plugins. Yoast SEO WordPress plugin [29] is used on millions of WordPress sites for improving sites presence on search engines [30]. In 2015. blind SQL injection vulnerability was discovered in that plugin. WPScan Vulnerability Database, which is Vulnerability Database for WordPress its plugins and themes, issued an advisory after responsibly disclosing the vulnerability to the plugins author. By having a logged-in author, editor or admin, and by visiting malformed URL, a malicious hacker could change sites database. This is a case where CSRF [8] vulnerability allowed blind SQL injection. Blind SQL injection is a type of SQL Injection attack that asks the database true or false questions and determines the answer based on the applications response. This attack is often used when the web application is configured to show generic error messages, but has not mitigated the code that is vulnerable to SQL injection.Hosting companies that are dominantly hosting WordPress sites (like Pressable and Siteground) reacted in two ways: Pressable initiated automatic update of all affected sites and Siteground has added a temporary fix to tide customers over in the meantime before they have the chance to update. The company added new security rules to its WAF[9], which actively filtered any possible incoming hacking attempts that try to exploit the vulnerability.

This is classical situation where flaw in one plugin can compromise millions of WordPress sites. In this chapter we are going to analyze response of the various actors and show how joint efforts of the members of community can produce excellent results. Team from WPScan Vulnerability Database promptly alerted Yoast plugin team about issue first and waited for release and deployment of update before releasing details of the issue. Immediately after update for compromised plugin was released, WordPress pushed that update as platform security update for all WordPress sites that were using Yoast plugin. Users that were using cloud based Wordpress.com plugin called Jetpack could automatically update all critical plugins. Users can turn this option on in Jetpack dashboard administration [15]. Obviously choosing a professional hosting service can greatly improve on security of the site as demonstrated in this case. Site owners are usually concerned with bandwidth speed and space on servers without desire to spend extra money on securing their web sites. This is common problem and it's frequently neglected. Many realize potential and scope of threats only when they feel consequences on their own situation.

We can conclude that most frequently reported types of attacks are mainly using SQL injection vulnerabilities or CSRF in combination with SQL injection. Also we can conclude that having organized and interactive community (like one present around WordPress development) and professional hosting is essential for security of CMS solution.

IX. CLOUD PROTECTION FROM CONTENT DELIVERY NETWORKS

Additional layer of security can be achieved with CDN. CDN is a system of distributed servers that deliver webpages and other Web content to a user based on the geographic locations of the user, the origin of the webpage and a content delivery server [31]. CDN WAF offers protection against XSS and SQL injection attacks described in TABLE 4. WAF is frequently updated in cooperation with leading CMS providers so it can prevent new attacks on sites in CDN [32, 33]. Such secure delivery networks can help prevent or mitigate the most common attacks against websites. A case study from a leading provider of content delivery services (Akamai Technologies) illustrates CDN effectiveness, as authors in [34] described it. Akamai Technologies reported that distributed denial-of-service (DDoS) attacks against its customers are increasing in terms of both the bandwidth and the number of requests generated by the attackers. From 2009 to 2014, the size of the largest attack grew year by year from 48 to 68 to 79 to 82 to 190 to 321 in Gbit/s. Another global leader in CDN technologies, CloudFlare, confirmed that they mitigated the largest-ever recorded DDoS attack, which peaked at 400 Gbit/s in February 2014 [35]. At the same time, the number of packets per second in the largest attack grew from 29 million to 169 million. These kind of web application attacks are the most common cause of data breaches today [34].

In light of that, Google also joined this battle recently by filing patent request for "Cloud based firewall system and service". System will provide protection to customers sites from attacks, leakage of confidential information and other security threats. Such firewall system can be implemented in conjunction with a CDN [36]. In time to come cooperation between Google and major CDN providers like Akamai or CloudFlare will probably intensify. Collaboration recently extended beyond just business partnership and into joint collaboration on new patents like "Supporting secure sessions in a cloud-based proxy service" which is Google patent but original assignee of the document is CloudFlare, Inc [37].

Unfortunately cloud solutions do not yet offer ideal environment. Some researchers have explored novel "origin-exposing" attack vectors which can be used to discover the IP address of the server where website is hosted [38]. To assess the impact of the origin-exposing vectors on the security of CDN protected websites

[8] Cross-Site Request Forgery
[9] Web Application Firewall

researchers consolidate all vectors into CloudPiercer. This is automated origin-exposing tool, which they use to conduct the first large-scale analysis of the effectiveness of the origin-exposing vectors. Results show that the problem is severe: 71.5% of the 17,877 CDN protected websites tested, expose their real IP address through at least one of the evaluated vectors [38].

X. Best practices and Content Management System protection discussion

Based on research from previous chapters we've defined list of best practices for securing CMS. Most important of them are:

- Regularly updating CMS, programming and database support, web server and server itself. If all of that is done regularly it's almost impossible to hack major CMS applications,
- Limiting logging attempts, obscuring admin page and installation itself, enforcing strong passwords, use discreet error messages, regularly check for proper permissions, limit IP and country access if necessary.
- Deleting known IDs and names for administration access (like Admin) and use random IDs for admin accounts,
- Using professional hosting services with security teams specially dedicated to monitor security issues with CMS of a choice,
- Using professional CDN service with option to automatically add new and recommended firewall rules.

Using web firewall and security plugins (like iThemes Security or Wordfence) is not effective in case previously defined important factors are not covered [39]. Experts demonstrated how fully patched systems with all security plugins in place can still be compromised [40]. This could be countered if hosting is moved in cloud environment. Since attackers are using standardized exploit kits and standardized codes massively [41] it's easy to spot breach in cloud environment if cloud hosting service is hosting thousands of WordPress installations for example. If there's communication with development team of CMS application than rules can be quickly added or changed easily. For specific most used web applications attacks are detected daily and rules are added immediately to prevent them [42].

XI. Conclusion

Based on findings of this research we can conclude that servers, web servers and database servers used to host most popular CMS discussed in article are not less secure than those used to host other CMS if proposed alternatives are used. We also concluded that number of discovered CMS vulnerabilities in three major CMS is correlated with their market share (chapter 7.). Analysis of most frequently reported types of attacks showed that attacks on three major CMS are mainly using SQL injection vulnerabilities or CSRF in combination with SQL injection. Analysis of community responses in case of security incidents showed that it's important to choose CMS which has live and interactive community. WordPress development team is good example of such community where automated security updates are created and deployed as soon as threat is discovered. Research also showed that CDN WAF offers high quality protection against main vulnerabilities described in chapter 7. All other recommendations for further securing CMS are summed up in chapter 11. where we concluded that even if CMS is fully updated it can still be compromised when it's not hosted in CDN with cloud based protection like WAF. Current research confirms that hosting CMS without cloud web protection is becoming increasingly vulnerable even if all other measures are taken.

References

[1] J. Ji, J. Kim, Y. Kim, S. Jung, C. Lee, D. Kim, *et al.*, "Design and Implementation of the Korean Style Plug-In using the Wordpress."

[2] S. a. t. i. t. u. o. c. m. s.-. W3Techs. (2016). *Usage Statistics and Market Share of Content Management Systems for Websites*. Available: http://w3techs.com/technologies/overview/content_management/all/

[3] S. K. Patel, V. R. Rathod, and J. B. Prajapati, "Comparative analysis of web security in open source content management system," in *Intelligent Systems and Signal Processing (ISSP), 2013 International Conference on*, 2013, pp. 344-349.

[4] A. Onishi, "Security and Performance," in *Pro WordPress Theme Development*, ed: Springer, 2013, pp. 297-332.

[5] R. Ismailova, "Web site accessibility, usability and security: a survey of government web sites in Kyrgyz Republic," *Universal Access in the Information Society,* pp. 1-8, 2015.

[6] S. Mansfield-Devine, "Taking responsibility for security," *Computer Fraud & Security,* vol. 2015, pp. 15-18, 2015.

[7] J. C. Coelho Martins da Fonseca and M. P. Amorim Vieira, "A Practical Experience on the Impact of Plugins in Web Security," in *Reliable Distributed Systems (SRDS), 2014 IEEE 33rd International Symposium on*, 2014, pp. 21-30.

[8] T. Koskinen, P. Ihantola, and V. Karavirta, "Quality of WordPress plug-ins: an overview of security and user ratings," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, 2012, pp. 834-837.

[9] A. Algarni and Y. Malaiya, "Software vulnerability markets: Discoverers and buyers," *International Journal of Computer, Information Science and Engineering,* vol. 8, pp. 71-81, 2014.

[10] A. Mirdha, A. Jain, and K. Shah, "Comparative analysis of open source content management systems," in *Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on*, 2014, pp. 1-4.

[11] P. J. Costa Nunes, J. Fonseca, and M. Vieira, "phpSAFE: A Security Analysis Tool for OOP

Web Application Plugins," in *Dependable Systems and Networks (DSN), 2015 45th Annual IEEE/IFIP International Conference on*, 2015, pp. 299-306.

[12] T. Jensen, H. Pedersen, M. C. Olesen, and R. R. Hansen, "Thaps: automated vulnerability scanning of php applications," in *Secure IT Systems*, ed: Springer, 2012, pp. 31-46.

[13] P. Orduna, S. Botero Uribe, N. Hock Isaza, E. Sancristobal, M. Emaldi, A. Pesquera Martin, *et al.*, "Generic integration of remote laboratories in learning and content management systems through federation protocols," in *Frontiers in Education Conference, 2013 IEEE*, 2013, pp. 1372-1378.

[14] A. Iglesias, L. Moreno, P. Martínez, and R. Calvo, "Evaluating the accessibility of three open‐source learning content management systems: A comparative study," *Computer Applications in Engineering Education,* vol. 22, pp. 320-328, 2014.

[15] A. Singh, "How to Secure Your WordPress Site From Brute Force Attack," *watermark,* 2016.

[16] M. Pistoia, O. Segal, and O. Tripp, "Detecting security vulnerabilities in web applications," ed: Google Patents, 2015.

[17] A. o. t. U. S. D. o. Commerce. (2.2.2016). *National Vulnerability Database*. Available: https://nvd.nist.gov/

[18] K. Hashizume, D. G. Rosado, E. Fernández-Medina, and E. B. Fernandez, "An analysis of security issues for cloud computing," *Journal of Internet Services and Applications,* vol. 4, pp. 1-13, 2013.

[19] G. K. Chaudhary, "Development Review on Phishing: A Computer Security Threat," 2014.

[20] M. Yağci and M. Ünal, "Designing and implementing an adaptive online examination system," *Procedia-Social and Behavioral Sciences,* vol. 116, pp. 3079-3083, 2014.

[21] E. Weintraub and Y. Cohen, "Continuous Monitoring System Based on Systems' Environment," in *Proceedings of the Conference on Digital Forensics, Security and Law*, 2015, pp. 151-160.

[22] D. Bartholomew, *MariaDB Cookbook*: Packt Publishing Ltd, 2014.

[23] D. Bartholomew, "MariaDB vs. MySQL," *Dostopano,* vol. 7, p. 2014, 2012.

[24] CVE_Details. (2016). *Vulnerability Details : CVE-2015-5734*. Available: http://www.cvedetails.com/cve/CVE-2015-5734/

[25] http://www.cvedetails.com/. (2016, 3.32016). *Wordpress: Vulnerability Statistics*. Available: http://www.cvedetails.com/product/4096/Wordpress-Wordpress.html?vendor_id=2337

[26] V. Bittal and S. Banerjee, "Prevention Guidelines of SQL Injection Database Attacks: An Experimental Analysis," in *Emerging Research in Computing, Information, Communication and Applications*, ed: Springer, 2016, pp. 23-32.

[27] I. Lee, S. Jeong, S. Yeo, and J. Moon, "A novel method for SQL injection attack detection based on removing SQL query attribute values," *Mathematical and Computer Modelling,* vol. 55, pp. 58-68, 2012.

[28] CVE_Details, "Security Vulnerabilities Published In 2015 (SQL Injection)," 2015.

[29] J. Aull, *WordPress SEO Success: Search Engine Optimization for Your WordPress Website Or Blog*: Pearson Education, 2014.

[30] S. Fernandes and A. Vidyasagar, "Digital Marketing and Wordpress," *Indian Journal of Science and Technology,* vol. 8, p. 61, 2015.

[31] E. V. J. Aikas, A. Agarwal, and B. N. Bershad, "Content delivery network," ed: Google Patents, 2015.

[32] T. L. Longren, *Wordpress Multisite Administration*: Packt Publishing Ltd, 2013.

[33] J. Stearns, "Create a Faster WordPress Website & Blog," 2013.

[34] D. Gillman, Y. Lin, B. Maggs, and R. K. Sitaraman, "Protecting Websites from Attack with Secure Delivery Networks," *Computer,* vol. 48, pp. 26-34, 2015.

[35] N. Mosharraf, A. P. Jayasumana, and I. Ray, "A Responsive Defense Mechanism Against DDoS Attacks," in *Foundations and Practice of Security*, ed: Springer, 2014, pp. 347-355.

[36] J. A. Dilley, P. Laghate, J. Summers, and T. Devanneaux, "Cloud based firewall system and service," ed: Google Patents, 2015.

[37] M. B. Prince, L. H. Holloway, S. N. Rao, and I. G. Pye, "Supporting secure sessions in a cloud-based proxy service," ed: Google Patents, 2015.

[38] T. Vissers, T. Van Goethem, W. Joosen, and N. Nikiforakis, "Maneuvering Around Clouds: Bypassing Cloud-based Security Providers," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1530-1541.

[39] A. K. Kyaw, F. Sioquim, and J. Joseph, "Dictionary attack on Wordpress: Security and forensic analysis," in *2015 Second International Conference on Information Security and Cyber Forensics (InfoSec)*, 2015, pp. 158-164.

[40] S. Sreedharan, "Does your WordPress Security Plugin really secure your site?," 2014.

[41] T. Taylor, X. Hu, T. Wang, J. Jang, M. P. Stoecklin, F. Monrose, *et al.*, "Detecting Malicious Exploit Kits using Tree-based Similarity Searches," in *Proceedings of the Sixth ACM on Conference on Data and Application Security and Privacy*, 2016, pp. 255-266.

[42] S. Prandl, M. Lazarescu, and D.-S. Pham, "A Study of Web Application Firewall Solutions," in *Information Systems Security*, ed: Springer, 2015, pp. 501-510.

Can malware analysts be assisted in their work using techniques from machine learning?

Igor Novković, Stjepan Groš
Faculty of Electrical and Computing Engineering
University of Zagreb
Unska bb, 10000 Zagreb, Croatia
E-Mail: {igor.novkovic, stjepan.gros}@fer.hr

***Abstract*—When a malware analyst analyzes some code to determine if it's malicious or not and what it is doing, he has to overcome protections built in by malware writer that tries to make it as hard as possible to get to the main functionality of the malware code. In practice that means that when malware analyst, while stepping through the code in some debugger like OllyDbg, hits call instruction or something similar has to decide if he is going to follow call or skip over it. Obviously, if the call is unimportant the best would be to skip it, but at that point analyst doesn't know if it is important or not. The problem is that creative malware writer can use anti debug techniques in such a way that they are hard to recognize and analyze, they can even come up with a new ways to make malware analysis harder. So, the question is is it possible to write a plugin for a debugger that, based on the current call instruction and data behind it, can suggest malware analyst what to do? In this paper we present a system that we are designing and developing that would allow experiments to be performed to find out the answer to the aforementioned question.**

***Index Terms*—machine learning, malware, malware analyst, anti-debug technique, debugging**

I. Introduction

Malicious software [1] is a burning problem for today's ICT industry. There is an ongoing cat-and-mouse game between malware authors, who are writing more sophisticated code every day and returning to life old malware with new protection methods, and malware analysts who are struggling to analyze malicious code and find a way how to stop it. Sophisticated malware has multiple defense layers which serve to fool anti-virus (anti-malware) software, malware analysis platforms, as well as to make it harder for malware analysts to analyze malware internal workings [2]. Malware analyst has to go through mostly manual process of reverse engineering [3] malware, document malware internal workings, write malware signature (or some other form of malware detection technique) and to see if the damage malware has caused is reversible and how to remove it from the system. Malware, besides the malicious code itself, often contains garbage code, which essentially does nothing and it's sole purpose is to confuse malware analyst, harmless code, which is not important for understanding malware workings and can be safely skipped over during analysis, as well as various anti-debug and anti-disassembly techniques which are meant to make malware impossible to analyze. Malware analysts loses precious time while fighting his way through various anti-debug and anti-disassembly techniques and analyzing code which turns out to have nothing to do with malicious activities. The biggest problems are caused by various anti-debug techniques, which can terminate the program being analyzed and force malware analyst to analyze program from beginning. Number of plugins already exist which detect and even prevent certain anti-debug techniques, but malware authors can evade these plugins if they know how anti-debug code is being detected. To be able to detect new versions of anti-debug techniques these plugins need to be updated. We tried to solve this problem by designing the system which would learn to detect anti-debug code dynamically, and thus to remove the need of updating plugins.

The idea is to use machine learning techniques to try to predict whether given segment of a code contains mechanisms which are meant to make malware analysis harder. In other words, at each CALL instruction system collects designated data and feeds it into classifier. Classifier returns output which marks code segment as interesting (important) or non-interesting (non-important). Code segment is defined as interesting (important) if it contains some form of anti-debug techniques and non-interesting (non-important) if it doesn't contain any anti-debug technique. We use both static data (code) and dynamic data (state of registers, memory content, events) to perform classification. Depending on classification

result, system suggests to analyst either that the segment of the code in a callee function should be investigated or that there is nothing interesting in a callee function and that it should be skipped over.

There are two additional goals that are set forth for the system. The first is that it should monitor behavior of the malware analyst and try to infer from its behavior which code is important and which is not. This is necessary since there isn't much data available for the training, and probably there won't be much more in the foreseeable time. This is the problem since machine learning works the better the more data is available for training and testing. The second requirement is that the prototype system should allow for experimenting since at this point of time it is not clear which features and machine learning algorithms are appropriate for the given task. We decided to name system *Machine Learning Aided Reverse Engineering System*, or short MLARES.

In this paper we present a current state of the prototype implementation of MLARES which we are developing that could help shorten time needed for a analyst to analyze malicious code.

The remainder of this paper is structured as follows. We describe MLARES architecture and present data and heuristics which could be used for making suggestion to malware analyst in Section II. Then, in Section III current state of implementation is given and in the Section IV we discuss future work. Finally, Section V concludes the paper.

A. Related work

Machine learning techniques are usually used to classify whether software is malicious, given set of features. Our work and similiar works use machine learning techniques to classify and detect code properties and code segments rather than to classify whole file.

Anderson et al. use machine learning techniques to classify individual subroutines of a program into the appropriate class of functionality [4]. Categories used by authors in a paper are file I/O, process/thread, network, GUI, registry and exploit. As features authors used instructions and API calls.

Opcode histograms have been use to detect computer architecture and endianess [5].

Optimization algorithms were used for malware deobfuscation [6]. In a work control-flow optimization, peephole optimization, constant propagation, constant folding and dead code removal were used to make binary program easier to raed and analyze.

Dynamic translation and pagefault handler modification methods have been shown to be able to help with problem of software packing [7].

One of the common ways to detect malicious software is to run it inside virtual environment and to see whether it triggers heuristics based detection. System Panorama [8] runs software inside virtual environment and it uses taint analysis to see if software accesses data which it should not access.

II. MLARES Architecture

MLARES consists of two components, front-end and back-end. Front-end is implemented as a plugin for a debugger. Its task is to collect and store data needed for classification and to display result of a classification to analyst. Back-end is further decomposed into data processing module and machine learning module. Data processing module extracts features from the data received from the front-end. It then sends this data to the machine learning module. Machine learning module performs classification upon received data and return classification result to the front-end.

Remainder of this section presents data which we store from a reversing session using front-end, and information which we extract from that data upon which classification is performed.

A. Data collecting system

Data collecting system is a part of the front-end which monitors analyst's behaviour during reversing session and stores designated data needed for classification. This subsection describes data which data collecting system stores during a reversing session and which we thought it could matter for classification purpose. Static data (instruction opcodes) is supplemented by a dynamic data (state of registers, memory content, software events) from which additional features can be extracted. At this point we do not know which data is going to prove out to be useful for classification. Because of that we wanted to store as many information as we can from a reversing session. We expect that the experiments with a system after it is finished will allow us to determine which subset of collected data allows classification to be performed. This in turn will make data collection phase to be more efficient.

Instruction opcodes in a callee function are the base feature used for classification. In addition to the opcodes in a callee function, we decided to store opcodes behind CALL instruction, as well as opcodes in a callee function after its execution. There are two parameters associated

with this phase of collection, collection width and collection depth. *Collection width* tells how many opcodes from the beginning of subprogram (the target of the CALL instructions has to be collected). The *collection depth* tells how deep collection should be performed, i.e. if, among the collection width of instructions there is another CALL instruction, it should be also followed or not.

It is obvious that the control flow is depended on data associated with the program at the time of the execution. Because of that it made sense to collect dynamic data that could provide additional features for the classification process. Some of the data collected marks occurrence of certain software events, such as exceptions, which could indicate presence of anti-debug techniques.

We also wanted to make the system which would learn by tracking malware analyst while analysing malicious code. In other words we wanted to have labeled data in a form (opcodes, dynamic data; important/not important). The labeled data would be used to train classifier. The problem is that we can not expect expert malware analysts to waste time with our system in order to generate data for us. Thus, the idea is to obtain this labeled data by observing malware analyst's behavior in specific moments. More specifically, whenever malware analyst hits the CALL instruction, the front-end takes a note weather the analyst entered it or skipped over it.

Based on the previous analysis, the following is the list of dynamic data we chose to save:

- register values,
- do registers point to the memory location,
- value of memory locations pointed to by registers,
- stack base,
- stack data,
- loaded modules before CALL instruction,
- loaded modules after CALL instruction,
- was CALL instruction followed (i.e. entered by malware analyst or not),
- did exception occur,
- was any breakpoint set,
- was any register modified by user.

B. Machine learning

In this subsection we describe problems we anticipated while planning machine learning module and our solutions to those problems. The main problem is, how to decide whether at some point data (i.e. opcodes behind CALL instruction and additional dynamic data) contains something important that requires malware analysts' attention or not. Going through the data by hand or asking malware analyst to tag the data while analysing malware is both impractical and tedious. We intend to solve this problem by designing heuristics which are used to decide whether code segment is important based on some actions taken by malware analyst once he comes across anti-debug technique. Also, interesting are software events which are used in anti-debug techniques and thus could indicate their presence.

There are some typical actions which malware analyst will take once he comes across anti-debug technique or when he comes across interesting segment of code. For example, he'll modify register values or instructions. One typical event which could indicate presence of anti-debug technique is exception triggering.

So, malware analyst's actions and software events are extracted from the data received from the front-end and used to tag data. At this time it is unknown how any of this heuristics is good, but this is something that we'll know when we start to experiment with the prototype after it is finished.

Following is the list of heuristics with the rationale why they could be useful for the described purpose:

- *User has modified registers*. One of the usual ways to bypass some anti-debug techniques is to change value of the registers with the purpose of skipping security checks, so if registers have been modified by user that is an indication of possible anti-debug tecnique in current segment of code, it could be considered as a stronger heuristic.
- *User has modified code*. Also one of the frequently used means to bypass anti-debug techniques. For example, code which performs anti-debug checks will be overwritten with NOP instruction by malware analyst to disable it. This could be considered as a stronger heuristic.
- *If CALL is skipped over at first but in the new run same CALL is entered*. If a call is skipped over at first run but then entered in a new run it could mean that there is anti-debug technique which crashed program in the first run.
- *Number of times CALL was followed*. It could be expressed as a percentage, and it is perhaps a weaker heuristic. Just because CALL was followed many times it doesn't mean there is something important in that segment of the code, so this heuristic can be used to supplement other heuristics.
- *Breakpoint was set*. Breakpoint is set when malware analyst wants to make sure that program stops at chosen location, it indicates he wants to investigate segment of code after breakpoint which could con-

tain anti-debug technique. Nevertheless, the code from the current point to the breakpoint is probably not important.
- *Exception was triggered*. Some anti-debug techniques use structured exception handling (SEH) [9] to make it harder to debug program, so if exception was raised it could indicate there is anti-debug technique in current segment of code.
- *New module was loaded*. If application contains new module after fragment of code was executed, it means that fragment of code required functionality which wasn't available prior to it and it could indicate something related to anti-debug techniques. This is is also only a weaker heuristic.

First five heuristics try to guess the importance of certain parts of the code based on malware analyst's behaviour, while letter two use events which could be interesting. Some heuristics are not useful on their own but they could be used to supplement other heuristics.

III. Current state of development

MLARES is in early phase of development, and only the data collecting part has been implemented so far. This section contains description of the current phase of implementation and it will mention some of problems we have encountered while implementing it.

Data collecting module is implemented as a plugin for a debugger. Ideally, plugin should be developed for all debuggers currently use by malware analysts, but since this is a prototype system, first debugger supported is OllyDbg [10]. The most important thing that had to be designed and implemented in data collecting module was the way of storing data, i.e. the data format.

Data is stored in blocks. Each block describes code segment which CALL instruction points to. Block is determined by it's start address and with the number of bytes we store from code segment. Ideally each block segment would correspond to the whole function that it is being called, but it is difficult to determine where function ends without heavy processing (RET instruction is not necessarily end point because of anti-debug techniques). So, the code segment contained can correspond to the whole callee function (when function size is exactly equal to the number of bytes we have stored), part of a callee function (when the function size is bigger than the number of bytes we have stored) or whole callee function with other code at it's end (when function size is smaller than number of bytes we have stored).

Certain parts of data collecting module implementation were quite challenging, since process of reverse engineering is very nondeterministic and because MLARES needs to act quickly, which required data collecting module to be implemented as light-weighted as possible. The parts are:

- *Determining if reverser left current code segment*. This is important to know because some of data we want to store requires given code segment to be executed for us to be able to store it. That data includes instructions in code segment after its execution, loaded modules after code segment's execution, was there breakpoint set, did malware analyst modify any register during it's execution and did exception happen. We assume that malware analyst is in current code segment while current address is in range between the segment's start address and end address, which is calculated as start address to which is added assigned number of bytes which we store per block. But that is not completely correct solution since real return to a calling function can happen before our end address or return to a calling function can be behind our end address. Second problem is that malware analyst can be in an other segment which is called from current code segment, but it's address is not in current range (that should still register as investigating current code segment). First problem will partialy be solved with data processing in back-end, where we will regulate number of bytes which will be sent to machine learning module. Second problem is partially solved by remembering a calling block ID, but if malware analyst did not step through the code but rather jumped to the breakpoint there is no way of remembering which function he came from.
- *Determining if reverser modified registers*. With support from the debugger for which plugin is developed this would be easy, but we have had problems implementing this in ollydbg.
- *Catching exceptions and breakpoints which happen at address which does not belong to any block*. It is a common case for a malware analyst to put a breakpoint in the middle of a function where some segment is which he is interested in and wants to avoid stepping every instruction to it. It is also common for exception to happen in a part of a code which we are not stepping through. These events could hold important data and we want to capture them. We solved this by creating a new

block around the address at which this kind of event happens. Additional parameter here which needs to be designed is how many bytes before this address to store and how many bytes after this address to store.

Data format had two important requirements:

- Each debugging session had to be uniquely identified as to allow differentiation between malware analyst's various attempts to analyse program. We solved this by having session ID which is pseudo-randomly generated.
- Each block had to be uniquely identified since block's data is basis for learning. We solved this by having block ID which is generated based on it's contents.

MLARES monitors instructions which malware analyst stops at. Every time malware analyst hits CALL instruction MLARES creates block which contains data given in section II.

Each time new program is run or current program is restarted new session data is created. Session data begins with session header. Session header contains time when debugging session was run and a unique session ID (session signature), based on time debugging session was run, configuration data for that session. Configuration data includes:

- number of opcode bytes which should be stored when CALL instruction is hit (collection width)
- number of calls MLARES should follow recursively when saving instructions (collection depth)
- number of bytes which should be stored from stack and memory.

Each block is described by its block header. Block header contains block starting address (address which CALL instruction points at), callee block ID, belonging session ID, block ID and signature which is a hash function of selected data from block body. Block body contains data mentioned in section II. Also, since some data is generated dynamically and it's size is unknown, fields which describe size of such data are added. List of modules before CALL instruction and after CALL instruction belong to that kind of data, and the number of modules for each field is added in every block.

IV. Future Work

This section describes future development of MLARES.

A. Future implementation

This subsection describes modules left to be implement to make MLARES functional. Modules left to be implemented are data processing module and machine learning module.

Data processing module will receive a file with data formatted as described in Section III. It's job is then to extract features from this raw data which can be used for learning and to feed it to the machine learning module. Currently we are not planning on employing any heavy-weight data processing methods because speed is an important factor. We will use functions which will select from input file which data to forward to machine learning module and we will regulate length of some data (instructions and memory locations). Also, this module will extract information about analyst's behaviour and it will tag blocks as important/unimportant upon heuristics described in Section II.

Machine learning module performs classification upon received data from data processing module. Since this is an experimental system multiple classificators will be supported and it should be seen which one will give the best results. MLARES is going to support both the batch learning and online learning. Online learning is required because we want MLARES to be able to learn in real time while malware analyst is analyzing malware.

After these modules have been implemented connection between front-end and back-end needs to be designed and implemented. MLARES needs to be able to give a feedback fast in real time.

B. Experiments

This subsection describes experimentations to be run after MLARES is complete.

Data for learning is going to be collected by analyzing various malware samples. File itself being analyzed is not important for MLARES, since the purpose of MLARES is to detect anti-debug techniques, which can be found even in goodware software. Part of data can be tainted, both because of malware analyst's mistakes as because of our heuristics not being accurate enough. We intend to solve this problem in data processing module by looking at all blocks with same signature and taking into account heuristics which were applied to the block, block's session id (thorugh which time at which block was recorded can be accessed) and frequency of each label. Heuristics are going to be ranked according to how significant they can be to tag block of data correctly. Between the two blocks with the same signtaure but different label, the label of one with stronger triggered

heuristics is going to overwrite the label of other block. Frequencies of labels combined with time block was recorded will help us guess block's true label. The label that is more frequently recorded later in time will be the one which will overwrite the others.

There will be two phases of experimentations. In the first phase multiple classificators are going to be learned using batch learning. In this phase new heuristics and features are going to be designed if there will be need. From the first phase classificator, features and heuristics with the best result are going to be selected and implemented into a MLARES test implementation. In the second phase MLARES will be tested in real time analysis scenario. Other criteria, beside the correctness of classification will be observed in this phase. This criteria includes response time and capability to learn in the real time.

V. Conclusions

Once finished, MLARES should allow experiments to be run which will show if there is a subset of data which can be extracted from information got from reversing sessions which could be used for teaching classificator to recognize anti-debug techniques dynamically. If experiments show success additional decisions need to be made which would make MLARES ready for use. These decisions include location of data processing module and machine learning module with respect to flexibility, speed and load carrrying, which can be placed locally on every station or can be placed on a server which stations would communicate to. For a former option it needs to be designed how knowledge gained from one station watching one malware analyst analyse malware would be conveyed to other researchers and their stations.

References

[1] D. Goodin, "You're infected – if you want to see your data again, pay us $300 in Bitcoins," Ars Technica, October 2013. [Online]. Available: http://arstechnica.com/security/2013/10/youre-infected-if-you-want-to-see-your-data-again-pay-us-300-in-bitcoins/

[2] M. Sikorski and A. Honig, *Practical Malware Analysis*. no starch press, February 2012.

[3] E. Eilam, *Secrets of Reversee Engineering*. Wiley, April 2005.

[4] B. Anderson, C. Storlie, M. Yates, and A. McPhall, "Automating reverse engineering with machine learning techniques," in *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop*. ACM, 2014, pp. 103–112.

[5] J. Clemens, "Automatic classification of object code using machine learning," *Digital Investigation*, vol. 14, pp. S156–S162, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1742287615000523

[6] B. Spasojevic, "Using optimization algorithms for malware deobfuscation," Master's thesis, FER, Sveuciliste u Zagrebu, 2010.

[7] D. Quist and V. Smith, "Covert debugging circumventing software armoring techniques," 2007. [Online]. Available: https://www.blackhat.com/presentations/bh-usa-07/Quist_and_Valsmith/Whitepaper/bh-usa-07-quist_and_valsmith-WP.pdf

[8] H. Yin, D. Song, M. Egele, C. Kruegel, and E. Kirda, "Panorama:capturing system-wide information flow for malware detection and analysis," in *Proceedings of the 14th ACM conference on Computer and communications security*. ACM, 2007.

[9] M. Pietrek, "A Crash Course on the Depths of Win32™ Structured Exception Handling," Jan 1997. [Online]. Available: https://www.microsoft.com/msj/0197/exception/exception.aspx

[10] O. Yuschuk, "Ollydbg." [Online]. Available: http://www.ollydbg.de/

Performance Evaluation of a Rule-based Access Control Framework

S.A. Afonin*
* Lomonosv Moscow State University, Moscow, Russian Federation
serg@msu.ru

Abstract - Attribute-based access control is a flexible approach to security policy specification in an information system: access permission for particular operation on an object is granted depending not only on user's membership in a role, but on the object attributes as well. As object attributes could be difficult to compute (e.g. complex SQL queries might be involved in the computation), the performance of rule-based access control systems is a serious concern in real life applications. In this paper the evaluation results for a rule-based access control system are presented. This access control system dynamically translates access rules into SQL queries and uses various heuristics in order to minimize overall database workload induced by access control checks. The evaluation was performed using the real workload to an information system with millions of objects and thousands of users divided into a dozen roles.

I. INTRODUCTION

Modern information systems require complex access control policies that are hard to implement in the classical Role-based access control (RBAC) model [1]. For example, in a cloud application access rules of the form "a registered user can modify *his own* objects" are quite natural. Checking such constraint can not be realized in terms of RBAC because it requires a check that the user belongs to the role "registered user" and to access the object's attribute "creator" every time the operation is requested. Similar access rules arise in other domains, including information systems, database systems, GRID and Cloud computing, Linux kernel, social networks, and XML document-security [2]-[7]. In the sequel we will refer to web information systems as an example of applications requiring complex access policies.

In the real-life information systems complex access control policies are frequently implemented at the application logic level, e.g. [3],[8]. Some application development frameworks, such as Django, provide means for separation of application and access control logic. Nevertheless, in contrast to the RBAC model, the access control rules are implemented in the same language as the main application. Utilization of a general purpose programming language provides greater flexibility to rule expressiveness, while losing the support easiness and ability of formal verification of access control policies. Moreover, in a cloud system it could be desirable to allow users define their own security policies. This scenario is not possible in case of hard-coded rules because the application owner will not allow users to modify the code.

Attribute-based access control (ABAC) is an extension of RBAC aimed to establish a formal basis for many programming solutions. In this model, the *rule* depends on application level objects' *attributes*, as well as roles and operation context. The following classes of attribute based control systems are identified [9], [10]:

- Dynamic roles. In this approach, the RBAC model is extended by dynamic computation of the role membership relation. Roles are assigned to user dynamically, depending on user and/or context attributes.
- Attribute-centric. Roles are considered as subject attributes. In contrast to RBAC model this approach does not assume permission-to-role relation.
- Role-centric. Permissions are assigned to roles, like in the RBAC model. Attributes are used to constrain RBAC roles. Constraint rules can only reduce user's permissions.

An ABAC system should access object's attributes regardless of the above-mentioned particular class it belongs to. As attribute evaluation could be costly, e.g. intensive computations or complex database query evaluation might be involved, the performance issues are always a serious concern. Compared to hard-coded access control checks, any general purpose access control framework provides more threats to performance degradation because it can not utilize application-specific heuristics known to the developer.

Attribute-based access control received significant attention during the last decade. A promising line of research was developed in ABACα [2], HGABAC [11], EBAC models [12] and [13]. HGABAC model is based on hierarchies of objects and roles, and uses logical expressions over the alphabet of object attribute names to define access rules. Attribute values evaluation procedure is not specified by that model, so the value may be computed by an arbitrary function. EBAC extends this result by introducing path expressions for attribute specification. Very similar idea was introduced in [13], where a more flexible path expression was used for attributes definitions. In contrast to previous works on ABAC models, both [12] and [13] define specific procedures for attributes evaluation. Roughly speaking, attributes are defined as path expressions over the database Entity-Relation model and attributes values calculation reduces to SQL queries evaluation.

Specification of attributes computation procedure makes static policy analysis possible. Attribute definition language should find a balance between expressive power and evaluation complexity.

In this paper we evaluate the performance of the attribute-based control system introduced in [13][1]. Effectively, this system "compiles" the access control policy, defined in a special language, into a set of parameterized SQL queries. When a permission check request arrives, relevant queries are populated by the request parameters and evaluated on the database. In general, more than one rule should be evaluated in order to determine whether or not the operation is permitted. As both permissive and restrictive rules are supported by the model, the list of rules that should be checked during the request processing could be large. The system chooses the rules checking order using an internal heuristic depending on the complexity of the queries. The main goal of this paper is to check the efficiency of such approach in real settings.

The layout of the paper is as follows. In the next Section we briefly describe the ABAC system used. Section III contains the description of the testing application, its properties, security policy, and workload. Evaluation results are presented in Section IV. Short discussion of the results and directions for future work conclude the paper.

II. Description of the Access Control System

In this work we use the access control system introduced in [13]. It defines access control policy definition language and provides runtime support for permissions checks. The system follows the role-centric approach: objects attributes are used to constraint permissions assigned to the roles.

Policy definition language includes two parts: definitions of application *models* (or entities), and definitions of roles and access control rules. Model definition may be considered as a simplified form of SQL CREATE TABLE clause. It includes the name of a table and attributes, where each attribute described by name, column name. If an attribute represents a foreign key, the name of referenced table is also provided. Example of model definition:

```
(defmodel article
 (table "article")
 (attributes (
  (id f_article_id primary key)
  (title f_article_name)
  (owner f_article_user
         references user))))

(defmodel author
 (table "author")
 (attributes (
  (article f_article_id
           references article)
  (department f_department_id
           references
department))))
```

[1] https://github.com/sergadin/SecurityServer.

TABLE I. Rule definition grammar [13]

RULE	::=	'(rule' {'(' clause ')'} ')'
clause	::=	'comment' string | 'model' string ['object alias' string] | 'access' ('allow' | 'deny') | 'operations (' operation {operation} ')' | 'grantees (' grantee {grantee} ')' | 'strict' ['True' | 'False'] | 'condition (' expr ')'
operation	::=	string
grantee	::=	'(users' user {user} ')' | '(roles' [role | '(' role {selector} ')'] ')'
user	::=	string
selector	::=	('object' | alias) [filter] { '.' attribute [filter] }
alias	::=	string
filter	::=	constant relation (attribute [filter] { '.' attribute [filter] }) | '(' filter ')' | filter ['and' | 'or'] filter
expr	::=	'(' relation constant selector ')' | '(' ['and' | 'or'] expr {expr} ')'
relation	::=	'>' | '<' | = | '<=' | '>=' | 'like' | 'in'

The second part of the language is the rule definitions. Each access control rule is associated with one or more grantee, which is either a role, or a particular user. Roles could be parameterized by elements of the data model. For example, the following statement defines a role that restricted to particular department.

```
(defrole "admin" department)
```

When user-to-role membership is verified, a reference to particular department object should be provided in addition to the user name.

A rule is characterized by:

- rule access type (allow/deny),
- the model this rule is associated with,
- list of operations,
- list of grantees (roles or users), and
- additional constraints specifying conditions that have to be satisfied by the access request.

Additional constraints constitute the core of the attribute-based access control. Each constraint is a logical expression on object's attributes. If the constraint is not satisfied by the request, then the rule is not applicable. The grammar of rule definition language is represented in Table 1. Let us illustrate it by the following example.

```
(rule
 (comment "Delete an article.")
 (access allow) ;; allow/deny
 (model "article" object alias "a")
 (operations ("delete"))
 (grantees
  (roles (admin a.authors.department)
         superuser))
```

```
 (conditions
  (and
   (> a.journal[title like "Inform%"]
       .impact[year=2010] 1.5)
   (= 2011 a.year))))
```

This rule defines conditions when an instance of the Article model can be deleted. It states that members of superuser and admin roles can delete articles, if the following conditions are met. An article should be published in year 2011, in a journal with the title starting with "Inform" having 2010's impact factor grater than 1.5. Moreover, admin role is parameterized by an instance of department model. User's membership in the admin role is checked dynamically using the set of departments associated with authors of the article (this is an illustrative example and the real database scheme might contains additional relations, such as many-to-many relation between authors and organizations). Note that foreign keys may be traversed in both directions: *a.authors* traverses foreign key defined in the *author* model in the "reverse" direction, while *authors.department* follows the foreign key in the usual way. Such syntax usually appears in object-relational mapping frameworks. Square brackets are used for extra filtering constraints.

An access control policy is a set of such rules. Rules are compiled into parameterized SQL queries at the policy loading stage. If a rule contains an attributes-aware condition, then it may be reduced to the evaluation of corresponding SQL query. In the above example of a rule, the query generated from rule's condition clause accepts article identifier as a parameter and could be as follows:

```
SELECT article.id
FROM article
  JOIN journal
    ON journal.id = article.journal_id
  JOIN impact
    ON impact.journal_id = journal.id
WHERE article.year = 2011
  AND journal.title LIKE 'Inform%'
  AND impact.year = 2010
  AND impact.value > 1.5
  AND article.id = ?
```

At runtime, the application checks for specific permission by calling *has_permission* function. Input parameters for check operation are the *subject* (user), the *object*, and the name of the requested *operation* (required permission). Checking for particular permission leads to the evaluation of a number of SQL queries. Logically, permission check is equivalent to the following procedure:

- find all *applicable rules*, i.e. rules that are assigned to specified object with matching constraints, and
- verify the there exists at least one permissive applicable rule and there are no restrictive rules.

This framework assumes that an access policy may be defined at different levels of the administrative structure in a federated manner, so each rule is associated to a node of the administrative units hierarchy. It is possible to overwrite an access rule in a subordinate administrative unit. Such possibility is managed in the style of [8]: a rule has special boolean property "strict" that states whether or not this particular rule can be redefined. The value of this property itself can be lowered from True to False by a subordinate policy writer. Overloaded rules are evaluated on per-object basis: when two or more rules, defined at comparable nodes of administrative hierarchy, are relevant for the request, then the most specific rule takes the precedence.

Access control requests are processed using the following procedure.

1. Compute the list of all rules with matching grantees (user names or group memberships), operation name and model.
2. Sort applicable rules by their usefulness (an integrated measure taking into account checking time and frequency of the positive answer).
3. For every rule selected at step 2 do steps 4-10.
4. If the current rule is permissive (positive) and a permissive rule was matched before, go to step 10.
5. Evaluate constraints defined by the condition clause and measure time statistics.
6. Update rule statistics (time spent for computation and the obtained result).
7. If the rule does not match the request (constraint are not satisfied by the access request parameters), then go to step 10.
8. It the current rule is a denying rule, then return False (access is denied).
9. Mark that an permissive rule was found.
10. Proceed to the next rule.
11. Return True (access is allowed).

This procedure verifies that there exists a matching permissive rule and there are no matching denying rules. The search is linear on the number of rules and search order is defined by the measure of rule "usefulness", i.e. a heuristically evaluated numerical value that increases for "fast" rules that match frequently.

The access control system uses automatically generated queries, collects running time statistics and uses heuristics for the optimization of rules processing order. The overall performance of the system in case of real security policy is the subject of evaluation. The situation is pretty much the same as for object-relational mapping frameworks. Such frameworks increase programmer's productivity by providing means for SQL query construction using higher level abstractions available in the host language, e.g. classes and objects of an object-oriented programming language. At the same time, the performance of generated queries is usually much lower compared to the "hand-written" SQL for the same problems.

III. TESTING ENVIRONMENT

For the testing purposes we use workload data collected for the current research information system

(CRIS) of the Moscow State University (http://istina.msu.ru/). The general statistics of the system is presented in Table 2. This system, like any other CRIS system, contains information about researchers' professional activities. At the top level, the system contains about 40 models, such as article, research project, collection editorial membership, textbook, course, conference committee membership, award etc. Each researcher is allowed to modify his own data. For every organization unit, a number of administrators might be assigned and these administrators are allowed to perform some operations on behalf of users associated with their departments. Large number of system's functions are available to administrator only.

Finally, the system is provided as cloud service and participating organizations are expected to use their own security policies.

We believe that the described CRIS system is suitable for access control framework testing due to the following reasons:

- this is a real dataset with real distribution of records among users;
- total number of objects and size of the database scheme leads to relatively complex queries;
- most of the workload is generated by the registered users working simultaneously;
- the administrative hierarchy contains four levels where the access control rules could be defined: system defaults, university, faculty, and department.

As it is not possible to incorporate the access control system into the target application just for the testing purposes, we create the access control policy and run the collected workload on the real database. Actually, the most important parts in this testing scheme are the access control policy and the dataset, because the number of rules and database content affect system performance most. Removing of normal operations evaluated by the application code should not influences the results. During the evaluation procedure we are measuring number of requests processed per second. This value may be compared with the number of request at application level (see Table 2) for estimation of the percentage of time required for security-related tasks.

The security policy constructed from the hardcoded rules used in the CRIS application contains plenty of rules with the same constraint pattern. Most of the models in the database contain a list of authors as their attribute, and this list determines who can manage instances of these models. The most common rule is the following: the object can be modified by any of its authors, or by an administrative user of any author, at any level of administrative hierarchy (administrator of a faculty is an administrator of all its departments as well). Such a rule was defined for almost every data model. As some records in the system pass human control (in order to generate more reliable organization's statistical reports), the confirmed records should be “locked”. This feature is modeled by the denying rules, that are also defined for many data models.

TABLE II. STATISTICS OF THE CRIS SYSTEM

Parameter	Approximate value
Number of models	40
Total number of objects	1 200 000
Maximum number of objects of a model	400 000
Number of profiles (authors)	250 000
Number of roles	15
Number of organization units	900
Registered users	45 000
Active users (per week)	10 000
Requests per minute	350

Nontrivial rules reflect special operations. For example, every institution could define a formula for numerical evaluation of researcher's results. The usefulness of scientomentrics for administrative decision making is far outside the scope of this paper, but this feature gives an example of a more complex access control rule. Formulas are created by users and might be published for public use. When someone computes researcher's report for a given formula, the following conditions should be satisfied: (1) the requester should either requests for his own report, or be an admin for the report's researcher, and (2) the formula should be either published, or be authored by a user for which the requester is an admin. These conditions, defined as the system defaults, require evaluation of complex queries and could be redefined in some departments. (The default system's behavior is to hide a personal report from anybody but the owner and his admins; some departments might use more openness policy.)For the testing purposes we have choose ten most frequently referenced data models and defined access control policy containing 42 rules that were replicated to the total of 164 rules in order to simulate security policy modification in subordinate departments.

IV. EVALUATION RESULTS

The experiments were conducted using two machines: the access control tool was running on an Intel(R) Core(TM) i5-4590 CPU @ 3.30GHz connected over a 100Mb/s switched network to the Oracle 11g database server running on Intel(R) Xeon(R) CPU E5649 @ 2.53GHz with 16GB of RAM. All requests to the database server were evaluated using single thread. SQL queries were not parsed in advance.

Using a trace file with 20 000 requests, that approximately corresponds to a one hour workload, we found that average performance of the access control system is about 6500 requests per minute (the whole file was processed in 3.1 minutes with about 90% load of CPU at the database server) with average 2.1 SQL queries per access control request. The maximum number of SQL queries required to perform the operation was 15 that were computed in 0.4 seconds. It is clear that these numbers are relevant the particular application only. At the same time the access control policy used in the evaluation procedure

covers real rules and access checks took about 5 percent of the normal application's request processing time.

Relatively small number of SQL queries per request is determined by the fact that most requests correspond to a situation when a user modifies his own items. This checks are quite fast and evaluated first due to access control system's heuristics. One extra query was required almost every time, because one should validate that there are no conditions preventing the edit operation. It is worth noticing that if a policy contains large number of denying rules associated with some model, then all these rules should be checked if the access will be granted.

V. CONCLUSION

In this paper we preset evaluation results for the attribute-based access control system [13]. The goal of this work was to check the performance of this system in real settings and to validate the expressiveness power and flexibility of the access control policy definition language.

The performance results are quite predictable. The overall performance of the access control system is mostly limited by the database server. As for the policy definition language, it lacks constructs for property inheritance. Many rules in the access control policy were manually replicated with simple replacements of attribute names. The situation when similar objects, such as book and article, are stored in different database tables sharing the same access control pattern seems to be a rule, rather than an exception. Internally such rules may be replicated, but manual replication in the policy definition file is a source of frequent errors.

As a result of this work we may conclude that an attribute-based access control system that uses automatically generated SQL queries for access control conditions checks may be useful and efficient in real applications. The rule-based access control is not suitable for defining denying rules on particular object/user pairs. If such conditions will be defined using the rule-based approach it would either lead to a large number of rules checking for equality of two identifiers (subject and object), or to a single rule with a very long logical expression.

The main direction of future work is the extension of the access control definition language [13] by some means for access condition grouping. For example, access rules might be associated to application-level concepts instead of database tables. Such concepts might include common parts of relevant access rules. Integration with the DAC model would be appropriate, if large number of exceptions from the general access rules are expected. It also seems reasonable to develop the gold standard, including the database and access control policy, for ABAC systems performance comparison.

REFERENCES

[1] D. F. Ferraiolo, J. F. Barkley, and D. R. Kuhn, "A role-based access control model and reference implementation within a corporate intranet," ACM Trans. Inf. Syst. Secur., vol. 2, no. 1, pp. 34-64, Feb. 1999.

[2] X. Jin, R. Krishnan, and R. Sandhu, "A unified attribute-based access control model covering DAC, MAC and RBAC," in Data and Applications Security and Privacy XXVI, ser. Lecture Notes in Computer Science, N. Cuppens-Boulahia, F. Cuppens, and J. Garcia-Alfaro, Eds. Springer Berlin Heidelberg, 2012, vol. 7371, pp. 41-55.

[3] J. Fischer, D. Marino, R. Majumdar, and T. Millstein, "Fine-grained access control with object-sensitive roles," in ECOOP 2009 Object- Oriented Programming, ser. Lecture Notes in Computer Science, S. Drossopoulou, Ed. Springer Berlin Heidelberg, 2009, vol. 5653, pp. 173-194.

[4] B. Carminati, E. Ferrari, and A. Perego, "Rule-based access control for social networks," in On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, ser. Lecture Notes in Computer Science, R. Meersman, Z. Tari, and P. Herrero, Eds. Springer Berlin Heidelberg, 2006, vol. 4278, pp. 1734-1744.

[5] B. Lang, I. Foster, F. Siebenlist, R. Ananthakrishnan, and T. Freeman, "A flexible attribute based access control method for grid computing," Journal of Grid Computing, vol. 7, no. 2, pp. 169-180, 2009.

[6] E. Yuan and J. Tong, "Attributed based access control (ABAC) for Web services," in IEEE International Conference on Web Services (ICWS), July 2005, pp. 561-569.

[7] J. Crampton and J. Sellwood, "Path conditions and principal matching: A new approach to access control," in Proceedings of the 19th ACM Symposium on Access Control Models and Technologies, ser. SACMAT' 14. New York, NY, USA: ACM, 2014, pp. 187-198.

[8] G. Motta and S. Furuie, "A contextual role-based access control authorization model for electronic patient record," Information Technology in Biomedicine, IEEE Transactions on, vol. 7, no. 3, pp. 202-207, Sept 2003.

[9] D. R. Kuhn, E. J. Coyne, and T. R. Weil, "Adding attributes to role-based access control," Computer, vol. 43, no. 6, pp. 79-81, 2010.

[10] X. Jin, R. Sandhu, and R. Krishnan, "RABAC: Role-centric attribute-based access control," in Proceedings of the 6th International Conference on Mathematical Methods, Models and Architectures for Computer Network Security: Computer Network Security, ser. MMM-ACNS'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 84-96.

[11] D. Servos and S. L. Osborn, "HGABAC: Towards a formal model of hierarchical attribute-based access control," in Foundations and Practice of Security. Springer, 2014, pp. 187-204.

[12] J. Bogaerts, M. Decat, B. Lagaisse, and W. Joosen, "Entity-based access control: supporting more expressive access control policies," in Proceedings of the 31st Annual Computer Security Applications Conference. ACM, 2015, pp. 291-300.

[13] S. Afonin, "Towards a rule-based access control framework for distributed information systems," in Proceedings of the 38th International convention on information and communication technology, electronics and microelectronics (MIPRO), 2015, pp. 1284-1289.

SOCIAL ENGINEERING TRACK

Going White Hat: Security Check by Hacking Employees Using Social Engineering Techniques

Zrinka Lovrić Švehla, Ivan Sedinić and Luka Pauk
Hrvatski Telekom d.d./ Data and Cyber Security Section, Zagreb, Hrvatska
zrinka.lovric@t.ht.hr; ivan.sedinic@t.ht.hr; luka.pauk@t.ht.hr

Abstract – Information security is built on top of three basic building blocks: people, processes and technology. Even when organization has very refined processes and state of the art technology, people usually remain the weakest link in the security chain. Social engineering is the practice of obtaining confidential information or valuable assets by manipulation of legitimate users or owners. Large telecommunication companies, having large customer base and a broad collection of sensitive data, present a very interesting target for social engineering attacks. In order to identify the risks for the business, penetration testing can be used as a method for determining organizational resilience against malicious attack attempts. Usually these tests are done using technical methods, but non technical methods like social engineering can also be employed. In this paper we present a case study of a penetration test using social engineering techniques targeting a large telecommunication in Croatia. We describe the preparation process, techniques used for testing, results and lessons learned.

I. INTRODUCTION

The need for protection of assets is present in every business. Companies create security policies and other documentation to set the rules that should help in building the level of security awareness and to help protect company's valuable assets. Numerous security processes are being defined to prevent and to minimize the risk related to both internal and external misuse and fraud, as well as to minimize the damage when misuse or fraud occurs. State of the art security solutions are being implemented to assure protection and security of assets long term. If people who work for and in the company are not aware of security threats and importance of company's assets and if they are easily susceptible to manipulation [4], there is no security tool or document that will prevent leakage of information, misuse of assets or fraud. One of the hardest threats to handle is social engineering. The art of manipulation put in practice, with the goal of obtaining confidential information can easily cause serious damage to the company that is vulnerable to social engineering. Social engineering is a blend of science, technology and art [5]. It is not always negative and there is no person in the world that did not use it at least once, in order to get what he or she wants. When put in business perspective, it is among the highest risks for the company, especially if the target is being hacked live – by other person.

In this article we present a short overview of what social engineering is, how penetration testing by using social engineering techniques was performed, what was the outcome of it and lessons learned.

II. SOCIAL ENGINEERING

Social engineering is combination of techniques used by malicious attackers to divulge victims in enclosing confidential information or performing actions to compromise information security, by exploiting the flaws in human logic [2]. It is categorized into two categories: human based and technology based. Three main forms are:

- Phishing - The practice of sending emails appearing to be from reputable sources with the goal of influencing or gaining personal information.
- Vishing: The practice of eliciting information or attempting to influence action via telephone, may include such tools as "phone spoofing"
- Impersonation: The practice of pretexting as another person with the goal of obtaining information or access to a person, company, or computer system. [5]

The behavioral component is very important for success of performing social engineering activities [1].

There are different categories of people who perform social engineering activities: hackers, penetration testers, spies, identity thieves, evil employees, recruiters, sales people, governments, and ordinary people [5].

Common social engineering vector attacks are: impersonation, dumpster diving, shoulder surfing, false trust, and web sites [5]. Each of them has its upsides and downsides and that is the reason why the common case is the combination of those.

In order to achieve their goals, the attackers are using different tools: applications, burner phones, caller ID spoofing, cameras, GPS trackers, recording devices and similar [5].

Victims will be more susceptible to successful social engineering attack if there is psychology driven personality trait like diffusion of responsibility, chance for integration, guilt or trust relationship [3].

Typical social engineering attack consists of four steps:

1. Information gathering
2. Relationship development

3. Exploitation

4. Execution

III. RESEARCH PROBLEM

Being one of the largest telecommunication companies in Croatia, the organization is continuously facing numerous threats to its security. Significant part of threats is targeted towards obtaining customer data. Organization has more than one thousand employees who, based on their job description, need to have access to personal data of customers or employees, via databases, self-care portals, social networks or interactive voice response tools (IVR). These people are trained to be helpful and to put the service to the customer (internal or external, depending on the position) as a first priority. The level of their security awareness is being constantly built through various educations, e-learning sessions, posters and workshops. The data that these employees have access to is a very valuable asset to the company and therefore any misuse could cause serious reputation and financial damage. Since many employees have access to one of the most sensitive category of data that company has, information security department decided to conduct a penetration test that will specifically asses the organizational security posture against social engineering attacks.

IV. TEST PREPARATIONS

Penetration test was planned and managed by department responsible for information security. The organization has established internal procedure for planning and conducting penetration tests, which defines that following needs to be defined in the preparation phase:

- scope of the testing,
- tester's initial level of access,
- testing methodology,
- primary goals of the test,
- next steps.

Most critical part of the preparation phase is defining the scope of the test. Improperly defined scope can cause serious problems, because testers could unexpectedly go "out of scope" and perform potentially damaging tests on production systems and other critical organizational assets. When defining the scope of the technical penetration test, this is usually done by listing systems and their location (IP addresses, URLs, etc.). With social engineering this is much more challenging, because boundaries are not so well defined. In our particular case, the scope was defined by listing contacts (email, phone, social networks) testers can use, which were also publicly available on company's portal.

Defining tester's initial level of access is also important, because it drastically influences testing scenarios. For example, we could give testers the role of the complete outsider (having access to nothing), employee (having access to internal resources), vendor (having limited maintenance access to some resources), etc. In our case, it was decided that testers will have customer's role. This means that they already use some of the company's services and have access to a customer self-care portal.

Testing methodology was, of course, social engineering, with the primary goal of unauthorized access to customer's data.

Before the test begins, it is crucial to have appropriate authorizations. In our case, the authorization was requested from the director of the department in the scope of the test, and from responsible board member. However, the decision still remains if personnel "targeted" with a test should be informed. When testing using technical methods, it is usually advisable to inform the system operators and other personnel, so they can properly monitor the system and report any potential problems. However, one could argue if this approach is better, because it could skew the testing results. When testing using social engineering, this argument is even more convincing. Employees with the knowledge they are being "tested" could became more distrusting and significantly alter their behavior. In our case, it was decided that employees will not be informed, but we believe this isn't necessary the recommended approach and that decision should be done on a case by case basis.

Partner company specialized in providing information security services was hired to conduct the test. Appropriate written authorization with clearly defined scope was issued to them, in order to minimize the chances of misunderstandings and legal issues.

It was decided that no special accounts will be opened for the purpose of the testing. Partner company themselves ensured they have appropriate customer accounts which were belonging to real persons. These persons authorized the testing using their data and it was decided that all test will be done using or targeting their data.

V. METHODOLOGY

Penetration testing was done in the real environment and on the real data.

The goals of the testing were:

- check the possibility of access to data that should not be available without clear verification and authentication of the customer,
- check the possibility of customer's data manipulation or change by avoiding to provide the data needed for verification and authentication,
- based on the penetration testing results, provide the guidelines for improvement of processes.

The testing lasted 6 days with the schedule presented in Table 1. Penetration testing started by gathering information via phone calls, social networks and e-mail. During all four phases (information gathering, building relationship, exploit and execution), different social engineering methods were used, as described in Table 2. These methods were selected in order to assure that

collected information is true and independent. No special tools were used, just the land line, mobile phone with SIM cards from different telecom operators, laptop and Internet services.

Examples of attacks include:

- attempting to reset customer's password without appropriate authorization,
- ordering / buying services without appropriate authorization,
- changing tariffs, changing options, etc.

Usually the tester would try to convince the employee he needs to do this while claiming he doesn't have access to data or resources needed for authorization (lost ID card, lost phone, etc.).

Day 1	➢ Information gathering ➢ Initial (test) calls and calls related to identification of protection process ➢ Emails sending planning ➢ Start of communication with social networks services
Day 2	➢ Phone calls (daily) ➢ Emails
Day 3	➢ Planning of escalations and new calling models, selection of key methods for additional testing ➢ Emails analysis ➢ Phone calls
Day 4	➢ Phone calls (daily and nightly)
Day 5	➢ Phone calls (daily and nightly)
Day 6	➢ Making report

Table 1: Testing schedule

Method	Characteristics
Trust	Creation of trust with the person that tester is speaking to in order to get confidential information
Help	Reverse engineering
Access to easily accessible information	Access to information that seem easily accessible
Knowing internal processes	Access to information pretending to be internal employee Usage of information that point to knowledge about internal processes
Clout	Access to information pretending that higher instance/manager is included in communication

Table 2: Testing methods

VI. RESULTS

Penetration testing resulted in six identified vulnerabilities: one high, three medium and two low risk. Risk methodology was provided by the partner company and it used a qualitative approach. The standard quantitative vulnerability risk scoring (like CVSS, usually requested for technical tests) wasn't used, because it was difficult to apply it to non-technical findings Risk level description is given in Table 3. For each one of findings the recommendation in sense of a countermeasure was given.

Somewhat unexpected finding was that most successful exploits were not based on persuasion techniques, but on employees' mistakes and lack of proper knowledge of the process. In some cases, documented instructions were not clear enough and it was apparent that employees are overwhelmed with information in a relatively complex environment. These problems cannot be addressed with security awareness trainings, since lack of awareness is not the primary cause, but more analytical approach is required with the primary goal of making the employee's work more simple and straightforward.

Risk level	Description
Critical	By exploiting critical risk level vulnerability, privileged access to data or systems can be gained.
High	By exploiting high risk level vulnerability, non privileged access to data or systems can be gained, with customers' privileges, as well as access to customers' sensitive data.
Medium	By exploiting medium risk level vulnerability limited access with customer privileges can be gained or the attacker can access as a customer.
Low	By exploiting low level risk vulnerability, the attacker can obtain the information about systems, about customers or other information that can be useful in future attacks.
Informative	By exploiting informative risk vulnerabilities, no damage can be made and those are just for your information.

Table 3: Risk levels

Operational recommendations (OR), as well as strategic ones (SR) were also given:

- OR1: define and implement unambiguous customer identification process
- OR2: make security checklist for every process: if the customer wants to change something regarding his personal data or subscription to services, the employee has to fulfill the check list – to control if all the needed security checks have been done
- SR1: educate all personnel on security and social engineering
- SR2: monitor the implementation of security check lists from OR2
- SR3: continuous penetration testing on broader scope

In the follow-up meeting, these recommendations were used to create action points with appropriate responsible persons and deadlines. Customer's accounts used for the testing were reversed to original state. Partner company was asked to destroy all confidential data gathered during the test.

VII. CONCLUSION

There is no company that is immune to social engineering. Penetration testing done in one of the biggest telecommunication companies in Croatia proved it. Despite continuous educations and building up security awareness of all employees, especially ones who have access to customer data, the company is vulnerable. People are leaving the company, the new ones are coming. Everyone can have a bad day or just want to be as

efficient as possible in reaching his daily targets or simply the internal process is too complicated. All of this increases chances for successful social engineering attacks. Besides security checkpoints in everyday processes and powerful technology, the continuous education is the key to minimizing the risk related to social engineering.

LITERATURE

[1] K.D.Mitnick and W.L.Simon, The Art of Deception: Controlling the Human Element of Security, Wiley Publishing, 2002

[2] X.R. Luo, R Brody, A. Seazzu, S. Burd Social engineering, „ The Neglected Human Factor for Information Security Management“, in Managing Information Resources and Technology: Emerging Applications and Theories, IGI Global, Chapter 11, pp. 151-157, 2013

[3] T.Peltier, Social Engineering: Concepts and Solutions, Information Systems Security, Vol.15, Issue 5, pp.13-21, 2006

[4] F.Mouton, L.Leenen, M.M.Malan and H.S.Venter, Towards an Ontological Model Defining the Social Engineering Domain, ICT Society, 11th IFIP TC 9 International Conference on Human Choice and Computers, HCC11 2014, Turku, Finland, July 30 – August 1, 2014. Proceedings , Series Vol. 431, pp.266-279, 2014

[5] www.social-engineer.org

Analysis of phishing attacks against students

Jakov Andrić*, Dijana Oreški**, Tonimir Kišasondi***
* NTH, Varaždin, Croatia
** Faculty of Organization and Informatics, Varaždin, Croatia
*** Open Systems and Security Laboratory, Faculty of Organization and Informatics, Varaždin, Croatia
andric.jakov@hotmail.com ; dijana.oreski@foi.hr ; tonimir.kisasondi@foi.hr

Abstract - Aim of this research was to examine familiarity of students in Croatia with threats in form of social engineering and phishing attacks. To obtain our research data, a practical assessment of student's capabilities to identify phishing attacks against them was conducted with the help of a graphical questionnaire that presented real phishing messages that tried to identify what kind of scams or attacks were successful against students and what security topics are needed to identify and protect against phishing attempts. This paper shows the result of our work, some thoughts on phishing research and the identified features that are problematic for students' detection of phishing and social engineering attacks.

1. INTRODUCTION

One of the more pressing issues in computer security are targeted attacks against users with the help of methods that rely on social engineering to have an effect against their targets, where that effect is usually an attempt to collect sensitive information like access data for certain systems [1]. Social engineering attacks exploit the inherit trust that people have in communication, since an average computer user that is not trained to identify social engineering attacks will most likely be a victim to such targeted attacks.

Depending on their level of sophistication, when we talk about phishing attacks we think about mass scale, non-targeted attacks against multiple individuals in one or more organizations. Such attacks are usually low in sophistication, and can be extremely effective if the user is gullible enough to divulge his username/password combination or any other piece of sensitive data. This usually helps attackers in multiple ways: users usually reuse passwords, so by obtaining one set of credentials, there is a high probability that those credentials will work on multiple services, and the second one is, if the user can be tricked into running a malicious file that contains a trojan or any other type of malware on his computer, attackers can easily pivot their attacks inside the organization.

As with all classes of attacks, attackers usually improve their attacks by gaining familiarity and additional data about a target. Spearphishing attacks refer to the class of attacks where attackers collect additional data about a target usually from public sources or social networks to make sure that they can craft a persuasive e-mail that would increase the chances that a user will divulge sensitive data or execute a file that contains malware. If the target of the attack is a executive who has access to classified or sensitive data in the organization or a system administrator who has administrative access to computer systems where the attack is specifically crafted against a high value target, we call this type of attack whaling (as a reference that the attackers are trying to catch a really big fish) [2, 3, 4].

When we try to estimate the damage that phishing attacks cause, there are multiple interesting cases we can refer. One of the most interesting cases is the recent attack against the bitcoin startup and exchange Bitstamp. Bitstamp is a bitcoin exchange, meaning the users can use Bitstamp to convert between the the digital currency Bitcoin and fiat currency (like euros or dollars). In January 2015, Bitstamps operations were suspended as a part of an investigation of a possible intrusion in Bitstamps systems. The attackers used a series of targeted phishing attacks against multiple employees, where a system administrator opened a malicious attachment in a personalized phishing mail, and infected the machine of the system administrator. With the help of this access, the attackers managed to steal 19.000,00 BTC (bitcoin) with an estimated value of five million USD [6].

2. RELATED WORK

Phishing research is not a novel research branch. There are multiple related papers that cover susceptibility of individuals to phishing attacks where the research area focuses on the question of why do people fall for phishing and what kind of people fall for phishing. Other papers focus on development of methods that enable identification of phishing e-mails with varying amounts of success. The third interesting branch is user education, where there are multiple papers that focus on how to educate users to be more resistant to phishing attacks.

Mohebzada, Zarka, Bhojani and Darwish [7] conducted two large scale phishing experiments on 10.000 individuals that were students, alumni, faculty and staff of the American University of Sharjah in Australia. They did not inform their subjects that they will be a subject in their experiment, but they did obtain the authorization from the Faculty's Ethical board to proceed with their research. Their research concentrated on the demographics of the subjects that fell for phishing attacks where they showed

This work has been supported by the Croatian Science Foundation under the project IP-2014-09-3877.

that there is no strong correlation to demographics and their susceptibility to phishing attacks.

Sheng, Holbrook, Kumaraguru, Cranor and Downs [8] tried to link demographics and phishing vulnerability, where they found that female and younger test subjects are more susceptible to phishing attacks. Dodge, Carver and Ferguson conducted similar studies against military cadets [9] where about 80% of cadets failed to identify phishing messages. A similar research was conducted in Indiana University by Jagatic, Johnson, Jakobsson, and Menczer [8] where the researchers tried customizing their attacks by using information from the research subject's profiles on social media, in this research, their success rate was 72%, which showed that targeting a user with additional data can provide a significant advantage for an attacker.

Sheng, Magnien, Kumaraguru, Acquisti, Cranor, Hong and Nunge [9] successfully implemented gamification concepts to phishing education which they described in their paper "Anti-phishing Phil: the design and evaluation of a game that teaches people not to fall for phish" where they created an online game that educates users about phishing. The game focused on identifying URL's that might be malicious, where verification showed that the game had an effect that the users were more cautious when identifying future phishing attacks. This research gives an idea to "gamify" potential phishing attacks to test users, as we did in our paper.

Wang, Herath, Chen, Vishwanath and Rao [10] researched how people process phishing mails' and found out that reduction of the depth of information processing induces recipients to make errors, in which they fall for phishing attacks which is tied to deception indicators that help subjects recognize deceptive measures. By increasing the subjects' knowledge on deception indicators and how their knowledge about scams their attention is heightened against their visceral triggers on the likelihood to respond.

Other papers have explored automated methods to detect phishing messages, like the use of multi-label rules in associative classification to detect potential phishing messages like in [11] or usage of supervised learning algorithms like multilayer perceptrons, decision tree induction or naïve Bayes classification like in [12] or heuristics like in [13] where there are multiple interesting avenues for phishing research like feature extraction, where extracted features can be used in other experiments and tests.

3. RESEARCH LIMITATIONS

Research into targeted attacks against users that falls into the social engineering class of attacks has a number of interesting issues that can influence the results of such studies. The one area is the legal and ethical issues. Any social engineering attack that collects potentially sensitive or private data with the help of a computer system can fall and usually falls under the area of potential violation of cyber crime laws, since a computer system is used to unlawfully obtain sensitive data without the user's consent. If we want to lawfully conduct phishing attacks against users, we need to either obtain the permission and consent from their organizations management or we need to obtain permission from the users themselves. In case of obtaining the permission form the users, the scientific validity of the testing can be skewed since the users know that they will be tested and they need to correctly identify the attack as malicious. This fails the requirement that all experiments with users should be blind to eliminate any bias with testing. If we obtain the permission from the board or management of an organization as in a case of a penetration test, the management or someone in the authorization or procurement chain can intentionally or unintentionally leak the information that a test will be conducted, where again the test will have a bias if the organization wants to present itself as secure.

If we want to conduct a realistic phishing test, we need to collect sensitive data to realistically show that a user was indeed successfully exploited or "trick" the users into opening a malicious attachment. The collection of sensitive data is a security problem, since when the data is collected, we need to notify the user that the data is compromised, which in the case of usernames and passwords usually means that the user or organization will have to rotate that compromised set of credentials, also when we collect data, there is a possibility of unintended data leaks or that the test credential set can be stolen, leading to a security incident. An additional problem is the fear of reprisal the users might have if their management / executives find out that they failed a security test and they are a weak link in the security chain. This makes testing difficult, since the users are influenced and biased before the test.

The problem of identifying successful attacks in phishing tests is way simpler with a "malicious" attachment where the attachment is not real malware, but "testware" where the crafted testware won't infect the system, but just send a identifying ping or request to the researchers server. The problem with this approach is that it doesn't account for host based security measures that might stop the infection.

The third problem is how should the users react to phishing attacks? If the organization doesn't have an implemented information security management system and doesn't train their employees to detect phishing or social engineering attacks, usually their employees will be far more susceptible to such attacks and will become victims of such attacks. If they do identify the phishing mails correctly, usually users will simply ignore the e-mail because they correctly identified the e-mail as malicious or simply ignored or didn't see that e-mail, which doesn't help us with identifying the exact reason why an attack failed or succeeded. Developed organizations have systems that the users can use to flag phishing e-mails or report them to the security department, which is especially useful for spearphishing e-mails and heightens the security of the whole ecosystem.

The fourth problem is the availability of test cases and realistic phishing mails as research corpora. If we want to have relevant, unbiased and realistic data for research, we have to collect data continuously with the help of a honeypot e-mail, and usually that will only get us low

quality mass phishing mails and no e-mails that are used in advanced or targeted attacks. And most of the time, if there are interesting cases, organizations that had breaches with targeted attacks classify the phishing e-mails as sensitive and are reluctant to share that information with third parties, which means valuable and interesting research data is hard to obtain, which means that we need to collect large data sets of phishing e-mails that would enable us to research users behavior against known samples to obtain more information about how users perceive phishing attacks and the effectiveness of such attacks.

One important measure is the gradation of the sophistication the phishing e-mails can have. On the low end of the spectrum, we have the mass attacks that try to net the most gullible of users into divulging their PayPal accounts or any other account that can be used instantly to obtain some kind of benefit for the attacker, where even slightly trained users will know how to avoid such attacks. The connection we see from real life cases shows us that the most interesting result is the point that combines the effort the attackers invested into the preparation of the attack, the elements (hooks / deception elements) they attackers used to create the phishing e-mail and finally, the information if the attackers succeeded or not. The interesting attacks are the ones that are marginally targeted, where an attacker did minimal preparation and obtained even one compromised target.

4. RESEARCH GOALS

The goal of our research was to examine the students' familiarity with threats in the form of phishing attacks conducted via the Internet. In accordance to the reasons we wrote earlier in the earlier chapter, a practical test was conducted to determine the student's ability to recognize phishing attacks and if they know how to protect themselves.

In order to present the main guidelines for protection against phishing attacks we tried to answer the following research questions:

- In what extent are students targeted by phishing attacks?
- Do awareness and habits of students help prevent phishing attacks?
- Based on which elements can we determine the students' knowledge about phishing attacks?
- Is there a statistically significant correlation between the education of students about phishing attacks and their resistance against phishing attacks?
- How well do students identify phishing attacks?
- Is there a significant difference between students in different areas of study with regard to knowledge of how to protect themselves from phishing attacks and knowing the essential characteristics of phishing attacks and social engineering?

The motivation behind this research is to explore the Croatian student population's self assessment in regard to phishing attacks and to assess their capability on a limited data set for purposes of obtaining a baseline for future research.

5. METHODOLOGY

We created an online questionnaire that contained 23 questions of which five are pictures of e-mails where the students could see all relevant data in the picture (sender, subject, addresses, URLs etc.). Four were examples of realistic, high quality phishing attacks that were conducted against test subjects' institution that were collected with the help of incident response efforts and reports by users. One sample was a legit email from a company. Participants were asked to determine if each sample is malicious, legit or if they are not sure.

First sample was an example of phishing email sent by a PayPal imitator. The main indicators of fake email were misspelled words, poor imitation of instructions, link that redirect to the fake site and link to log-in that usually PayPal does not add to content of their emails.

Second sample was a phishing email that looked like it was from IRS (Internal Revenue Service). The main indicators of fake email were that the e-mail was not addressed to the recipient and a form that demanded personal and private user data, and submit button that sent data via an unsecured connection.

Third sample was a legit email about confirmation of a recent transaction from a bank. The main indicators of legit email were the last four digits of account number, email was addressed with name of account holder and legit domain name in the URL.

Fourth sample was a phishing email sent by someone imitating “PayPal Dispute Transaction”. The main indicators of this fake email were that the e-mail was not addressed to the correct recipient, large amount of details that should convince recipient about legitimacy, and a fake URL that is similar to the legit URL, and connection via an unsecured connection.

Fifth sample was an example of a phishing email about account access. There were a few phishing indicators that are hidden and hardly noticeable. Main indicators were wrong usage of casing (uppercase/lowercase) of text in the e-mail text and one fake URL among legit content.

Based on the results of the survey we analyzed the data with the help of correlations and analysis of variance. Pearson correlation coefficient was used, which is sensitive only to the linear relationship between the variables. The correlation was performed in order to show the link between education about phishing attacks and prevention of this same attacks.

All research samples and data can be obtained in raw form by contacting the authors.

6. RESULTS

We collected 342 valid responses and determined the percentage of targeted students (students that identified that they were targeted by phishing attacks) and students who were victims of phishing attacks, where they are sure that they were successfully targeted and exploited.

Our sample consisted of students from the Faculties from seven different scientific fields of study: technical, social, humanities, natural, medical, agricultural and art sciences as shown in table 1 which were part of the University of Zagreb. The distribution of their study years is evenly distributed (see table 2).

Table 1: Distribution of students fields of study

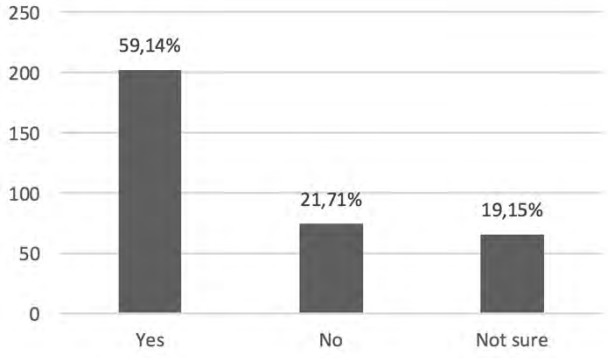
Scientific field	Students
Arts	1 (0,29%)
Technical sciences	124 (36,26%)
Natural sciences	53 (15,50%)
Agricultural sciences	23 (6,73%)
Humanistic sciences	21 (6,14%)
Medical sciences	16 (4,68%)
Social sciences	104 (30,41%)

Table 2: Distribution of students study years

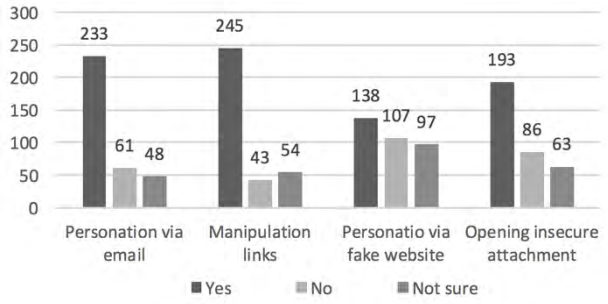
Year of study	Students
First year of study	68 (19,88%)
Second year of study	61 (17,84%)
Third year of study	69 (20,18%)
Fourth year of study	67 (19,59%)
Fifth year of study	69 (20,18%)
Postgraduate study	8 (2,34%)

Most of our test subjects, 59.14% stated that they were targets of phishing attacks while 19.5% are not sure if they were targeted so we can say that more than 59.14% students were targets of some class of phishing attack.

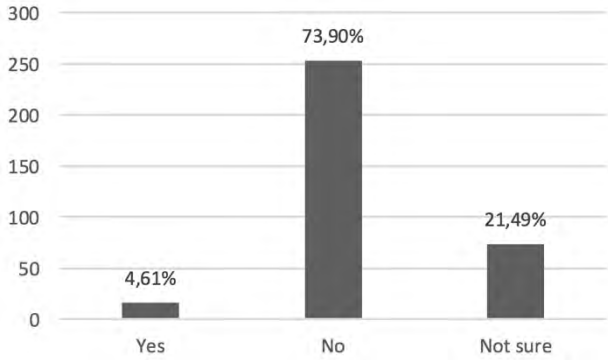
Graph 1 shows the students answers depending on their response if they were targeted, where graph 2 shows the response depending on the attack type. Graph 3 shows the overall proportion of students who were victims of phishing attacks, where students answered that 4.61% of them were victims. Most of students, 73.90% answered that they were not victims of phishing attacks while 21.49% are not sure if they were victim so we can say that more than 4.61% of students were affected by some kind of phishing attack. Graph 4 shows the proportion of students were victim of phishing attack separated by type of attack.



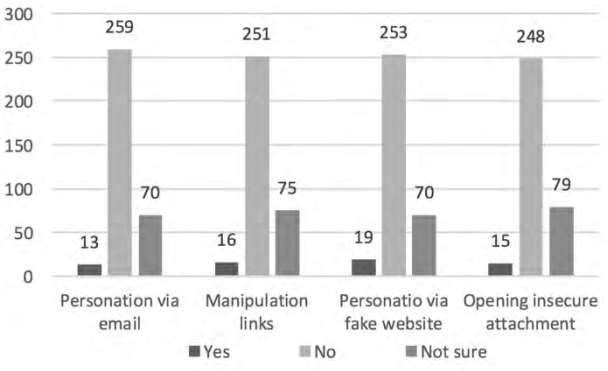
Graph 1: Proportion of students that were targeted by phishing attacks



Graph 2: Proportion of students that were targeted according to attack type



Graph 3: Proportion of students that were victims of phishing attacks



Graph 4: Proportion of students that were victims according to attack type

One interesting piece of data is the success or failure of students to identify correct phishing cases. Table 3 shows the students answers to test samples where we have underlined the wrong answers to those test cases. From table 3 and 4 we can see that the in average failure rate was 27%. The most interesting case is the fifth sample (Phish4), where we see that almost 33% of the students failed to identify the last sample as malicious, which shows that the best strategy in phishing attacks is to create a perfect clone of a legitimate web or mail and just change the core, single element with a malicious one.

Table 3: Students answers to phishing samples

Answer	Phish1	Phish2	Legit	Phish3	Phish4
Phishing attack	195 (57%)	130 (38%)	120 (35%)	127 (37%)	91 (27%)
Legit website	50 (15%)	98 (29%)	81 (24%)	78 (23%)	113 (33%)
Not sure	97 (28%)	114 (33%)	141 (41%)	137 (40%)	138 (40%)

Table 4: Students answers to phishing samples

Answer	One correct answer	Two correct answers	Three correct answers	Four correct answers	Five correct answers
%	79 (23,10%)	88 (25,73%)	65 (19,01%)	31 (9,06%)	10 (2,92%)

70% of students (7 out of 10) with all correct answers stated that they were not targeted by phishing attacks.

We also investigated relationship between self perception of knowing how to protect and number of correct answers given on five questions recognizing phishing.

Table 5: Students self assessment and capability to identify phishing attacks correlated

Answer	Number of correct answers	Number of students
Do you know how to protect yourself from phishing attacks - YES	5	9 (6,98%)
	4	21 (16,28%)
	3	35 (27,13 %)
	2	40 (31,01%)
	1	17 (13,18 %)
	0	7 (5,43 %)
Do you know how to protect yourself from phishing attacks - NO	5	1 (0,47%)
	4	10 (4,69 %)
	3	30 (14,08 %)
	2	48 (22,54 %)
	1	62 (29,11 %)
	0	62 (29,11 %)

Table 5 shows significantly higher percentage of students with no correct answers in the group of students evaluating themselves as the one that "don`t know how to protect". Furthermore, there is "higher percentage of correct answers in the group of students with self evaluation as "we know how to protect". These results demonstrate validity of students self evaluation.

We used Pearson's *r* [16], a correlation to see what variables have strong linear relationship between themselves. Correlation analysis reveals high and statistical significant correlation between students answers and the fact are they target of the attacks or not: students recognizing phishing attacks were not target of the attacks (*r = 0,68, p < 0,002).*

One field we wanted to explore is if education has any effect against phishing attacks, since it acts as a preventive measure. The following correlations between two variables and their Pearson's r value and p value indicate the importance of education in protection against phishing:

- Are you familiar with the term phishing attacks? - Do you know how to protect yourself from phishing attacks? (r = 0,5865 p = <,0001)
- Do you know the difference between http and https protocols? - Do you know how to protect yourself from phishing attacks? (r = 0,5089 p = <,0001)
- Do you know what shortened link is? - Do you know how to protect yourself from phishing attacks? (r = 0,4813 p = <,0001)

The correlation coefficient is positive if students are familiar with the concept of phishing attacks, shortened links and the difference between http and https protocols, which strongly correlates with the variable that shows that they know how to protect themselves from phishing attacks. All three correlation coefficients are between 0,4 and 0,7 which indicates that they are significant. These correlations provide confirmation to the fourth research question: Is there a statistically significant correlation between the education of students about phishing attacks and protect against phishing attacks?

The following correlations indicate importance of awareness and habits that prevents user from becoming a victim of phishing attack:

- Are you familiar with the term phishing attacks? - Do you know how to protect yourself from phishing attacks? (r = 0,5865 p = <,0001)
- Do you know what are shortened links? - How often do you check the URL after the opening new website link? (r = 0,4142 p = <,0001)
- Do you know how to protect yourself from phishing attacks? - How often do you check the URL after the opening new website link? (r = 0,4419, p = <,0001)

Other interesting correlations are shown in table 6. If student has a habit to check URL after opening the website link, he is more likely not to become a victim of phishing attacks. Also, not knowing about URL shorteners and not having habits like checking URLs after opening new website link, leads to students' vulnerability and students are more exposed to attacks. With these correlations we confirm research question: Do awareness and habits of students helps prevent phishing attacks?

Table 6: Interesting correlations

Variable X	Variable Y	r	p - value
Do you know how to protect yourself from phishing attacks?	Are you familiar with the term phishing attacks?	0,5865	<,0001
Are you familiar with the term social engineering?	Are you familiar with the term phishing attacks?	0,5233	<,0001
Do you know the difference between http and https protocols?	Do you know how to protect yourself from phishing attacks?	0,5089	<,0001
Do you know what the shortened links are?	Do you know how to protect yourself from phishing attacks?	0,4813	<,0001
Do you know what the shortened links are?	Are you familiar with the term phishing attacks?	0,4588	<,0001
Do you know the difference between http and https protocols?	Do you know what the shortened links are?	0,4450	<,0001
Do you know the difference between http and https protocols?	Are you familiar with the term phishing attacks?	0,4392	<,0001
Do you know what the shortened links are?	How often do you check the URL after the opening new website link?	-0,4142	<,0001
Do you know how to protect yourself from phishing attacks?	How often do you check the URL after the opening new website link?	-0,4419	<,0001

Analysis of variance [16] was used to indicate whether there is a difference between the students' areas of study and Faculty in understanding the characteristics of phishing attacks. We saw that variability inside groups is bigger than variability between groups and that indicates difference in understanding the characteristics which is shown in table 7. The results of the ANOVA include Sum of Squares, df (degrees of freedom), Mean Square, F (for F-ratio). Critical for interpretation is statistical significance of the results. Thus, table 7 highlights only statistical significant differences.

The question "Do you know how to protect yourself from phishing attacks?" showed promising variable quality between and inside groups.

Between groups:

- SumOfSquares = 11.167,
- df = 6,
- MeanSquare = 1.861
- F = 9.013 and p = 0.000.

Inside groups:

- SumOfSquares = 69.175
- df = 335
- MeanSquare = 0.206
- F = 9.013
- p = 0.000

Table 7: ANOVA for interesting variables (p=0.000)

		Sum of Squares	df	Mean Square	F
Are you familiar with the term phishing attacks?	Between/ Within Groups	27.759/ 168.081	6/ 335	4.626/ 0.502	9.221
How often do you check the URL after the opening new website link?	Between/ Within Groups	49.055/ 597.498	6 /335	8.176/ 1.784	4.584
Do you know how to protect yourself from phishing attacks?	Between/ Within Groups	11.167/ 69.175	6 /335	1.861/ 0.206	9.013
Do you know the difference between http and https protocols?	Between/ Within Groups	26.422/ 56.084	6 /335	4.404/ 0.167	26.304
Are you familiar with the term social engineering?	Between/ Within Groups	10.393/ 117.770	6 /335	1.732/ 0.352	4.927

Table 7 indicates highest differences between students in different scientific fields in the URL checking after the opening new link. Furthermore, the lowest difference within the students in the same scientific fields is in the knowing of the difference between http and https protocols. With this results we can say that there is a statistically significant difference between students in different areas of study, where as expected, computer science and informatics students had the best resistance against phishing attacks. Such trends show the need for information security awareness education for students that study in other areas of study.

7. FUTURE GOALS

Research into phishing and social engineering is a very interesting area since a significant number of attacks are conducted with the help of social engineering and phishing as the main vector to either obtain credentials or trick the user into executing a malware infected file. Our research wasn't designed to deliver significant state of the art results, but it was designed to explore the area of research for

potential future studies. Such a limited pilot study showed interesting challenges with this area of research, mostly because high quality phishing samples that are used against real users is hard to obtain, and 'invented' research data is still biased by the researcher. We see two possible improvements in this field if we collect a significant amount of valid, realistic phishing and spearphishing e-mails for a research corpus. First would be the possibility to identify interesting patterns in phishing e-mails that would help us to train users to detect deception or to develop new methods that would rely on machine learning and pattern recognition to filter out and tag possible phishing mails in a way that would not rely on definitions like antivirus programs.

8. CONCLUSION

Our analysis showed disturbing results: more than 59% of students were targeted by phishing attacks. The answered research questions indicate the crucial need for additional education and guidelines for prevention of phishing attacks for students in Croatia. With given correlations we can see that students without theoretical knowledge about phishing attacks are gullible and susceptible to phishing attacks. Some of the interesting results our research are:

- 21,71% of the students' state that they haven't been targets of phishing attacks, 19.5% of the students' state that they are not sure
- In average 25% of the students failed to correctly identify phishing attacks
- 4,61% of the students say they were a victim of phishing attacks, 21% are not sure.
- The most effective attack was a page that was a perfect clone of a regular, valid mail with only the main action URL rerouting to a phishing domain.
- The single most relevant fact that helped users correctly identify phishing attacks was the familiarity with phishing attacks and if they were educated about those classes of attacks and if they are aware if the correct URL or protocol is used for communication.

REFERENCES:

[1] Stavroulakis, Peter, and Mark Stamp, eds. Handbook of information and communication security. Springer Science & Business Media, 2010.

[2]. Nohlberg, Marcus: Securing Information Assets: Understanding, Measuring and Protecting against Social Engineering Attacks, Stockholm 2008. Accessed: 16.02.2016. Online:http://www.diva-portal.org/smash/get/diva2:200190/FULLTEXT01.pdf

[3]. The Most Popular Social Engineering Lures Used in 2014, Trend Micro, 2015. Accessed: 16.02.2016. Online: http://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/the-most-popular-social-engineering-lures-used-in-2014

[4]. Croatian National Cert: Phishing napadi CCERT-PUBDOC-2005-01-106, Carnet, 2005. Accessed: 16.02.2016. Online: http://www.cert.hr/sites/default/files/CCERT-PUBDOC-2005-01-106.pdf

[5]. The State of Security: 5 Social Engineering Attacks to Watch Out For, David Bisson, Tripwire, 2015. Accessed: 16.02.2016. Online: ttp://www.tripwire.com/state-of-security/security-awareness/5-social-engineering-attacks-to-watch-out-for/

[6] Higgins, Stan: Details of $5 Million Bitstamp Hack Revealed; Online: http://www.coindesk.com/unconfirmed-report-5-million-bitstamp-bitcoin-exchange/ (Available on 18.2.2016)

[7] Mohebzada, J. G., El Zarka, A., BHojani, A. H., Darwish, A. (2012, March). Phishing in a university community: Two large scale phishing experiments. InInnovations in Information Technology (IIT), 2012 International Conference on (pp. 249-254). IEEE.

[8] Sheng S, Holbrook M, Kumaraguru P, Cranor LF, Downs J. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. InProceedings of the SIGCHI Conference on Human Factors in Computing Systems 2010 Apr 10 (pp. 373-382). ACM.

[9] Dodge RC, Carver C, Ferguson AJ. Phishing for user security awareness. Computers & Security. 2007 Feb 28;26(1):73-80.

[10] Jagatic, T. N., Johnson, N. A., Jakobsson, M., & Menczer, F. (2007). Social phishing. *Communications of the ACM*, *50*(10), 94-100.

[11] Sheng S, Magnien B, Kumaraguru P, Acquisti A, Cranor LF, Hong J, Nunge E. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. InProceedings of the 3rd symposium on Usable privacy and security 2007 Jul 18 (pp. 88-99). ACM.

[12] Wang J, Herath T, Chen R, Vishwanath A, Rao HR. Research Article Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email. Professional Communication, IEEE Transactions on. 2012 Dec;55(4):345-62.

[13] Abdelhamid N. Multi-label rules for phishing classification. Applied Computing and Informatics. 2015 Jan 31;11(1):29-46.

[14] Lakshmi VS, Vijaya MS. Efficient prediction of phishing websites using supervised learning algorithms. Procedia Engineering. 2012 Dec 31;30:798-805.

[15] Rao RS, Ali ST. PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach. Procedia Computer Science. 2015 Dec 31;54:147-56.

[16] Field, Andy P, Jeremy Miles, and Zoë Field. Discovering Statistics Using R. London: Sage, 2012. Print.

What do students do with their assigned default passwords?

L. Bošnjak* and B. Brumen*
* University of Maribor, Faculty of electrical engineering and computer science/Institute of Informatics, Maribor, Slovenia
leon.bosnjak@um.si, bostjan.brumen@um.si

Abstract - Despite being the most widely used method of authentication, passwords still pose a significant threat to an information system's security. This threat is mostly attributed to the human factor, as users tend to select passwords that are easy to remember, but are not resilient to brute force or dictionary attacks. Worse yet, when not prompted to change their passwords on a regular basis the users tend to keep their original passwords, or even default passwords set by the system. These bad practices have been addressed over the decades, with the intention to educate the users on the security risks associated with them. A case study on passwords, used by the students at a Slovenian university to access the online grading system was conducted to examine whether passwords have improved over the course of the years. The results have shown that the vast majority of students continue to use the generated default passwords. Of the rest of the students who have changed their passwords, a large percent use short, simple passwords, consisting mainly of alphabetic or numeric characters. With no specific password policies enforced, user-created passwords remain weak, showing that users are still the Achilles' heel of information security.

I. INTRODUCTION

The fact that password authentication schemes have always been and continue to be one of the more insecure methods of authentication, is by no means a new discovery. Already more than three decades ago, Thompson and Morris had observed some of the users' password creation habits. They found out that over 71% of all the passwords were only up to six characters long, and limited to fairly small search spaces. Furthermore, a dictionary search managed to retrieve about a third of the passwords, bringing the total amount of passwords found to about 86% when taking the overlap into account. [1] Ten years later, a follow-up study concluded passwords remained weak, while cracking techniques are getting progressively more successful, due to the growing online word lists and increasingly faster computers. [2]

That password security has been unable to keep up with the exponential increase in computer processing speed, was later shown in several subsequent studies. In a 1999 survey of military personnel, a large majority (80%) of users said they were using passwords of four to seven characters in length. In regards to password composition, around 80% used alphabetic characters, and less than a percent included special characters in their passwords. [3] While user-created passwords have been getting stronger over time [4][5], the password cracking methods remained a step ahead. More than ten years ago, a program designed to crack passwords using the brute-force method would be capable of discovering a lowercase five-character password in less than two seconds, and a lowercase eight-character password within merely ten hours. [6] In 2012, at the Passwords^12 Conference in Oslo, Jeremi Gosney's team demonstrated extreme password cracking speeds. A cluster of five servers, using Virtual OpenCL's, equipped with 25 AMD Radeon GPUs, communicating at 10 Gbps over the Infiniband switched fabric, and running an optimized implementation of the HashCat password cracking program, they were able to achieve 348 billion tries/sec, when cracking NTLM hashes. Using this setup, it would take only six minutes to crack a 14 character long WinXP password. [7]

On top of that, users continue to be the root of the problem: aside from creating short, simple and easy-to-guess passwords, they tend to re-use them across several secured systems, [8] and do not change them on a regular basis, unless prompted to do so. [9] However, when strict password policies are enforced they often write their passwords down to avoid forgetting them. [3]

Nonetheless, the studies conducted suggest that the increasingly demanding password policies, along with the users' growing awareness of the password security issues, are the main reason for the improved complexity of user-created passwords over time. [10][11] The aim of this study is to examine the passwords used by the students to access the university's online grading system, in order to observe their password habits when no specific password policies are enforced. We were interested in finding out how many students changed their default passwords on their own. We observed the characteristics of the user-created passwords, namely their length and composition. Finally, we compared them to the password characteristics observed in the core study conducted by Thomson and Morris, to determine how much have the passwords improved over the course of the last 37 years.

II. RELATED WORK

Past research on the security of passwords created by the students has already been conducted. In some cases, they were asked to create passwords specifically for the purpose of the study, [12] while the other studies focused on students' password habits, associated with their existing user accounts. [13] In the former case, the

passwords were not used outside the study itself, which means that some users might have created arbitrary passwords that do not fully reflect their actual password creation habits. In the latter case, the users' perception of the account's importance varies, meaning that a particular user account (e.g. e-mail client) can be perceived as more or less important, depending on the user. A secondary e-mail account, intended for inessential and spam messages, will likely not be valued at the same level of importance as the user's primary e-mail account; likewise, an e-mail account could be more important to a user than their social media account, but the security importance priorities may differ for another user entirely. [13]

Mazurek et. al studied the characteristics of more than 25,000 students' and faculty members' passwords, used to access various online services within the university. The passwords were created with the intention to protect accounts of high value, and were used frequently over a longer period of time. [14] However, since strict password creation policies were enforced, the users were obliged to meet specific criteria when designing their passwords. The overall strength of the passwords thus mainly depended on the information system in place, rather than the security knowledge of the users themselves.

III. DATA COLLECTION

A total of 185,643 textual passwords, used by the students of a Slovenian university to access several online university services (i.e. online grading system, student e-mail, learning material repository, university Wifi network certificate,…) were obtained and analyzed. Upon account creation, students are assigned default passwords that can be changed. The information system in place does not implement any password policies. As such, the users were not prompted to change their passwords, allowing us to observe their attitude towards their assigned passwords.

A. Password data and default password generation

The information system generates default passwords according to a predefined pattern. Initially, computer-generated passwords were consisted of six characters, with the first two being lowercase initials of the student and the last four being digits. With the knowledge of the pattern, the entire search space can be reduced to only $10^4 \times 29^2 = 8{,}140{,}000$ possible combinations. Knowing the student's initials, which is a valid assumption when trying to recover a specific student's credentials, can bring the total number down to merely $10^4 = 10{,}000$. For that reason, the pattern was later updated to eight characters of length, consisting of one uppercase, one lowercase letter and six digits in an arbitrary order: $10^6 \times 29^2 \times 8! = 33{,}909{,}120{,}000{,}000$ possible combinations. Despite the substantial enlargement of the search space, it can be seen from Table 1 that the times required to retrieve a plaintext password from its hash value have improved drastically since ten years ago [6]. As the password cracking speeds differ depending on the hash function used, as well as the computer's processing speed, the approximated times were estimated based on some of the reported cracking speeds on typical, commercially available computers: 3 million tries per second 10 years ago [6], and 8.5 billion tries per second today. [15] It should also be noted that it is often not necessary to traverse the entire search space, as the correct password can be found sooner.

TABLE I. TIME TO CRACK DEFAULT PASSWORDS NOW AND 10 YEARS AGO, DEPENDING ON PASSWORD LENGTH AND PRIOR KNOWLEDGE OF THE PATTERN

Default password	*Possible combinations*	*Time to crack (10 years ago)*	*Time to crack now*
6 characters, pattern & initials	10000	0.003 sec	1.1E-6 sec
6 characters, pattern knowledge	8.140E+6	2.713 sec	9.6E-4 sec
6 characters, **no** pattern knowledge	1.804E+11	16.7 hours	21.1 sec
8 characters, pattern knowledge	3.391E+13	130.8 days	1.1 hours
8 characters, **no** pattern knowledge	1.015E+15	10.7 years	1.4 days
10 characters, **no** pattern knowledge	5.707E+18	60287 years	21.2 years

B. Data acquisition and ethical aspects

All of the plaintext passwords were obtained from the university under a strict security policy and as a result of a several months long negotiating process. A computer unit with restricted physical access was devoted exclusively to storing and processing the data. The unit's web access was physically disabled at all times, and the operating system was protected by a username and password. The password data stored on the device was encrypted; the processing and statistical analysis was performed on temporarily unencrypted data, which was erased after processing.

As we were operating with real, sensitive data, ethical concerns were addressed as well. The obtained data was anonymized by the university security service's personnel, and contained only plaintext passwords. The use of such personal data without prior or written consent of a subject (e.g. student) is allowed for the purpose of research under Article 11 (2), Article 13 (2), and Article 32 (3) of the Data Protection Directive [16].

Establishing a link between a student and their password would require access to their personal information (such as name, surname or student ID), which is kept in university's databases; however, if a malicious party gained access to the database through any means, reverse engineering the obtained passwords would be unnecessary as the records would be accessed directly. Furthermore, we believe other factors assist in reducing a potential adversary's chance of an unauthorized breach of security and/or privacy to a minimum. The majority of passwords belong to students that no longer use them (graduates or drop-outs), and whose accounts are thus no longer active. Finally, the password analysis was performed in a secure, off-line environment (as described above).

IV. DEFAULT PASSWORDS AND PASSWORD POLICIES

All of the students' passwords were tested against both of the default patterns, to determine whether the students have considered and gone through with modifying their assigned default passwords since the creation of the high-value account. The results, summarized in Table 2, showed that an outstanding 83.2% of the students have not changed their passwords and continue to use the simple, alphanumeric passwords that were assigned to them by the

information system. 144,204 of the passwords, or a total of 77.7% of all students' passwords fit the old, 6 character long pattern, whereas the rest (10,255 passwords) conform to the improved, 8 character long pattern.

The results are more than concerning; 10 years ago, it would have taken up to 51,650 years to crack all 154,459 of the default passwords, whereas today, the same task can be completed in less than 18 years on a typical computer. With the Moore's Law still in effect [17], in another 10 years we can expect the computer processing power to be at least 100 times higher than it is today [18], which would take the task of breaking all of the default passwords only about 66.5 days to complete.

In order to ensure the effort to crack a single textual password becomes impractical, it has been established that password length should increase over time. Whereas traversing the entire search space for 10-character long passwords might still be infeasible today, it will no longer be the case in 10 years. Similarly, 12-character passwords might be sufficient to prevent brute-force attacks ten years from now, [18] but as indicated in the example above, passwords of such length will not be a permanent solution.

It should be noted that the extrapolation above is based on the estimates of the typical computer processing speeds and is prone to change depending on the attacker's setup, as well as the actual processing speed at any given time. Likewise, the estimation does not take into account other improvements such as hash computations [2] or quantum computing [19]. Finally, the cracking time can decrease substantially if patterns are used when generating default passwords.

The designers of information systems are therefore encouraged to keep in pace with the advancements made in computing, so as to always stay a step ahead. Should they decide to generate default passwords for their users upon account creation in the first place, these passwords should be drawn from a pool that is far larger than the one several modern computers even a few years from now can traverse in reasonable time (e.g. 16-character long default passwords). Users, who would decide not to change their assigned passwords, would thus remain safe from brute-force and dictionary attacks; but due to possibly frequent typographical errors made during the user's input of the password [6], and especially reduced memorability and the consequent need to write their password down [20][21], a much larger percent of users than the one reported in this paper would likely be inclined to change their assigned default password.

It has been reported in several studies, however, that users tend to select passwords that are easy to remember. The limitations of the human memory; such as the short-term capacity to remember sequences of items that also need to be somewhat meaningful, as well as the humans' predilection towards redundancy, reduce the high entropy required for secure, hard-to-guess passwords. [21] While coping techniques to help with memorability issues have developed over the years (e.g. cognitive [22], mnemonic passwords [23], etc.), little research has been done to find out how much they are actually in use and whether they are even recognized by the wider population. Instead, research done in the field suggests that password security has not changed much in the last three decades. The users continue to make short, simple passwords that consist mainly of alphabetic and/or numeric characters, follow predictable patterns or exist in available dictionaries.

TABLE II. STUDENTS' PASSWORDS BY CHARACTERISTICS, CONSIDERED IN THE STUDY

Passwords	*Count*	*%*
ALL passwords	185,643	100.00
Default passwords	154,459	83.20
Old pattern (2 lowercase, 4 digits)	144,204	77.68
New pattern (1 uppercase, 1 lowercase, 6 digits; random order)	10,255	5.52
User-generated passwords	31,184	16.80
By composition		
Alphabetic	11,325	6.10
Numeric	8,600	4.63
Alphanumeric	10,845	5.84
ASCII	414	0.22
By length		
4 or less characters	2,129	1.15
5 characters	3,768	2.03
6 characters	10,582	5.70
7 characters	3,640	1.96
8 characters	5,350	2.88
9 characters	2,255	1.21
10 characters	1,455	0.78
11 characters	776	0.42
12 or more characters	1,229	0.66

To prevent the users from choosing such passwords and to ensure the search space of available passwords is large enough, password policies are often employed. A simple example would be a series of rules that determines the minimum length of a password, types of characters that must be included (such as uppercase and lowercase letters, digits and special characters) and prevents the user from including words that appear in a given dictionary. Well-structured password policies have been shown to increase the security of a system [10], however bad practices and habits can often emerge due to the imposing security constraints. [24] It is thus vital to find the balance between the creative freedom, given to users, and the strict policies that maintain high password entropy.

But what happens, when no specific password policies are enforced? With people continuously being educated about the importance of good password creation, as well as information security being a hot topic in the last few years, it could be expected that the quality of passwords would increase over time.

In the second part, we studied the characteristics of the passwords that the students had changed on their own. We were particularly interested in two aspects, related to the overall password strength: its length in characters, and the

composition (denoting the pools individual characters were drawn from to make up the said password). Both of the observed features influence the size of the search space, which should ideally be large enough to prevent several modern computers from traversing it in any reasonable amount of time.

The distribution of students' passwords by character length shown in Table 2 shows that nearly two thirds of the students who have changed their default passwords, chose passwords of 7 characters or less. Taking Table 1 into consideration, it can be seen that these users reduced the number of possible combinations needed to crack their password by a substantial margin, as opposed to leaving their original default password unchanged.

To put things in perspective, if all 20,119 passwords of 7 characters or less had remained unchanged, it would take around 76 years to crack all of them without knowing the default password's pattern, and up to 2.5 years with the pattern's knowledge. On the contrary, the insufficient password length of the modified passwords would allow for an average computer to retrieve all of the passwords in less than 70 days, *even* if all of the passwords were drawn from the pool of alphanumeric and special characters.

A brief examination of the actual passwords revealed that the majority of them appear to be simple, meaningful words. While a more detailed study would be necessary to inspect the semantic meaning behind each of the chosen passwords, it is believed that the students, who have decided to change their default passwords, did so for the sake of convenience (short, simple passwords that are easy to type), and memorability (non-random, redundant words that appear in dictionaries), rather than to actually increase the security of their accounts.

Most of the students (more than a third) created 6-character passwords, presumably because a large number of the convenient and memorable passwords fell into that category. Another, albeit smaller peak can be seen at 8 characters; possibly taking into account the students that have remembered and considered the minimum password length requirements proposed in recent literature [10][11], and often employed in practice. [25]

Nonetheless, with the gradually increasing processing speeds, passwords of 8 and even 9 characters in length are becoming easier to crack. Consequently, any single one of nearly 90% of all the user-generated passwords could be uncovered in a reasonable amount of time. With more time and computing effort, even individual 10-character passwords could be cracked, bringing the total to more than 93.5% of all the passwords, made by the students.

These estimations are based on the assumption that the password is consisted of alphabetic, numeric and special characters, which typically brings the character pool size to 75. However, as users are inclined to select passwords that are simple to type and easy to remember, the pools their passwords are drawn from, are usually smaller. For the purpose of the study, we established four categories, as seen in Table 2, and classified each of the 31,184 user-generated passwords into one of them.

Then, all of the four categories were further broken down into sixteen mutually-exclusive pools of characters, as seen in Table 3. The pools were identified based on the different types of characters the passwords can contain: the first four of the subcategories represent passwords, drawn from the pools that contain only numeric (0-9), alphabetic (A-Z, a-z) or special (i.e. !#$%?,.) characters, whereas the other twelve pools cover various combinations of them. Furthermore, some of the latter twelve subcategories were further divided (e.g. singlecase letters with digits in front, back or in between), as they give additional insight into the patterns that users might default to when making their passwords.

The subcategories presented in Table 3 are roughly ordered by the size of the character pool, meaning that the smaller pools appear near the top of the table.

The majority of the students' changed passwords are composed of characters from the smaller available pools: they use either only digits, or lowercase characters, or they are made of both digits and singlecase (mostly lowercase) letters. The typical character pool of 75 characters is then reduced to 10, 25 and 35 respectively, greatly diminishing the necessary effort to find each individual character within the password. To illustrate how vastly character pool size affects the cracking speeds, the following case of a single 10-character password should be considered. The example in Table 1 assumed that the password was drawn from the pool of 75 characters, meaning that it would take up to 21.2 years to crack it – which is relatively safe for a single password! However, reducing the pool to only 35, 25 or even just 10 characters in size, would allow for the same 10-character password to be uncovered in less than 4 days, 3 hours or 2 seconds, respectively.

Furthermore, targeting the smaller and most often used pools of characters is not the only way to speed up the attack. By knowing how people tend to choose their

TABLE III. BREAKDOWN OF CHANGED PASSWORDS BY COMPOSITION CATEGORIES AND SUBCATEGORIES (N = 31,184)

Category	*Subcategory*	*%*
Numeric	Digits only	27.6
Alpha	Lowercase only	36.3
Alpha	Uppercase only	0.006
ASCII	Special characters only	0.026
Alphanum	Singlecase letters, 1-2 digits in front	0.93
Alphanum	Singlecase letters, 1-2 digits behind	12.2
Alphanum	Singlecase letters, 1-2 digits in between	2.4
Alphanum	Singlecase letters, 3 or more digits	19.2
ASCII	Digits, special characters	0.038
Alpha	Mixed letters	0
ASCII	Singlecase letters, special characters	0.57
ASCII	Singlecase letters, digits, special characters	0.67
Alphanum	Mixedcase letters, 1-2 digits somewhere	0
Alphanum	Mixedcase letters, 3 or more digits	0.013
ASCII	Mixedcase letters, special characters	0
ASCII	Mixedcase letters, digits, special characters	0.022

passwords and exploiting that knowledge, smart brute-force techniques can focus on specific patterns that are more likely to exist in user-chosen passwords. Students that decided to increase the security of their passwords by adding digits to alphabetic characters might have enlarged the character pool size; but a brute-force attack does not have to go through all of the combinations, if the attacker can predict the most likely positions for specific types of characters within the passwords.

For example, the majority of passwords that contained singlecase letters and one or two digits have them placed after the letters (a further breakdown of singlecase letters with three or more digits would show a similar case). With that knowledge, a smart brute-force attack can omit a great number of possible combinations, simply by assuming which character pools a specific character in the password would be drawn from. In our case of a single 10-character password, we might assume the first eight characters are only singlecase alphabetic, while the last two are numeric. That brings the total amount of possible combinations to $25^8 \times 10^2 = 1{,}525{,}878{,}906{,}250$ which would take about half an hour to crack on a typical computer.

Another interesting observation favors the belief that the users tend to select passwords that are easy to type. Of the 11,323 students that chose alphabetic passwords, only two of their passwords were uppercase alone, while none contained both uppercase and lowercase letters. Similarly, alphanumeric passwords mostly include singlecase letters (which are predominantly lowercase), whereas mixedcase alphanumeric passwords appear in just four cases. Finally, a very small percent (only about 1.32% of all user-generated passwords) include special characters. That means a total of up to 98.6% students that changed their passwords, selected passwords that would not require them to hold down any additional keys, such as Shift or Alt Gr, while inputting them.

V. NO CHANGE IN MORE THAN 35 YEARS?

Already in a 1979 study by Thompson and Morris [1], the authors observed the typical user's password creation habits, and determined them to be poor. Table 4 shows the relative amount of uncovered passwords based on their composition; the study also mentions an additional 15% of passwords that appeared in various dictionaries.

Table 4 suggests, that the passwords have improved to a certain degree over the last three and a half decades; a smaller percent of users tends to select shorter passwords, and they appear to select them from larger character pools. Additionally, only about 8% of passwords created by the students appeared in various dictionaries, as opposed to about a third of all passwords in [1].

Nonetheless, a direct comparison between the users' password habits in 1979 and today is not possible without taking the advances in computing into consideration as well. When conducting the study, Thompson and Morris were limited to the amount of processing power available to them at the time. They established several categories of different password lengths and compositions, each making up a search space of a small enough size that an arbitrary password could be recovered from it in a reasonable time.

TABLE IV. COMPARISON OF PASSWORD LENGTH AND COMPOSITION: 37 YEARS AGO AND TODAY

Password pattern	*Thomson, Morris (1979)*	*Students' passwords (2016)*
Single ASCII character	0.46%	0.50%
Two ASCII characters	2.19%	0.33%
Three ASCII characters	14.11%	1.04%
Four alphanumeric characters	14.50%	4.95%
Five characters, all uppercase/lowercase	21.47%	4.34%
Six characters, all lowercase	18.40%	11.37%

To make a valid comparison, we had to consider the contemporary processing speeds, as they set the current benchmarks for brute-force attacks. Since much larger search spaces can be traversed now as compared to more than thirty years ago, we created several new categories, using the password lengths and composition subcategories established in Table 3. To ensure any individual password can be recovered, every category encompassed a number of combinations that can be processed in less than an hour on a typical computer.

From 31,184 passwords, created by the students, the following fit into the listed categories:

- 2,129 (6.83%) were one to four ASCII characters,
- 3,768 (12.08%) were five ASCII characters,
- 10,582 (33.93%) were six ASCII characters,
- 3,640 (11.67%) were seven ASCII characters,
- 5,266 (16.89%) were eight lowercase alphameric characters,
- 884 (2.83%) were nine lowercase alphabetic or numeric characters,
- 49 (0.15%) were ten to thirteen digits.

An additional 127 (0.4%) lowercase passwords of 10 characters or more were uncovered, using a set of simple dictionaries. A total of 26,445 passwords, or 84.8% of the sample as such fell into one of the above classes. Had the more advanced dictionary attack techniques (such as word combinations, or appending of digits) been used, a larger percent of additional passwords would surely have been discovered. Comparing that to the 86% of passwords that were discovered by Thomson and Morris in [1], we can conclude that password security has not made a necessary shift forward in the last 37 years.

VI. CONLUSION

Considering the frequency of security breaches, some of which were quite prominent in the last few years [26], it is clear that password authentication still poses a serious threat to information security. Despite the continuous education of users about the importance of good password creation practices, not much progress seems to have been made in the last three decades. In a case study, in which no specific policies were enforced to support the students maintaining strong and secure passwords, we have shown that the results were just as concerning as in the study done by Thompson and Morris in 1979 [1].

When not coerced to change their default passwords, a large majority of students (over 83%) continued to use the

assigned passwords, relying on the level of security that was provided by them. Of the rest of the students that changed their passwords on their own, nearly 85% chose passwords that would have been easily discoverable in less than an hour on a typical computer. Only up to 4,739 (about 2.55% of all students) created passwords that could potentially not be uncovered in a reasonable amount of time with the current processing speeds available.

While some studies have shown that enforcing strict authentication policies can be counter-productive [11], we argue that giving the users absolute freedom when managing their passwords does not lead to a sufficient level of security. Despite being aware of the security risks, the users generally do not change their passwords, if they are not required to. The security of their passwords then depends on the strength of default passwords the system generates, giving the users a false feeling of security, if the information system's password generation algorithm does not meet the current necessary security standards, as seen in the study.

The users that change their passwords appear to do so mostly out of convenience: without adequate policies, they continue to make short and simple passwords that are easy to type and remember. In our study, the majority of the students that changed their passwords actually lowered the necessary effort to crack them, as opposed to already insufficiently secure default passwords.

To strike a necessary balance between the high-level security requirements and user creativity, we propose that password policies are always to be implemented. If default passwords are used, they should be drawn from pools that are large enough to prevent modern computers from cracking them in a reasonable amount of time – by taking the fast advancement of computer processing speeds into account, default passwords should provide a high level of security for years to come, as well. It is also strongly advised that no specific patterns are used to make up the default passwords.

Similarly, password policies should always ensure the passwords created by the users meet the minimum requirements necessary to establish a large enough search space. If the users' imagination causes them to exceed the requirements, their passwords should be even stronger; but they are never allowed to lower the strength of their passwords beyond that of their original default ones.

REFERENCES

[1] R. Morris and K. Thompson, "Password security: a case history," Communications of the ACM, vol. 22, pp. 594–597, November 1979.

[2] D. C. Feldmeier and P. R. Karn, "UNIX password security - ten years later," CRYPTO '89 Proceedings of the 9th Annual International Cryptology Conference on Advances in Cryptology, pp. 44-63, 1989.

[3] M. Zviran and W. J. Haga, "Password security: an empirical study," Journal of Management Information Systems, vol. 15, pp. 161–185, 1999.

[4] E. H. Spafford, "Opus: preventing weak password choices," Computers and Security Archive, vol. 11, pp. 273-278, May 1992.

[5] B. Schneider, "MySpace passwords aren't so dumb," Wired, December 2006, [Online] Available: http://www.wired.com/2006/12/myspace-passwords-arent-so-dumb, accessed February 2016

[6] M. Keith, B. Shao, and P. J. Steinbart, "The usability of passphrases for authentication: an empirical field study," International Journal of Human-Computer Studies, vol. 65, pp. 17-28, January 2007.

[7] P. F. Roberts, "Update: new 25 GPU monster devours passwords in seconds," The Security Ledger, December 2012, [Online] Available: http://securityledger.com/new-25-gpu-monster-devours-passwords-in-seconds/, accessed February 2016

[8] A. S. Brown, E. Bracken, S. Zoccoli, and K. Douglas, "Generating and remembering passwords," Applied Cognitive Psychology, vol. 18, pp. 641-651, September 2004.

[9] G. B. Duggan, H. Johnson, B. Grawemeyer, "Rational security: modelling everyday password use," International Journal of Human-Computer Studies, vol. 70, no. 6, pp. 415-431, June 2012.

[10] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor, "Encountering stronger password requirements: user attitudes and behaviors," Symposium on Usable Privacy and Security (SOUPS), pp. 1–20, July 2010.

[11] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman, "Of passwords and people: measuring the effect of password-composition polices," CHI '11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2595-2604, May 2011.

[12] D. S. Carstens, L. C. Malone, and P. McCauley-Bell, "Applying chunking theory in organizational password guidelines," Journal of Information, Information Technology, and Organizations, vol. 2, pp. 97-113, January 2007.

[13] B. Friendman, "A study of South African computer users′ password usage habits and attitude towards password security," Rhodes University, May 2014.

[14] M. L. Mazurek, S. Komanduri, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, P. G. Kelley, R. Shay, and B. Ur, "Measuring password guessability for an entire university," ACM CCS, October 2013.

[15] R. Dickey, "Taking a closer look at cracked Ashley Madison passwords," Avast, September 2015, [Online] Available: http://blog.avast.com/2015/09/07/taking-a-closer-look-at-cracked-ashley-madison-passwords, accessed April 2016

[16] "EUR-Lex," Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995, [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML, accessed: April 2016.

[17] G. E. Moore, "Cramming more components onto integrated circuits," Electronics, vol. 38, pp. 114-117, April 1965.

[18] B. Brumen and V. Taneski, "Moore's curse on textual passwords," MIPRO '15 Proceedings of the 38th International Convention on Information and Communication Technology, Eloctronics and Microelectronics, pp. 1627-1632, 2015.

[19] T. D. Ladd, F. Jelezko, R. Laflamme, Y. Nakamura, C. Monroe, and J. L. O'Brien, "Quantum computers," Nature, vol. 464, pp. 45-53, March 2010.

[20] A. Adams and M. A. Sasse, "Users are not the enemy," Communciations of the ACM, vol. 42, pp. 40-46, December 1999.

[21] J. Yan, A. Blackwell, R. Anderson, and A. Grant, "Password Memorability and Security" Empirical Results," IEEE Security and Privacy, vol. 2, pp. 25-31, September 2004.

[22] M. Zviran and W. J. Haga, "Cognitive passwords: the key to easy access control," Computers & Security, vol. 9, no. 8, pp. 723-736, 1990.

[23] B. F. Barton and M. S. Barton, "User-friendly Password Methods for Computer-mediated Information Systems," Computer Security, vol. 3, no. 3, pp. 186–195, 1984.

[24] D. R. Pilar, A. Jaeger, C. F. A. Gomes, and L. M. Stein, "Passwords usage and human memory limitation: a survey across age and educational background," PLoS One, vol. 7, pp. 1-7, December 2012.

[25] Google Sign Up, [Online] Available: https://accounts.google.com/SignUp, accessed February 2016

[26] A. Keszthelyi, "About passwords," Acta Polytechnica Hungarica, vol. 10, no. 6, pp. 99-118, January 2013

Analysing Real Students' Passwords and Students' Passwords Characteristics Received From a Questionnaire

Viktor Taneski*, Marjan Heričko* and Boštjan Brumen*

* University of Maribor, Faculty of electrical engineering and computer science/Institute of Informatics, Maribor, Slovenia

viktor.taneski@um.si, marjan.hericko@um.si, bostjan.brumen@um.si

Abstract - Measuring strength of passwords is important in order to ensure the security of password-based authentication. Since passwords are still the most widely used method for authentication, there has been a considerable research on passwords and password strength. Yet, studies related to password still lack of access to plaintext passwords that are created under a specific password policy. Our research explores the connection between real students' passwords used for managing students' university account and students' passwords characteristics received through a questionnaire. The objective of this paper is to explore whether the characteristics of passwords received through a questionnaire are in line with real university passwords. We analyze real students' university passwords, using access to plaintext of these passwords, and compare the results to the ones as reported by students of the Faculty of Tourism and Faculty of electrical engineering and computer science, collected through the questionnaire. We find that there is a significant connection between the reported and the directly analyzed university passwords for the Faculty of Tourism, but no significant connection for the Faculty of electrical engineering and computer science. Our results for the Faculty of electrical engineering and computer science suggest that students' answers to our questionnaire regarding password characteristics are not in-line with actual university passwords collected in plaintext.

I. INTRODUCTION

Text-based passwords are the most common and widespread authentication method for granting user access. Even though there are other alternative authentication methods like biometrics, graphical passwords, tokens, and lately two-step verification and Single Sign-On, text-based passwords remain the most widely used method for authenticating users. Text-based passwords are also the most debated authentication mechanism through the years, since researchers have addressed numerous problems related to their usability and security.

Password-related problems like passwords being easy to guess, hard to remember, easily stolen etc. are firstly identified by Morris and Thompson [1]. They found that users of the system chose passwords that are short, simple and contained only lowercase letters and digits, or appeared in various dictionaries. These findings and similar were later confirmed by Zviran and Haga [2]. They concluded that one of the biggest vulnerabilities to a computer system's security is the user.

Facing the identified problems with passwords, many approaches and solutions, for increasing the security and usability of textual passwords, were proposed through the years but none have proven widely acceptable [3]. One major weakness with which we still encounter today is that users and their textual passwords are still considered "the weakest link". Users tend to choose weak passwords and passwords that are easy to guess and can be found in a dictionary [4].

The security and strength of passwords is important in order to ensure the security of password-based authentication. Since passwords are still the most widely used method for authentication, there has been a considerable research on passwords and password strength. Yet, studies related to passwords still lack of access to plaintext passwords that are created under a specific password policy.

This study explores the connection between real students' passwords used for managing students' university account and students' passwords characteristics received through a questionnaire. The objective of this paper is to explore whether the characteristics of passwords received through a questionnaire are in-line with actual university passwords. We used the data collected by means of an online questionnaire in our previous work and actual university passwords. We analyze real students' university passwords, using access to plaintext of these passwords, and compare the results to the answers as reported by students of the Faculty of Tourism and Faculty of electrical engineering and computer science, collected through the questionnaire.

The rest of the paper is organized as follows: In Section II we review and analyze related work, while in Section III we describe the research method. In Section IV we present the results of our study. We discuss our findings in Section V and conclude in Section VI.

II. Related Work

Since the users and their passwords are considered the weakest link in the security of information systems, it is important to assess the quality of the passwords that user choose. The most common method for retrieving and analyzing characteristics of users' passwords is a survey/questionnaire. One of the first studies that used a survey as a method for assessing password quality was the study made by Zviran and Haga [2]. They used a questionnaire as an instrument to investigate in more details the characteristics of user-selected passwords and the relationships between key password characteristics.

After many years of research in the area of passwords and passwords security done by many different researchers [3], the question about the ecological validity of password studies based on a survey/questionnaire arises. Fahl at al. [5] addressed the very same question. In their study they manually compared passwords that users made in a laboratory and online study with actual passwords. They found that passwords collected through a survey have some problems, but they can still provide a reasonable approximation for real passwords.

Another similar study is made by Michelle at al. [6]. The authors of this study investigated the gap, which is a result of the lack of access to plaintext passwords in the research on passwords. They analyzed and compared plaintext passwords of entire university to sets of passwords previously collected and leaked from low-value accounts. They found more similarities between the university passwords and passwords collected for research studies, than between the university passwords and the passwords leaked from low-value accounts.

Our approach can be viewed as a combination of the aforementioned approaches. Our questionnaire is an improved version of the one created by Zviran and Haga in [2]. Our university data contains ten times more records that the one of Michelle at al. Furthermore, we compared our real university data to the passwords characteristics that we collected through the questionnaire, in order to assess the validity of the questionnaire as an instrument for collecting reliable password data.

III. Method

A. The Survey

Part of the data for this study was collected through an online questionnaire. The questionnaire allowed us to determine the characteristics of textual passwords used by individuals to log in to their university account, and also to other accounts, such as Facebook, Twitter, Google, or their own mobile devices and phones. We did not ask the participants about their actual passwords, but rather about the characteristics of their passwords for different accounts. We classified the questions into six categories:

- Account information – which services or devices does the student have a password for (desktop computer, notebook computer, mobile phone, tablet, University account, Facebook account, Twitter account, Google account)
- Password information – the characteristics of the selected passwords (first character, length, how did the user choose the passwords etc.)
- Password recall – how often do users forget their passwords, how often do they change them and how often do they write their passwords down
- Frequency of password use – information about how often do users log in by using the password, and whether they use the password to log in to multiple accounts
- Data importance and data sensitivity – how important and how sensitive are the data to users they are protecting
- Demographic data about the user – gender, year of birth, university, faculty, postal code etc.

The questionnaire is based on the questionnaire that was performed by Zviran and Hagga [2]. It has additional questions and additional answer options about password characteristics for different accounts for different services or devices. This allows us to compare characteristics of different passwords for different accounts. The completion of the questionnaire took between 15 and 20 minutes.

B. Survey Procedure

The survey was conducted in the spring term of 2014 and consisted of two phases. The first phase of the questionnaire was performed among students that did not receive the lessons. After the first phase of the survey, the students attended lectures dedicated specifically to passwords and related issues, like password creation, management and security. The education consisted of topics about the importance of creating strong and secure passwords, how to choose such passwords and how to manage them (frequent changes, no writing down, no reusing one password for more than one different accounts etc.). After the education part followed a two-week period of learning decay, followed by the second phase of the survey.

C. Participants of the Survey

The participants were undergraduate and postgraduate students from the Faculty of Tourism and the Faculty of electrical engineering and computer science at a Slovenian university. Table I presents the number of participants in the first and the second phase of the study. A total of 76 students participated and completed the first phase of the web-based survey (43 students from the Faculty of Tourism and 33 from the Faculty of electrical engineering and computer science). 67 of them also completed the second phase of the study (37 from the Faculty of Tourism and 30 from the Faculty of electrical engineering and computer science). After excluding those participants that did not complete one of the two phases (either the first or the second phase), we ended up with a total of 64 legitimate answers for performing our analysis, as shown in the Table I. The last two columns in Table I represent the number of total participants in the study, after we filtered the results. Ages of the participants range from 21 to 51 years for the Faculty of Tourism

TABLE I. NUMBER OF PARTICIPANTS IN THE STUDY

Faculty	Initial results		Filtered results	
	Phase 1	Phase 2	**Phase 1**	**Phase 2**
Tourism	43	37	34	34
EE&CS[a]	33	30	30	30
Total	76	67	**64**	**64**

a. Electrical engineering and computer science

(M = 28.50; M_d = 24.50; SD = 9.046; a Shapiro-Wilk test of normality indicates a deviation from normality at $p << 0.05$) and from 21 to 26 years for the Faculty of electrical engineering and computer science (M = 22.93: M_d = 23; SD = 1.172; a Shapiro-Wilk test of normality indicates a deviation from normality at p = 0.015). Ten participants were male and 24 were female at the Faculty of Tourism, and 26 were male and 4 were female at the Faculty of electrical engineering and computer science.

D. University Data

We analyzed overall more than 250,000 passwords of the students at a Slovenian university. These are students' university account passwords that students use to access their university account, university email, grades, course information, and other restricted university resources. In addition to students' passwords the university data also contains additional information about the study program, gender, year of birth, the status of the student, year of first enrollment into the program, year of graduating, average grade, place of birth and post number. Please note that the collected data do not contain any information about name or surname, student ID, or any kind of identity number of the students.

The university data contains passwords that were collected for period of more than 20 years. The passwords were stored in plaintext format in the university's database until 2014. Thereafter the system was changed and passwords are no longer stored in the plaintext, but as salted hashes. Since our previous study was performed among students that were enrolled at the university in specific programs in the years 2012, 2013 and 2014, we had to filter the university real data in order to match the specifically selected programs and enrolment years from our previous study. For the purposes of our research, we only used students' passwords that are equivalent to the enrollment years of the students that attended our questionnaire earlier. After filtering the university data we ended up with a total of 84 records for the Faculty of Tourism and a total of 206 records for the Faculty of electrical engineering and computer science.

E. Ethical and legal aspects

The data on students' passwords were obtained from the university under very strict conditions. All data were anonymized by the university's IT personnel. The processing of such data is allowed under Article 11, Paragraph 2, Article 13, Paragraph 2 and Article 32, Paragraph 3 of the Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [7]. These paragraphs allow for using personal data for the sole purpose of research without prior and/or written consent of a data subject (i.e. person).

Ethical aspects were considered as well. The question arose whether it would be possible to reverse engineer a password in such a way that the link to a person's identity could have been established and hence such a password would enable an unauthorized breach of security and/or privacy. Such a possibility always exists, yet the probability of such an event is, in our opinion, very low due to several factors. Firstly, the passwords were anonymized at the source: we only obtained records consisting of a plaintext password and other related data, as described in subsection D. Reverse engineering would require an access to student's personal records, which are kept in the university databases. But, if the access to database of personal records is possible, there is then no need for reverse engineering as personal records are available by means of direct access. Secondly, most of password records belong to students which are no longer active (graduates or drop-outs), meaning their accounts are locked out and hence inaccessible. Thirdly, passwords used in the old information system were not transferred to the new system; each user was assigned a new identity and default (very different) password. Fourth, no emails or other similar identification fields were kept in the old system. This way an adversary cannot link an existing password to a person even if the person is reusing the old password in other systems. Fifth, the passwords are analyzed in a secured off-line environment (see below) Thus, in our opinion, the possibilities of a breach are reduced to an acceptable minimum.

F. Data acquisition and security measures

The acquiring of the real university data was a result of a several months long negotiating process between the security services of the Computer Center of the university, the management of the Faculty of electrical engineering and computer science, and the researchers interested in using the university data for research purposes only. We acquired the university data due to fortunate events, which may not be reproducible in the future.

We were obligated to manage with the university passwords and related data in a secure and predefined manner. The university data were stored on a special computer unit, which was devoted exclusively to the processing of these data, and on which was installed the necessary software for statistical analysis of the data. The computer unit was not connected to a network and the Web access was physically disabled. The unit was installed in a room with restricted physical access (locked). Physical access to this unit was limited by the granting of key followed by a signature. Furthermore, the access to the operating system and BIOS of the unit was secured with a user name and password. The data were stored on the device in encrypted form and the processing was performed on temporary unencrypted data, which was deleted after processing. Every transfer of the

TABLE II. MOST COMMON PASSWORD CHARACTERISTICS – FACULTY OF TOURISM

Password characteristics	Count
Questionnaire	
(1) numeric – digits only	**4 (14.8%)**
(2) single case letters and 1-2 digits in front	**4 (14.8%)**
(4) alphabetic - lowercase letters only	1 (3.7%)
(6) alphanumeric - single case letters and 1-2 digits in between	**3 (11.1%)**
(7) alphanumeric - single case letters and 3 or more digits	1 (3.7%)
(8) alphabetic - several mixed lower and uppercase	1 (3.7%)
(9) alphabetic - one letter uppercase, other lowercase	1 (3.7%)
(12) alphanumeric - mixed case letters and 1-2 digits somewhere	**3 (11.1%)**
(13) alphanumeric - mixed case letters and 3 or more digits	2 (7.4%)
(17) mixed case letters, several special characters and several digits	1 (3.7%)
(19) long sentence, uppercase only	1 (3.7%)
(24) I prefer not to disclose	**5 (18.5%)**
Real university data	
(1) numeric – digits only	31 (36.9%)
(13) alphanumeric - mixed case letters and 3 or more digits	**53 (63.1%)**

university data (only the results of the treatments) took place via an encrypted USB transmission medium.

IV. RESULTS

We retrieved the university data almost a year after we performed the questionnaire among the students. In order to have properly organized data and a valid study, we analyzed the results from the second phase of the questionnaire and compared them to the relevant (filtered) university data. We were interested in comparing the general password characteristics of the university passwords of both the Faculty of Tourism and the Faculty of electrical engineering and computer science to the ones of the university real data.

In our previous study [8], we asked our participants about their password characteristics and did not inquire their actual passwords. In order to properly compare the passwords characteristics from our questionnaire to the ones of the university data, we had to categorize the passwords characteristics into certain categories. In our previous work we categorized the password characteristics on an ordinal scale, from 1 = "digits only" to 23 = "long sentence, digits and special characters" and 24 = "I prefer not to disclose". We ordered this scale by calculating the number of possible password combinations for every password category. For the purposes of our study, we presumed that the average password length is approximately 9 characters, as stated in [9], a long sentence would have at least four words, and we have the English language and the English alphabet as a dataset (e.g. for a "one letter uppercase, other lowercase" type of password the number of possible combinations would be $26^1 \times 26^8 \times 9$).

TABLE III. MOST COMMON PASSWORD CHARACTERISTICS – FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Password characteristics	Count
Questionnaire	
(2) single case letters and 1-2 digits in front	**5 (17.9%)**
(3) alphanumeric - single case letters and 1-2 digits behind	**5 (17.9%)**
(6) alphanumeric - single case letters and 1-2 digits in between	1 (3.6%)
(7) alphanumeric - single case letters and 3 or more digits	**5 (17.9%)**
(12) alphanumeric - mixed case letters and 1-2 digits somewhere	**5 (17.9%)**
(13) alphanumeric - mixed case letters and 3 or more digits	1 (3.6%)
(15) ASCII - mixed case letters and one special character	2 (7.1%)
(16) ASCII - mixed case letters and several special characters	1 (3.6%)
(17) mixed case letters, several special characters and several digits	1 (3.6%)
(24) I prefer not to disclose	2 (7.1%)
Real university data	
(3) alphanumeric - single case letters and 1-2 digits behind	10 (4.9%)
(4) alphabetic - lowercase letters only	3 (1.5%)
(6) alphanumeric - single case letters and 1-2 digits in between	3 (1.5%)
(7) alphanumeric - single case letters and 3 or more digits	**190 (92.2%)**

We used the same categorization also for the university data in order to arrange the passwords into specific categories. We classified the real university passwords into the same 24 categories. When analyzing the passwords and comparing the results with the results from the questionnaire we excluded the category 24, which is "I prefer not to disclose".

The most common identified categories of password characteristics for the Faculty of Tourism were 1 = "numeric - digits only" and 13 = "alphanumeric - mixed case letters and 3 or more digits". Table II shows the comparison of the most common password characteristics from the questionnaire from our previous study and the real university data. The table contains only the categories that have count more than 0. The rest of the categories were not included in the table due to lack of space.

We find that 63.1% of the real university passwords of the Faculty of Tourism fall into the category of "alphanumeric – mixed case letters and 3 or more digits" and 36.9% fall into the category of "numeric – digits only", while the answers from the questionnaire mostly fall into the categories "numeric – digits only" (14.8%) and "single case letters and 1-2 digits in front" (14.8%).

We compared the sets of counts for different categories i.e. password characteristics for the university account from the questionnaire to the real university data for the Faculty of Tourism. If we take a look at Table II we can observe some differences that may indicate that there is no significant connection between the two variables. First, we can notice that there are more categories with real university passwords which count is 0, than the categories with the answers from the questionnaire. Second, the real university passwords are categorized in only two categories, while the answers from the questionnaire vary across more different categories. As it turns out, these differences are not significant. Fisher's Exact test showed significant dependency between the two variables ($p = 0.000223$, the null hypothesis being that the two variables are independent).

The most common identified categories of password characteristics for the Faculty of electrical engineering and computer science were 3 = "alphanumeric - single case letters and 1-2 digits behind", 4 = "alphabetic - lowercase letters only", 6 = "alphanumeric - single case letters and 1-2 digits in between", and 7 = "alphanumeric - single case letters and 3 or more digits". Table III shows the comparison of the most common password characteristics from the questionnaire from our previous study and the real university data. Similarly as with table II, due to lack of space the table contains only the categories that have count more than 0.

We find that 92.2% of the real university passwords of the Faculty of electrical engineering and computer science fall into the category of "alphanumeric - single case letters and 3 or more digits", 4.9% fall into the category of "alphanumeric - single case letters and 1-2 digits behind", 1.5% into the category "alphabetic - lowercase letters only", and 1.5% into the category "alphanumeric - single case letters and 1-2 digits in between". On the other hand, the answers from the questionnaire mostly fall into the categories "alphanumeric - single case letters and 3 or more digits" (17.9%) and into the category "alphanumeric - mixed case letters and 1-2 digits somewhere" (17.9%). Compared to the real university passwords of the Faculty of Tourism, we can observe that passwords of the Faculty of Tourism are slightly stronger (according to our ordinal scale) than the ones of the Faculty of electrical engineering and computer science.

Comparing the sets of counts for the categories i.e. password characteristics for the university account from the questionnaire to the real university data for the Faculty of electrical engineering and computer science, we find no significant dependency between the two ($p = 0.752381$, the null hypothesis being that the two variables are independent).

V. DISCUSSION

Our analysis shows different behavior among the students from the Faculty of Tourism, as opposed to students from the Faculty of electrical engineering and computer science. We can notice that there are some answers from the Faculty of Tourism that deviate from the real university passwords, despite the fact that the dependency between the real university passwords and the answers from the questionnaire is significant. Our deduction is similar to the one of Fahl at al. - some students did not behave realistically during our questionnaire-based password study and did not provide answers about their real university passwords.

The results of the comparison of the real university passwords to the answers from the questionnaire for the Faculty of electrical engineering and computer science tell a different story. The answers from the questionnaire about the password characteristics significantly deviate from the real university passwords. These results add weight to our assumption in our previous study where we argued that it is most likely that after attending the lectures the awareness of users about the importance of their data and the importance of not disclosing information about the passwords for their accounts, increased. This could lead to participants in the second phase deliberately falsifying data about their password characteristics.

While our results are statistically significant for our study, this should not be generalized without care, since our data was drawn from the student population of a single university and our study is focused on this population.

VI. CONCLUSION

In this study we explored the connection between real students' passwords used for managing students' university account and students' passwords characteristics received through a questionnaire. We analyzed the plain-text university passwords and explored whether the characteristics of passwords received through a questionnaire are in-line with real university passwords.

We used the data collected by means of an online questionnaire in our previous work, and actual passwords. We analyzed real students' university passwords, using access to plaintext of these passwords, and compare the results to the ones as reported by students of the Faculty of Tourism and Faculty of electrical engineering and computer science, collected through the questionnaire

We find significant dependency between the real university passwords and the answers from the questionnaire for the Faculty of Tourism. The students' answers to our questionnaire regarding password characteristics are in-line with the real university passwords. But, despite this significance, there are still some students that did not behave realistically during our questionnaire-based password study and did not provide answers about their real university passwords. On the other hand, we find that there is a gap between the reported and the directly analyzed university passwords for the Faculty of electrical engineering and computer

science. There is no significant dependency between the real university passwords and the answers from the questionnaire. Our results suggest that students' answers to our questionnaire regarding password characteristics are not in-line with actual university passwords collected in plaintext. Furthermore we find that students at the Faculty of Tourism generally create stronger passwords than students at the Faculty of electrical engineering and computer science.

We argue that even though some of our results show significant dependency between answers from our questionnaire and the real university passwords, our results should not be generalized without caution. One should be careful when using a questionnaire as a tool for analyzing characteristics of users' passwords, assessing password strength or password behavior.

While some of our results suggest that students' answers to our questionnaire regarding password characteristics are in-line with actual university passwords collected in plaintext, there is still a slight possibility that this is due to the university password policy itself. This possibility is still to be further researched in our future work.

REFERENCES

[1] R. Morris and K. Thompson, "Password Security: A Case History," *Commun. ACM*, vol. 22, no. 11, pp. 594–597, Nov. 1979.

[2] M. Zviran and W. J. Haga, "Password Security: An Empirical Study," *J. Manag. Inf. Syst.*, vol. 15, no. 4, pp. 161–185, 1999.

[3] V. Taneski, M. Hericko, and B. Brumen, "Password security - No change in 35 years?," in *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on*, 2014, pp. 1360–1365.

[4] S. Egelman, A. Sotirakopoulos, I. Muslukhov, K. Beznosov, and C. Herley, "Does My Password Go Up to Eleven?: The Impact of Password Meters on Password Selection," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2379–2388.

[5] S. Fahl, M. Harbach, Y. Acar, and M. Smith, "On the ecological validity of a password study," in *Proceedings of the Ninth Symposium on Usable Privacy and Security - SOUPS '13*, 2013, p. 1.

[6] M. L. Mazurek, S. Komanduri, T. Vidas, L. Bauer, N. Christin, L. F. Cranor, P. G. Kelley, R. Shay, and B. Ur, "Measuring password guessability for an entire university," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13*, 2013, pp. 173–186.

[7] "EUR-Lex," *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data*, 1995. [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML. [Accessed: 08-Apr-2016].

[8] V. Taneski, M. Heričko, and B. Brumen, "Impact of security education on password change," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings*, 2015, pp. 1350–1355.

[9] A. Adams, M. A. Sasse, and P. Lunt, "Making Passwords Secure and Usable," in *Proceedings of HCI on People and Computers XII*, 1997, pp. 1–19.

MISC TRACK

Using DEMF in Process of Collecting Volatile Digital Evidence

Miroslav Bača*, Jasmin Ćosić **, Petra Grd*
* Faculty of organization and informatics, Center for Forensics, Biometrics and Privacy, Varaždin, Croatia
** Independent researcher, Bihać, Bosnia and Herzegovina
miroslav.baca@foi.hr, jasmin.cosic.ba@gmail.com, petra.grd@foi.hr

Abstract – Acquisition of volatile data for further forensic analysis still represents a challenge to both practitioners and researchers. The current tools used for acquisition of such data are focused exclusively on a way to capture content. However, the development of forensic science, in particular in the area of digital evidence in terms of the admissibility in court, has introduced additional elements to be evaluated. Mainly, the integrity of the collected digital evidence, authenticity and other elements of the digital chain of evidence to be presented in court. This paper describes a framework for capturing volatile data using Digital Evidence Management Framework (DEMF) with regards to integrity of captured data.

I. INTRODUCTION

The problems researchers and practitioners face in the context of respecting digital chain of evidence[1] for the acquisition of volatile digital evidence lies in the fact that with existing tools it is not possible to present evidence in court and confirm their integrity and completeness. It is relatively easy, and so far it is accepted by the courts, to confirm the integrity of the file as well as the media using hash functions. But even "standard" static data analysis leads to different results from the same values, which can be seen in the analysis conducted by Zimmerman [1]. Exploring a variety of tools for media acquisition as well as different methods these tools use, results were quite different. For example, using the X-Ways Imager tool using DD format, obtained size was 1,000,204,886,016 bytes, whereas using the E01 method size was 1,000,449,415,330 bytes. The difference between formats can also be observed by using other tools like FTK where DD image was 1,000,204,886,016 bytes, while E01 image was 1,000,571,704,661 bytes. In instances where EnCase Imager tool was used for image format E01 value obtained was 1,000,571,589,032 bytes. Although it was the same media, the results indicate that different tools in the same format have very different values of the media image size. As the size of the media is directly related to the hash function calculation, conclusion is that neither the hash function for the same medium with different tools is the same.

Even greater difference occurs with the acquisition of volatile data, such as data collected from RAM or data collected from computer networks. Other that it is unknown how large the data is initially, it is rather difficult to calculate the hash function and confirm the integrity of the data. To solve this problem to some extent, researchers have begun to develop algorithms that would be able to calculate the hash function of data whose length is not known. Such functions usually take parts of the same size and sign them with hash functions.

The question posed here is whether it is possible, only on the basis of such data, to guarantee the integrity of the data.

II. DIGITAL EVIDENCE MANAGEMENT FRAMEWORK

In order to solve the problem described in the introduction, some solutions are suggested to deal with the digital chain of evidence of which the DEMF is one of the last and most comprehensive in terms of confirming the integrity of the collected data. Elements of DEMF are described in [2], [3] and they use hash functions as one of the mechanisms to confirm the integrity of the data. DEMF with its concept answers the questions who, what, when, how, where and why, which fully ensures the integrity of digital evidence, but also its confidentiality and availability. Figure 1 shows a rough diagram of the DEMF concept where two elements that are most important for the process of proving integrity are marked red and that are *report* and *judge*.

The concept of DEMF is universally applicable and allows different variations according to the area of

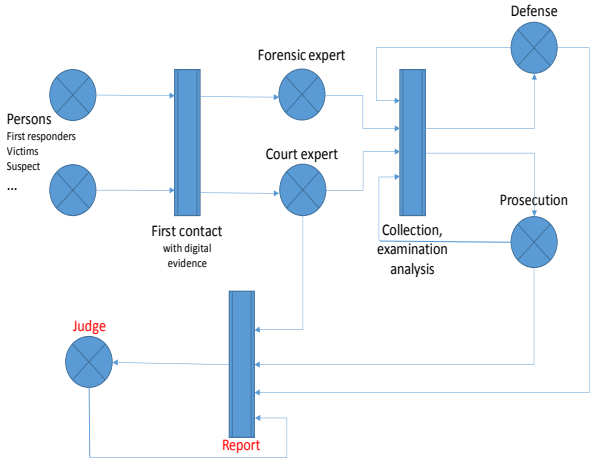


Figure 1 Processes in DEMF model

[1] Chain of Custody is sometimes called Chain of Evidence in literature and scientific papers

application. Further in this paper, the application of the DEMF concept on files whose content and metadata are not known will be described, which is especially important for collection, testing and analysis modules.

The importance of this is shown through the UML

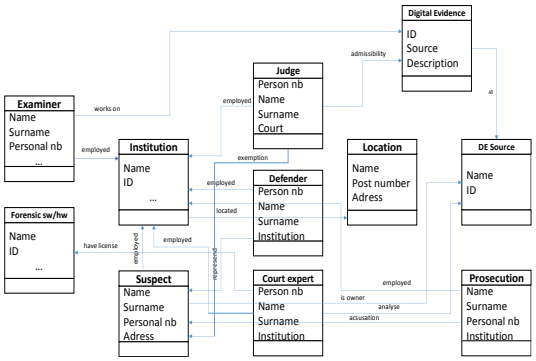


Figure 2 DEMF UML model

concept in Figure 2.

As it can be seen from the Figure 2, the system of integrity proofing in the court as well as the chain of evidence is based solely on the digital evidence and therefore the way to collect digital evidence is the most important factor in this process.

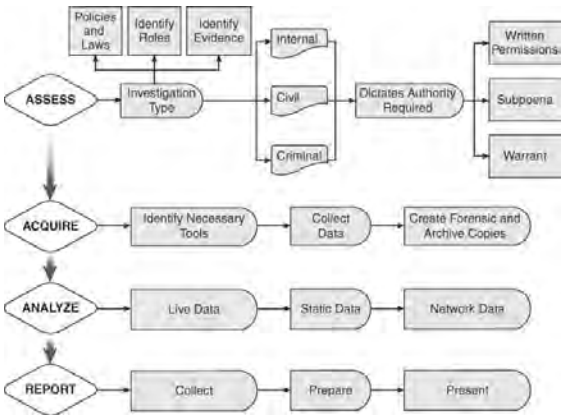


Figure 3 Forensic examination steps [6]

III. MEMORY DATA

Collecting memory content, along with the collection of the hard disk content and other holders of information, has become mandatory in the process of working with digital evidence. In order to collect the contents of RAM, it is necessary to run an application that will allow the collection of the data and that will affect the content of what is collected. Tools used to collect the contents of RAM are mainly open source based tools that do not necessarily comply with the technical standards regulating the collection of memory contents. A small number of commercial companies develop tools for the acquisition of RAM and are advertised as a standard for that purpose. Encouraged by the results obtained by Zimmerman[2] [1] and which are mentioned in the introduction and refer to the established methods, it can be concluded that the difference resulting from the use of various tools for the acquisition of RAM, is much higher than is the case with the acquisition of hard drives.

The volatility of the RAM content is not constant and depends on a number of different parameters like the amount of RAM and the number of running processes as well as to the characteristics of other peripherals that are in direct interaction with RAM memory. Therefore, the framework proposes, for working with the RAM memory, that the content is divided into blocks of fixed memory. This blocks will, during storage, be signed with all the elements contained in the DEMF.

This proposal has its negative aspects also. First of all, this refers to the fact that during memory acquisition a larger number of elements than is currently the case will be changed. The reason for this is the longer acquisition time required by DEMF in comparison with other state-of-the-art methods. Furthermore, an additional problem can be later work with a larger amount of elements. Today's tools save RAM memory as a single file which can facilitate the work if the files are of smaller memory range. However, in the case of files that have higher memory range (256 GB and larger), this can generate big problems.

In addition, there is the question, if the file is divided into smaller units, how to ensure the integrity of all the individual parts, and whether the process itself affects the quality of captured RAM images.

The methods used today for RAM acquisition have their advantages and disadvantages. There are two types of RAM acquisition: software and hardware. Software acquisition is widespread today, which can be used to acquire the content of RAM through tools such as dd. The problem with them is that one piece of content (possible evidence) will be rewritten because the tool itself must be installed in RAM. The amount that will be lost depends of course on the size of the memory itself, and the smaller it is, the larger part of content will be deleted. Hardware methods have different constraints and are less widespread than software. They rely on the fire wire for acquisition because it is possible to make a Direct Memory Access via IEEE 1349 [4]. This method is much better than software because it does not cause changes in RAM but it requires certain prerequisites that must be met and which are not common in the daily work (for example, the user should not shut down the computer, the computer has to be unlocked and the computer should have a specific type of fire wire adapter).

IV. COMPUTER NETWORK DATA

Network forensics deals with the capture, recording and analysis of network events in order to discover

[2] Results can be downloaded from https://docs.google.com/spreadsheets/d/1wXX5zYql7KIPgrsDdt6S5bTuGt_WRjWaBde1D0fhG5k/edit?type=view&gid=0&f=true&sortcolid=11&sortasc=true&rowsperpage=250&pref=2&pli=1#gid=0

evidential information about the source of security attacks in a court of law [5].

Depending on what network technology is applied to transfer data, different technologies for acquisition can be applied. For the purposes of this paper, the example of the most common method in Ethernet infrastructure will be given (for example, IEEE 802.11x WiFi). Traditional tool for data collection with this type of network technology is tcpdump. The output of this tool is PCAP file. Tcpdump can be used to filter out collected network traffic.

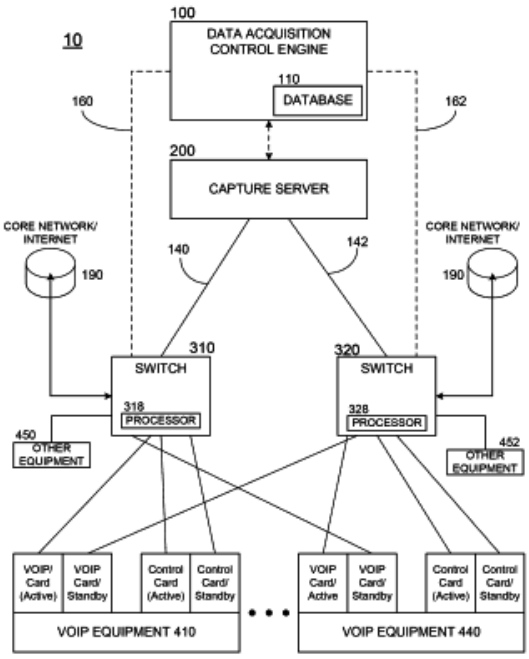


Figure 4 Patent network acquisition model [7]

Problems with data acquisition from the network lies in the fact that it is not known how much data can be collected. An example of such model is given in Figure 7.

„According to one aspect, the invention is directed to a communication network monitoring system that may include at least one switch serving as an intermediary to a plurality of data input streams and a plurality of data output streams; a capture server in communication with the at least one switch; and a data acquisition control engine operable to receive data acquisition instructions from a user and cause the received instructions to be implemented at the at least one switch. According to another aspect, the invention is directed to a method that may include presenting a graphical user interface (GUI) to a user by a data acquisition control engine, in a communications network; receiving data acquisition instructions from the user that specify a data acquisition plan; deriving commands to issue to one or more switches based on the data acquisition plan; and transmitting the derived commands to the one or more switches.“ [7]

In addition, there is a need for data acquisition from different services, for example IaaS (Infrastructure as a service). The paper [8] described a comparison of data acquisition with Amazon EC2, and it is shown how with present forensic tools data acquisition of that type is done. The differences between individual tools are pointed, from the amount of captured content to duration of data acquisition. Five different products showed similar acquisition results but still different enough that it can cause some doubt in court. On the side of the forensic process especially in an amount of acquisition time, the difference between the fastest and the slowest tool is 1:6.

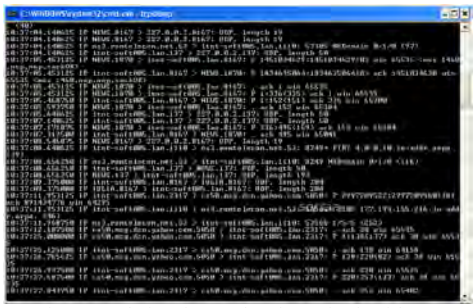


Figure 3 Chain of custody of acquired evidence

V. MOBILE DATA

The data collected from mobile phones and mobile devices are particularly sensitive to change. In the context of this study this passage refers to data that is collected during the "live" acquisition or with a mobile device that is operational and that cannot be turned off. Today's models are mainly based on two large commercial producers Cellebrite [9] and XRY [10]. In one part of mobile phones it is possible to make data acquisition using these tools without changing the content of mobile phones. It should also be noted that for certain mobile phones and operating systems on them, these commercial systems are unusable.

In order to conduct the acquisition from such devices it is necessary to use noncommercial tools, which implies the lack of standardization models as well as the testing of these tools, which can be a problem in court. One such system is Firefox OS [12], for which there is no commercial tool for the acquisition. For the purpose of the acquisition it is necessary to use alternative tools, such as [12]. However, when using such tools there is still a question as to what changed during installation of those tools on the mobile phone, whether it is possible to document and ultimately confirm the integrity of the data in court.

VI. PROPOSED MODEL

Regardless of whether the data is collected from RAM or from another source, difficulties that arise can be observed. In the first place is that the forensic scientist does not know the length of the memory to be collected, and that the length is not permanent. On the other hand, the installation of the tool itself causes changes in memory and proving the integrity of collected evidence is necessary.

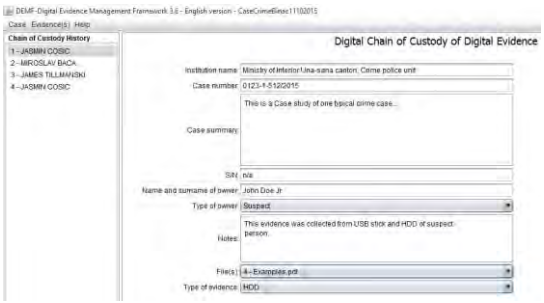


Figure 4 Chain of custody of acquired evidence

DEMF model can be used to prove the integrity of the collected evidence in court. In this sense, it is necessary to define a specific memory size of the package that will be collected. This step is necessary to reduce the impact of installed applications from the complete contents of the memory, to only part of it. As the memory changes, the overall size which is to be collected in this process should be defined.



Figure 5 Metadata acquired by using DEMF

When the individual package is collected, metadata from DEMF model need to be collected:

- specific task on the basis of which these data are collected, this may be a court order in case of official investigations by law enforcement agencies or management decisions in case of corporate investigations,
- first or/and last name of the person, or some other designation of the person who collects this data - identification can be done by a biometric characteristic or even digital signature stored on a personal documents.
- the geographic location of computer, or geo-data (latitude and longitude), and unique data on the computer to which the data is collected,
- a time stamp - the exact date and time obtained as a confirmation from an external TSA, as well as a public IP address.
- calculation of a hash function, selection from offered MD/SHA functions (to the level of 512 bytes)
- complete metadata files (author, time of creation, modification, last access, etc.)

Each individual package made is stored in a container. The containers are, at the end of the process, protected by a certain key - AES256 encryption. Key management procedures must be defined in institutions which are using DEMF, through a document called "Security policy". During the encryption process all metadata for the container, as well as for each individual package, are written (Figure 5).

In the process of proving the integrity it is possible, by analysis of each package as well as container, to read all metadata that are confirmation that the process is done in accordance with the preservation of the chain of evidence, which can guarantee the integrity of all the packages contained in a container.

Today, DEMF is initially tested in institutions in Bosnia and Herzegovina, and Croatia and some court expert witnesses for ICT are using DEMF tool to improve integrity of digital evidence acquired from seized devices.

VII. CONCLUSION

The idea of this paper was to give an overview of DEMF as a framework for managing digital evidence. The problems of collecting volatile digital evidence have been addressed, as were their possible solutions by using this model and software - DEMF. Memory and computer network data collection has been described. The main part of the paper was showing the usage of DEMF in collecting volatile digital evidence.

REFERENCES

[1] Zimmerman Eric, Speed testing of different forensic imaging tools, available at www.hecfblog.com, (accessed 01.02.2016.), 2013

[2] Ćosić, Jasmin; Bača, Miroslav; ,Do we have full control over integrity in digital evidence life cycle?,"Information Technology Interfaces (ITI), 2010 32nd International Conference on",,,429-434,2010,IEEE

[3] Ćosić, Jasmin; Bača, Miroslav; Leveraging DEMF to Ensure and Represent 5ws&1h in Digital Forensic Domain,International Journal of Computer Science and Information Security, Vol.13, No.2, pp. 7-10,2015

[4] Adam Boileau, Hit By A Bus: Physical Access Attacks With Firewire. Ruxcon, 2006

[5] Natarajan Meghanathan, Sumanth Reddy Allam and Loretta A. Moore: Tools and Techniques for Network Forensics, International Journal of Network Security & Its Applications (IJNSA), Vol. 1, No. 1, April 2009

[6] Graves, Michael Digital Archeology, The Art and Science of Digital Forensics, London, UK; O'Reilly-Media

[7] US Patent 9178791 B2, System and method for data acquisition in ana internet protocol network, 2015

[8] Dykstra, Josiah, Sherman, Alan Acquiring forensic evidence from infrastructure-as-a-service cloud computing: Exploring and evaluating tools, trust, and techniques, Digital Investigation 9 (2012) S90-S98

[9] Cellebrite forensic, available at www.cellebrite.com (accessed 30.01.2016.)
[10] MSAB Inc. availbable at www.msab.com (accessed 30.01.2016.)
[11] Mozilla, available at www.mozzila.org (accessed 15.01.2016.)
[12] GitHub repository, availbable at http://github.com/elninosi (accessed 30.01.2016.)

From Safe Harbour to European Data Protection Reform

Tihomir Katulić, Ph.D., Goran Vojković, Ph.D.

University of Zagreb,
Trg maršala Tita 14, Zagreb, Croatia
E-mail: tihomir.katulic@pravo.hr, goran.vojkovic@fpz.hr

Abstract – European personal data protection laws have set the electronic communication privacy standards for more than two decades. Among these standards, the Commission Decision 2000/520/EC of 26 July 2000 pursuant to Directive 95/46/EC of the European Parliament (The Safe Harbour Decision) stood out as a cornerstone of transatlantic data protection regime. The Court of Justice of the EU decision in Maximillian Schrems vs. Data Protection Commissioner in late 2015 has declared the decision invalid. In the light of the long standing legislative reform of the European Data Protection legal framework and the revelations of widely spread unauthorized electronic surveillance, data collection, interception and access by intelligence services and authorities of several countries, there is an urgent need for improved data protection rules, especially regarding collection and export data via cloud services established and hosted outside EU. The purpose of this article is to analyse publicly available reform proposals concerning in the light of the recent ECJ Safe Harbour decision, as well as the developments regarding the future EU-US Privacy Shield proposal.

I. Introduction

In October 2015, the European Court of Justice ruled that the long standing „Safe Harbour“ agreement concerning the transfer of European citizens personal data into the United Stated was no longer valid. The Court, among other concerns, raised the question of adequate protection of European citizens rights following numerous cases of state sanctioned surveillance of electronic communications perpetrated both in the US and among the EU Member States.

Invalidation of the Safe Harbour agreement is a last in the series of developments concerning the development of the European legal framework of personal data protection. Following the landmark Data Protection Directive (Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data) in 1995, the Union established what is probably the world's most comprehensive system of personal data protection.

With the exception of Argentina and several other states outside the EU, the level of protection envisaged by the EU framework has not been universally adopted. In this regard, the failure of the US to upgrade its privacy and personal data protection regulation to a more comprehensive system has presented a serious obstacle to future development of services based on collection and use of personal data.

While the new personal data protection agreement between the EU and the US is currently in the works, the strict implementation of EU personal data protection standards may impact the ability of US companies to collect, analyze and store personal data of European citizens.

The aim of this article is to analyze the current legislation reform proposals concerning the EU personal data protection framework with regard to collection, storage and use of personal data by entities based in countries whose legal standards of personal data protection differ from those established by the EU. We will also try to suggest solutions for facilitating a more balanced approach to protection of personal data from the perspective of European citizens having in mind the importance of cloud and personal data based services for the emerging information society economy.

II. Protection of Personal Data as a Fundamental Right in the EU

While it can be argued that most of the now recognized fundamental rights of the European citizens were codified even earlier, the Charter of Fundamental Rights of the EU proclaimed by the European Parliament in 2000 was a first EU document that expressively designates Protection of personal data as a fundamental rights. [1]

The protection is layed out in the provisions of Article 8 of the EU Charter titled "Protection of personal data":

1. *Everyone has the right to the protection of personal data concerning him or her.*
2. *Such data must be processed fairly for the specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right to access to data which has been collected concerning him or her, and the right to have it recitified.*
3. *Compliance with these rules shall be subject ot control by an independent authority.*

The provisions of the Article 8 effectively create protection of personal data as a fundamental right enjoyed by everyone [1], further enforced by the 2009 Treaty of Lisbon. [2]

III. Safe Harbour Privacy Scheme

After the Data Protection Directive went into effect in 1998, the pan-european data protection standards prohibited the transfer of personal data to countries that do not meet the EU standards of personal data protection.

Following a decision by the European Commission in 2000 that the US personal data protection principles comply with the standards adopted in the EU Data Protection Directive, the US Department of Commerce developed a framework of cooperation with the EU concerning the collection, storing, analysis and use of European citizens personal data under certain conditions.

These conditions were known as Safe Harbour principles: The principle of (giving) notice – the persons whose data was collected had to be informed that their data was being collected and how it was used. Persons subjected to data collection were to be provided with information how to contact the collecting organization with inquiries and complaints. The principle of choice – individuals were required to be allowed the option to opt out of the collection of their personal data and its forwarding to third parties. Principle of Onward Transfer stipulated that data transfers to third parties could only occur when those other parties observed adequate data protection principles. The principle of Security required that reasonable efforts should be undertaken to prevent loss of collected information. The principle of Data Integrity required that collected data had to be relevant and reliable for the purpose it was collected for. The principle of access stipulated that individuals had to be able to access information held about them and correct or delete it if it was inaccurate. Finally, the principle of enforcement required an effective means of enforcing the Safe Harbour principles.

As mentioned above, the purpose of the Safe Harbour scheme was to facilitate export of European personal data to the US in order to be stored, analyzed and used by the emerging information society services (often referred to as Big Data) such as social networks, digital content delivery services etc. European Data Protection Directive mandated Member States implement legal framework preventing export of personal data outside the EU unless an adequate level of protection is guaranteed. [3] In order to facilitate export of personal data to the US, the US companies were allowed to enter the certification program and be certified to adhere to Safe Harbour principles. This was available solely to the US companies and other organizations as regulated by the US Federal Trade Commission and the Department of Transportation and the certification process was not regulated by the Government of the United States. Instead, it was implemented as a self-regulating process through private sector entities with nominal oversight by the FTC and the US Department of Commerce.

The informal, self-certification process was subject to a substantial amount of criticism over the duration of the Safe Harbour scheme. We find these criticisms fully justified. The US Department of Commerce had originally published the set of rules a company or other organization had to follow in order to qualify for the Safe Harbour scheme. However, the self-certification model involved required that participants fill out questionnaires and confirm adherence to Safe Harbour rules. Those who did were entered into the Safe Harbour list without further auditing or supervision. This procedure was both practically unsafe and legally questionable.

IV. Criticism of the Safe Harbour scheme

During the course of the scheme, there have been three external evaluations regarding the compliance of companies and organizations participating in collecting and transferring European citizens data to be analyzed and stored in the United States.

In February 2002, the European Commission published a working paper fulfilling the obligation set in the EU Parliament Decision in July 2000 [4]. The Commission was required to ensure that the operation of the Safe Harbour was closely monitored and to make periodic reports. The report analyzed data from the US Department of Commerce web site, from US public authorities and private sector organizations involved in dispute resolution and enforcing Safe Harbour provisions and from the EU Member States data protection authorities. The Commission concluded that while the required elements of the Safe Harbour agreement are in place and individuals are able to lodge complaints if they believe their rights were being denied, a substantial number of organizations that have self-certified adherence to the Safe Harbour agreement have not established a degree of transparency regarding the fulfilment of their obligations and very few individuals have exercised their right to complain. Furthermore while a wide array of sanctions to enforce Safe Harbour rules under dispute resolution mechanisms were envisaged in the original agreement, not all dispute resolution mechanisms have indicated intention to enforce Safe Harbour rules or have set up in place practices applicable to themselves.

In October 2004, the European Commission published the follow-up working paper [5]. The Commission established that by 2003, there were over 400 US companies that self-certified to the Safe Harbour standards. While the Commission was pleased to see that the provisions of the Safe Harbour agreement were embraced by a large number of US companies, the Commission was concerned about the number of self-certified organizations that have not published a privacy policy or that have adopted a policy non-compliant with the Principles leaving FTC without jurisdiction to enforce the principles of the Safe Harbour agreement. Commission reiterated its previous finding that established alternative recourse mechanisms (dispute resolution mechanisms) still failed to comply with applicable Safe Harbour requirements.

Finally, in 2008 European Commission published a report prepared by an Australian consulting company Galexia. [6] The Galexia report was the most critical to date, revealing of the actual state of protection of European citizens personal data when collected and used by US companies, especially when obtained through new information society services. [7]

In essence, what was considered by many in the EU as a means of providing effective, wide spread protection of EU citizens personal data was in fact a very limited scheme with severe limitations.

The peak number of voluntary participants was less than 3000. [8] Many popular services used by European citizens, such as Instagram, Pinterest, Wikipedia etc., have simply avoided complying with Safe Harbour principles and officially joining the scheme. Additionally, Safe Harbour was originally not applicable to services such as

airlines, banks, credit card companies and telecommunication service providers. [8]

Additional points covered by the Galexia report include transient nature of Safe Harbour protection, false claims of Safe Harbour membership, failure of organizations to provide information to consumers regarding dispute resolution mechanisms, inaccessibility of selected dispute resolution providers to ordinary citizens and the failure of dispute resolution mechanisms to take into account specific Safe Harbour rules etc.

Foremost of them, as stated by the report was the finding that membership status of the organisations in Safe Harbour scheme was not permanent. The report states more than 1000 organisations have left the scheme, and additional number left and subsequently returned. The report stated that there was no accurate list or archive of historic membership and former entries have been known to simply dissapear. [8 p.4]

The report also mentioned a worrying number of false claims in relation to Safe Harbor membership with well over 400 organizations making false claims regarding adherence to Safe Harbour rules in 2013. [8 p.4]

The report states that many of the selected dispute resolution providers are inaccessible to ordinary consumers, jeopardizing one of the most important compliance requirements in Safe Harbour scheme which requires organisations to provide information to consumers regarding dispute resolution. Report finds that key services such as those offered by the American Arbitration Association or the Judicial Arbitration Meditation service are too expensive for ordinary consumers. [8 p.5].

In general, European economy is based on efforts of small and medium enterprizes: “Small and medium-sized enterprises (SMEs) are often referred to as the backbone of the European economy, providing a potential source for jobs and economic growth.” [9] Transition to information society economy has allowed massive outsourcing of previously in house activities, fostering development of micro enterprises (up to ten employees). Lean and flexible by nature these enterprises usually cannot afford information technology specialists. Large organisations (big multinational companies, financial sector companies such as banks and insurance companies etc.) develop their information systems according to current information security standards (ISO 27000 family of standards, PCI DSS etc.) and usually have ample regulatory consulting services.

SMEs, on the other hand, are almost always left to themselves to try to navigate increasingly complex regulatory demands. It is therefore not feasible to expect SMEs to self-regulate - they might be able to satisfy such regulatory demands but in the case they are not, there is no reliable way to ensure their compliance.

V. Personal Data Protection in Europe after Schrems v. PDC

Earlier this year, following a petition by Maximillian Schrems, an Austrian citizen and a user of the popular social networking service Facebook, the Court of Justice has declared that the existing Safe Harbour scheme based on the Commission Decision 2000/520/EC was invalid. Mr. Schrems filed a complaint with the Irish Data Protection Commissioner questioning the efficacy of the legal and practical protection of personal data collected by Facebook and stored and analyzed in the United States, especially against surveillance by the public authorities of the United States. The Irish Data Protection Commissioner rejected the complaint on the ground of the European Commission Decision that established the Safe Harbour scheme between EU and the US. The Commissioner reasoned that under that scheme the US ensured an adequate level of protection for the transferred personal data. [10]

The Court judgement held that the existence of a Commission decision finding that a third country ensures an adequate protection of the transferred personal data cannot eliminate or reduce the powers available to the national supervisory authorities under the Charter of Fundamental Rights of the EU and the Data Protection Directive. The Court stated that even if the Commission has adopted a decision, the national supervisory authorities, when dealing with a claim, must be able to examine whether the transfer of a person’s data to a third country complies with the requirements laid down by the Directive, although it is ultimately the Court's task to decide whether the Commission decision is valid. [112]

The Court further observed that the scheme is applicable solely to the United States undertakings which adhere to it, and that United States public authorities were not themselves subject to it. Since the US national security, public interest and law enforcement have precedence over the safe harbour scheme the Court explained that it was only logical to recognize tha United States undertakings were bound to disregard the protective rules laid down by that scheme where they conflict with such requirements. From the European perspective, this allows United States to interfere with the fundamental rights of European citizens. [10, p.2] The Court observed that any legal framework permitting the public authorities to access on a generalised basis the content of electronic communication to be compromising the essence of the fundamental right to respect for private life [10.p3].

The Court also observed that lack of possibility for individuals to pursue adequate legal remedies compromised the fundamental right to effective juridical protection, and also that denying national supervisory authorities powers to examine the framework of personal data protection was not within Commission competence. (10.p.3].

The outcome of the Court decision was that the Irish supervisory authority was required to examine Mr. Schrem's complaint and to decide whether the transfer of personal data of Facebook's European users to the US should be suspended on the grounds that the US legal framework does not provide an adequate level of personal data protection.

Following the decision by the European Court of Justice abolishing the Safe Harbour scheme the response by the European Commission and the specialized information security agency, ENISA (European Network Information Security Agency) has been lacklustre.

The modern internet is huge - some sources say there are: "over 284 million registered domains… on over 108 million hosts provided… by 5 million computers serving… over 876 million websites.” However, only a tenth of one

percent or less than 1 million account for over half of all web traffic. [11] The most popular Internet services, social networks like Facebook, LinkedIn, Instagram, Google and Youtube, and many other leading services are hosted by companies founded and headquatered in the US. The ruling of the European Court of Justice will have very limited effect on the business practices of companies largely outside of European jurisdiction and US companies will continue to collect and store data on European citizens.

We can observe this even locally. Croatian membership in LinkedIn numbers over 400,000 users, and well over a million and a half Croatian citizens use Facebook. Instagram has almost two hundred thousand users in Croatia [12].

VI. TOWARDS THE NEW SAFE HARBOUR AGREEMENT

In April 2016, the EU Parliament has adopted new data protection rules which now focus on giving the citizens a recourse to take back control of their personal data. The measures adopted also set minimum standards on use of data for policing and judicial purposes, affirming consumer rights and competition in the nascent European digital single market.

Among the new regulations, a few provisions immediately stand out. The new rules enshrine the right to be forgotten, a notion of clear and affirmative consent required by the person concerned as a prerequisite to the processing of private data, a right to transfer persona data to another service provider, the right to know when personal data has been hacked, a requirement ensuring that privacy policies are explained in clear and understandable language etc.

On the enforcement side, the Parliament has adopted measures allowing fines up to 4% of infringers total worldwide annual turnover. Historically, especially in competition cases against US technology companies such as Intel or Microsoft, fines limited to two percent amounted to hundreds of millions of euros, or even 1.06 billion euros (Intel). With this new upper limit, the fines in a potential case against Google could reach up to 6 billion euros.

The new data protection package also includes measures on data transfers for policing and judicial purposes, applying to data transfers across members states and setting minimum standards for policing purposes within each member state.[13]

. Earlier this year, the European Commission and the United States have agreed on a new framework for transatlantic data flows: the EU-US Privacy Shield.

The new arrangement will include the following elements:

First, strong obligations on companies handling Europeans' personal data and robust enforcement: U.S. companies wishing to import personal data from Europe will need to commit to robust obligations on how personal data is processed and individual rights are guaranteed. The Department of Commerce will monitor that companies publish their commitments, which makes them enforceable under U.S. law by the US. Federal Trade Commission. In addition, any company handling human resources data from Europe has to commit to comply with decisions by European DPAs.

Second, clear safeguards and transparency obligations on U.S. government access: For the first time, the US has given the EU written assurances that the access of public authorities for law enforcement and national security will be subject to clear limitations, safeguards and oversight mechanisms. These exceptions must be used only to the extent necessary and proportionate. The U.S. has ruled out indiscriminate mass surveillance on the personal data transferred to the US under the new arrangement.

Thirdly, effective protection of EU citizens' rights with several redress possibilities: Any citizen who considers that their data has been misused under the new arrangement will have several redress possibilities. Companies have deadlines to reply to complaints. European DPAs can refer complaints to the Department of Commerce and the Federal Trade Commission. In addition, Alternative Dispute resolution will be free of charge. [14]

These provisions represent a significant step ahead with regard to the old Safe Harbour self-evaluation model. Further research, especially of the practical side of the implemented model is required. The EU needs to adopt a model of regular independent audit concerning the implementation of the new regulation and the new data transfer scheme.

VII. CONCLUSION

The existing European legal framework regarding personal data protection is based on two main pillars, the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention No. 108) Council of Europe, and Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. The Directive provisions have been widely adopted and implemented as were the principles of data protection from the CoE Convention.

Faced with recent challenges, such as Snowden revelations and other instances of illegal and unsanctioned access and interception of personal data and communication it is clear that existing framework is showing its age.

While new personal data legislation, this time in form of a binding regulation is scheduled to be introduced very soon, the practice of personal data protection is something that needs to be addressed on political and economic levels as well as legislative. The measures adopted by the European Parliament and the future EU-US Privacy Shield might very well improve the state of personal data protection for the citizens of the European Union. However, in order for these improvements to truly come about a vigilant survey of the application of the new regulatory model is required, especially in the light of previous efforts and practical results.

REFERENCES

[1] Fuster, Gloria Gonzalez: The Emergence of Personal Data Protection as a Fundamental Right of the EU, Law,

Governance and Technology Series 16, Springer International Publishing Switzerland, 2014

[2] Treaty of Lisbon amending the Treaty on European Union and the Treaty establishing the European Community, [2007] OJ C306/1.

[3] Data Protection Directive, Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such dataEuropean Court of Justice 2000/520/EC: Commission Decision of 26 July 2000 pursuant to Directive 95/46/EC of the European Parliament and of the Council on the adequacy of the protection provided by the Safe Harbour privacy principles and related frequently asked questions issued by the US Department of Commerce (notified under document number C(2000) 2441)

[4] The application of Commission Decision 520/2000/EC of 26 July 2000 pursuant to Directive 95/46 of the European Parliament and of the Council on the adequate protection of personal data provided by the Safe Harbour Privacy Principles and related Frequently Asked Questions issued by the US Department of Commerce, SEC(2002) 196., available at: http://ec.europa.eu/justice/data-protection/international-transfers/adequacy/files/sec-2002-196_en.pdf

[5] The implementation of Commission Decision 520/2000/EC on the adequate protection of personal data provided by the Safe Harbour privacy Principles and related Frequently Asked Questions issued by the US Department of Commerce, SEC (2004) 1323, available at: http://ec.europa.eu/justice/data-protection/international-transfers/adequacy/files/sec-2004-1323_en.pdf

[6] "The US Safe Harbour – Fact or Fiction", Galexia Pty Ltd., 2008., http://www.galexia.com/public/research/assets/safe_harbor_fact_or_fiction_2008/safe_harbor_fact_or_fiction.pdf (22.2.2016)

[7] EU/US Safe Harbor – Effectiveness of the Framework in relation to National Security Surveillance, http://www.europarl.europa.eu/document/activities/cont/201310/20131008ATT72504/20131008ATT72504EN.pdf (22.2.2016)

[8] Connoly, Chris: EU/US Safe Harbor – Effectiveness of the Framework in Relation to National Security Surveillance, Galexia, Speaking / background notes for an appearance before the Committee on Civil Liberties, Justice and Home Affairs (the LIBE Committee) inquiry on "Electronic mass surveillance of EU citizens", Strasbourg, October 7 2013, available at: http://www.europarl.europa.eu/document/activities/cont/201310/20131008ATT72504/20131008ATT72504EN.pdf (22.2.2016)

[9] http://ec.europa.eu/eurostat/web/structural-business-statistics/structural-business-statistics/sme?p_p_id=NavTreeportletprod_WAR_NavTreeportletprod_INSTANCE_vxlB58HY09rg&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-2&p_p_col_pos=1&p_p_col_count=4 (22.2.2016)

[10] European Court of Justice, ECJ Case C-362/14 Maximillian Schrems v Data Protection Commisioner http://curia.europa.eu/jcms/upload/docs/application/pdf/2015-10/cp150117en.pdf (22.2.2016)

[11] http://tekeye.biz/2014/how-many-websites-are-there (22.2.2016)

[12] http://www.netokracija.com/broj-korisnika-instagrama-u-hrvatskoj-108541 (22.2.2016)

[13] Personal data protection: processing and free movement of data (General Data Protection Regulation), http://www.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2012/0011(COD)&l=en, (14.4.2016)

[14] EU Commission and United States agree on new framework for transatlantic data flows: EU-US Privacy Shield http://europa.eu/rapid/press-release_IP-16-216_en.htm (29.4.2016.)

Information Security Assessment in Nature Parks

Saša Aksentijević [1], Toni Đugum [2], Krešimir Šakić [3]
[1] Aksentijević Forensics and Consulting, Ltd.
Gornji Sroki 125a, Viškovo, Croatia
Tel: +385 51 65 17 00 Fax: +385 51 65 17 81 E-mail: axy@vip.hr
[2] National Park Krka
Trg Ivana Pavla II br. 5, Šibenik, Croatia
Tel: +385 22 20 17 77 Fax: +385 22 33 68 36 E-mail: toni.dugum@npk.hr
[3] National Park Krka
Trg Ivana Pavla II br. 5, Šibenik, Croatia
Tel: +385 22 20 17 77 Fax: +385 22 33 68 36 E-mail: kresimir.sakic@npk.hr

Abstract – In this paper, specific requirements for information security assessment will be identified along with proposal for the model specially tailored to suit the audit needs of the information systems within nature parks - protected and conserved areas. Elements of information security management system for nature parks are described along with specific information security elements called situations, to be evaluated. Proposed model is set in a way to provide quantitative evaluation of the overall system of information security management. The audit approach involves definition of cardinal events of information security, assets, threats and related vulnerabilities. Final outcome of the assessment is not only mark that represents overall state of affairs within information security management system, but also a set of recommendations for remedial measures and implementation of controls of information security in nature parks in order to elevate already achieved and determined compliance level. Proposed model is tested in case of nature park "Krka" in Croatia and the test results provide unique and adequate insight in achieved level of information security management system compliance.

Key words: information security, nature park, audit, quality systems, national park Krka

I INTRODUCTION

Standard definition of the nature park is that it is an area of countryside, or occasionally sea or fresh water, protected by the state for the enjoyment of the general public or the preservation of wildlife: commercial exploitation of natural resources in a national park is illegal [1]. Internationally recognized definition of the nature parks is the one according to the IUCN organization, International Union for Conservation of Nature whose core mission is to help the world identify pragmatic solutions to the most pressing environment and development challenges. IUCN considers national parks to be under the 2nd category of protection, and defines them as areas represented by large natural or near natural areas set aside to protect large-scale ecological processes, along with the complement of species and ecosystems characteristic of the area, which also provide a foundation for environmentally and culturally compatible, spiritual, scientific, educational, recreational, and visitor opportunities [2]. Primary objective of existence of nature parks is protection of biodiversity with underlying ecological structure, supporting environmental processes and promotion of education and recreation [3].

IUCN anticipates existence of two other forms of protected areas before national parks:

- Ia – Strict Nature Reserve, and
- Ib – Wilderness Area.

There is also a number of protected areas of lesser significance in categorization, just below nature parks:

III – Nature Monument of Feature,
IV – Habitat/Species Management Area,
V – Protected Landscape / Seascape, and
VI – Protected area with sustainable use of natural resources.

In this paper, a model of information security assessment especially suited for use in audits of information security management systems in nature parks is described. It is important to emphasize that according to Croatian Nature Protection Act [4], aligned with EU directives, the division of protected natural areas is slightly different. This law recognizes "national park" as a form of protected area one level below strict nature reserve, and two levels above nature parks. However, for the purpose of this research, theoretical model presumptions are taken from IUCN's internationally recognized nature parks, while field testing of model is performed using real processes of a national park according to Croatian legislation, in function, similar to IUCN's definition of nature park.

Ensuring information security compliance means adoption of a system that is able to respond to question – what is the achieved level of compliance to predefined set of rules of information security, and consequentially, identify existing gaps and measures used to close them in order to approach as much as possible the ideal model of information security management. Golden standard for information security in corporate environments today is ISO 27000 series of standards, that strictly define and govern this process, including document and process requirements, identification of scope of information security, information security assets and their characteristics (vulnerabilities and threats that can exploit identified vulnerabilities), checks of various controls that aim to address those vulnerabilities and threats and final conclusion and remedial actions. This process is performed in famous ISO PDCA cycle ("Plan-Do-Check-Act"), and they are subject to perpetual improvements and implementation of

checks and corrective actions. This cycle is shown in Figure 1.

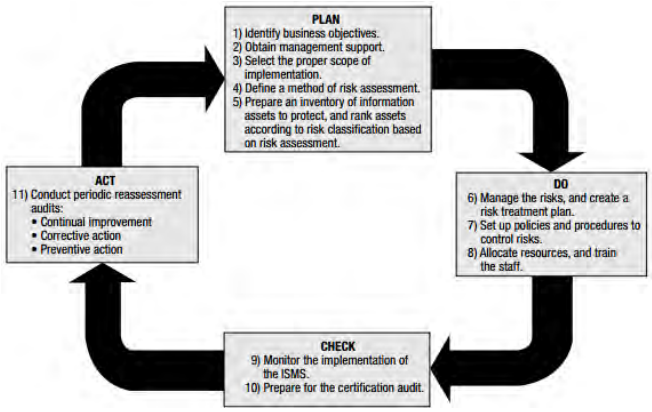


Figure 1. PDCA cycle of information security implementation [5]

Most information security management methods and underlying standards claim to be equally applicable to organizations of all sizes and types, regardless of ownership structure, goals or achieved maturity levels. However, practitioners of information security are well aware of the fact that *"one size fits all"* approach may sorely fail when faced with various facets of reality. Some of the reasons for this can be identified as follows:

1. Achieved level of information security,
2. Corporate culture,
3. Culture of information security,
4. Complexity of business processes,
5. Available financial resources,
6. Level of management maturity, and
7. Legislation.

Sometimes, detrimental influences to information security management whose sources lie in the DNA of organizations may be of such magnitude that they completely distort overall information security inside such organizations. Furthermore, it is possible to envisage, for example, small organizations that are financially and organizationally very weak, yet they strongly require adequate information security (for example, start-ups), because their sole existence relies on strong information security – and mainly information confidentiality. This is the reason why some authors have started to recognize that theoretical blueprints for information security cannot be used as a single unique model without adjustments aimed towards specific implementations.

During preliminary research of available online resources, it was not possible to identify a single model that would be adjusted for use in nature parks. Information security and nature parks can be identified only in terms of their information security policies that are publicly available. It seems that this activity is best underway in Japanese and Taiwanese national parks, while, incidentally, Japan was among the first countries to embrace ISO certification of information security managements systems, and most such certificates are issued exactly in Japan [6]. Some examples of such information security policies are those of select Taiwanese national parks: Yangmingshan National Park [7], Shei-pa National Park [8] and Taroko National Park [9].

Understanding business and process specifics of nature park operations on one side and information security compliance on another, our working hypothesis is the following: based on the available information security implementation models it is possible to propose a model for evaluation of information security best suited for use in nature parks. This model will be tested using available processes in National park Krka in Croatia.

II SPECIFICS OF OPERATIONS AND INFORMATION SECURITY IN NATURE PARKS

In order to propose a methodology for information security management in nature parks information systems', it is important to recognize what is specific for operations of nature parks, and what divides them from other organizations (corporations, NGOs or SMEs) implementing rules of information security and compliance. These characteristics will prove to be very important for creation of a specific model, best suited for nature parks.

Identified characteristics of management and operations of nature parks are the following:

1. *Dependence.* Nature parks are usually institutions founded by the state itself under separate legislative acts and therefore, nature parks do not have full management autonomy. They are subject to specific laws, regulations, directives and international laws,

2. *Financing.* Nature parks are subject to strict financing, but as autonomous subjects providing also tourist and visiting services, usually they have abundant own sources of income.

3. *Distribution.* Nature parks are often distributed across large geographic areas, and their significance is usually regional, limited to a single county, region or other form of local governance. Operations are executed from distributed centres, varying in size and architecture, while support operations and top management are usually in a separate location, making overview of all operations difficult. Distributed operations in the field of nature parks require a specific mix of BYoD technologies [10], nomad computing and distributed data processing.

Geographic layout of the National Park Krka including its main operations locations is shown in Figure 2.



Figure 2. Layout of National Park Krka, Croatia [11]

4. *Workforce*. Operations of nature parks are highly seasonal; depend on the geographic location, some nature parks have peaks of visits during summer and/or winter periods, with large parts of the year being relatively underutilized or scarcely visited. This means that most nature parks experience peaks in temporary employments during peak periods in the year. Introduction of temporary workers employed in nature parks presents a challenge for organization of information security

5. *Connectivity*. Operative locations of nature parks are usually in rural areas, out of the way of main dana links provided by large ISPs. Due to geographic traits (forests, large open sea areas), they are sometimes difficult to connect even by using mobile dana services. Propagation of wireless signal is also sometimes difficult and dampened by natural landscape (trees, hills, depressions). Nature parks' management often has to invest more in infrastructure, undertake complex connectivity projects and finally, pay more for operations of such systems, than some other similar organization that functions under urban scenarios.

6. *Strict adherence to SLAs*. Most nature parks experience high revenue from incoming visitors during peak times (and hours). From business continuity aspect, nature parks can easily quantify the cost of service disruption. Therefore, strict adherence to SLAs with providers of Internet and other services is of utmost importance for management of nature parks [12].

7. *Billing systems.* Billing systems, their availability, integrity and confidentiality of contained dana is of utmost important for nature parks. Most of the income of nature parks comes from tickets and entry fees for visitors, and related services provided to visitors (dining, entertainment and sports/recreation). Proper and uninterrupted functioning of billing systems is the most important single factor in operations of nature parks.

8. *Seasonality.* Seasonality of nature parks' operations is already mentioned in relation to workforce, but it is also important for the number of incoming visitors and planning of operations. Nature parks experience not only seasonal and annual peaks and lows, but also local peaks and lows during high and low seasons, depending on the precipitation and daily temperature. Therefore, weather forecast and extremes have high influence on operations of nature parks, but also present a difference in load on information systems and information security management system.

9. *Organization.* Organizational chart of nature parks is usually very complex despite relatively modest number of permanent staff, because nature parks need to attend to various field aspects of operations, and have a number of general service departments and groups. This requires formation of very diverse hierarchical structure. Usually, nature parks do not use matrix or project organizations. In the Figure 3., a typical organizational chart of the national park in Croatia is shown. While exact constituents of the chart are not that important for the context, the chart is included to show developed and distributed organization of a national park. While the typical number of employees is not large, there are many organizational units and sub departments. This fact requires inclusion of a large number of processes in the evaluation of information security and presents a special challenge in organization and subsequent assessment of information security in nature parks.

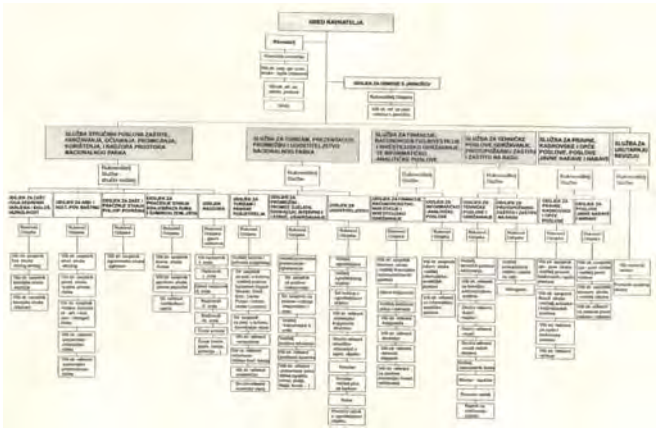


Figure 3. Typical example of the complex organizational chart – National park "Krka", Croatia [13]

10. *Compliance.* Owing to the dependant organizational nature, information security management systems of nature parks are subject to complex compliance requirements that are not only professional, but also legislative, and depend on the goals and mission set by their founders and the State.

These ten special requirements for information security audit in nature parks present a special challenge for those in charge of setting up and maintaining information security in nature parks. However, they also form a set of rules for visitors and transient IT system users and should be taken in consideration by those that are tasked with answering the most important questions for the management of nature parks:

1. How safe are our information and information processing facilities,
2. What should we do in order to increase our compliance, and
3. How much do proposed actions cost us in terms of financial impact and invested time for implementation?

III PROPOSAL OF MODEL FOR EVALUATION OF INFORMATION SECURITY IN NATURE PARKS

Thorough analysis of various systems of information security audit has lead the authors to work of David Brewer and Michael Nash from 2010 [14]. These two authors recognize that ISO 27001 standard requires organizations both to carry out a risk assessment and select controls and measures relevant to information security of their systems. Relationship between these two requirements is often unclear and murky.

The authors have spent almost four years investigating these relationships and demonstrated that ensuring the coverage of controls of Annex A of ISO 27001 standard limits the scope of risk assessment, while fulfilment of risk assessment of the standard may result in coverage of controls of Annex A but not necessarily providing addressing of the real exposure to information security risks. This relationship is shown in the Figure 4.

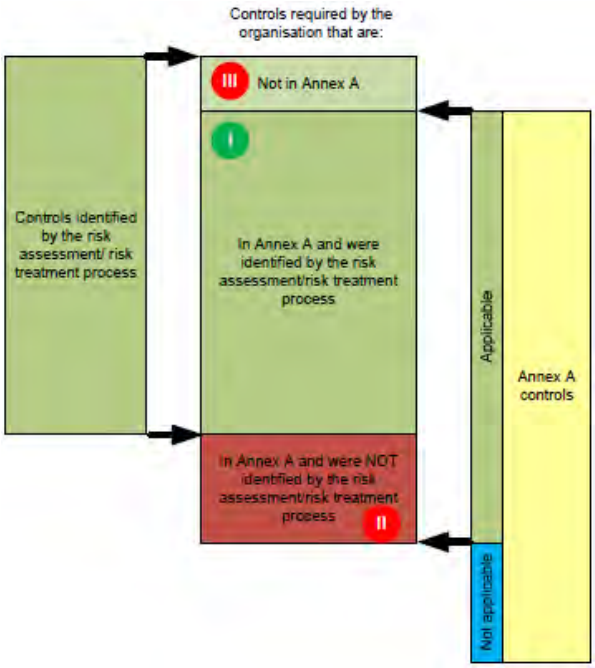


Figure 4: The cross-checking process [14]

In the cross checking process, there are controls identified by the risk assessment and risk treatment process (shown in the left column), controls of Annex A (divided into applicable and non-applicable in the right column). However, real situation of the information security management system is shown in the middle: clearly, there might be controls that are not present in the Annex A and those that were by mistake or negligence not identified by the risk assessment and/or treatment process.

As a result of this approach, concerns of the management related to information security are addressed, where concerns are mixture of *events* and *impacts* of those events on operations. This defines eight major events [14]:

S1 – Theft,
S2 – Acts of God, vandals and terrorists,
S3 – Fraud,
S4 – IT failure,
S5 – Hacking,
S6 – Denial of service,
S7 – Disclosure,
S8 – Law.

Three possible impacts of the major identified events are as follows:

B1 – Inappropriate deployment of people,
B2 – Failure to maintain proper records and
B3 – Issuance of wrong documents.

Relation plot of these events and controls of Annex A is shown in Figure 5. This model further recognizes that there is not one easily perceptible information security breach event in the sequencing story of information security, rather, it has to be broken down in some way. The authors have recognized three possible candidate (cardinal) events:

EI1 – Vulnerability exploitation,
EI2 – IT failure, and
EI3 – Disposession.

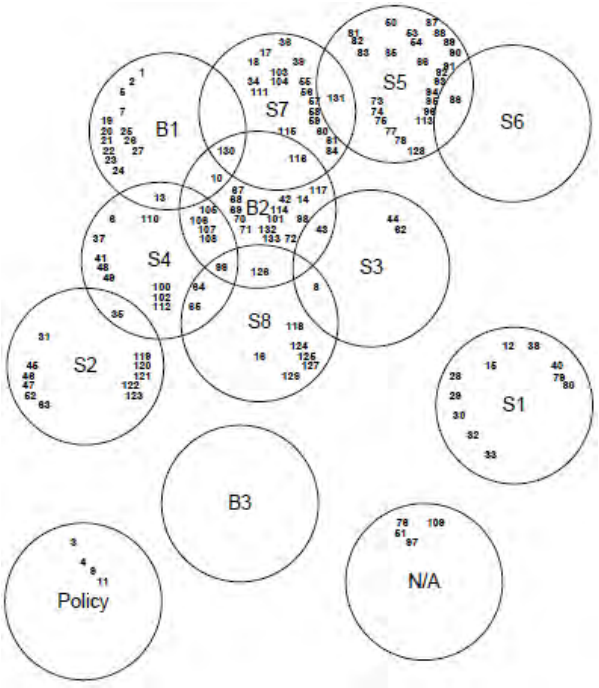


Figure 5. Relation plot of events and Annex A controls [14]

The authors of the model have further created a Venn diagram relating identified three cardinal events to Annex A controls, shown in Figure 6.

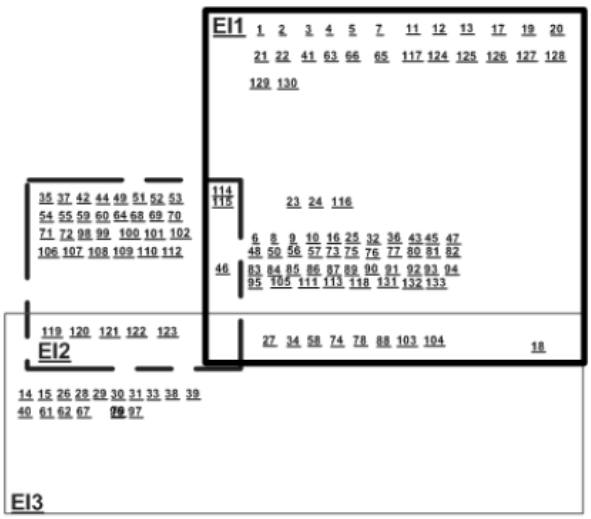


Figure 6: Relationship between the cardinal events and the Annex A controls [14]

In our approach, Brewer-List's methodology [15] was used superimposed over the scenario of the National park Krka's ISMS.

Nine different situations are anticipated:

1. *Solution implementation*: reduction of probability of information security incidents caused by solution vendors,
2. *Security of the workplace*: limiting access to informations in the workplace,
3. *Dislocated computing*: care about use of information and systems outside protected information security perimeter,
4. *Open computer access*: control of physical access to computers in the workplace,
5. *Remote actions*: protecting computer systems from cyber-attack,
6. *Applications*: ensuring security of computer applications,
7. *Working conditions*: ensuring uninterrupted functionality of the hardware;
8. *Information security status*: checking information security management system before the attack or incident occur, and
9. *Incident management*: undertaken steps in case of incident or attack.

For each of these nine situations, applicable situational fragments are identified and they are mapped onto applicable controls of annex A of ISO 27001 standard.

Due to restricted space available, not all situations and applicable controls will be discussed in this paper, yet it has to be clearly stated that described methodology covers all controls, control objectives and domains stated by Annex A of ISO 27001 standard. Also, one dilemma that authors faced was whether to use new version of the standard (ISO 27002:2013) or stick to the old one. It was decided early on to use the older standard because of two reasons: *first*, the model is well developed and described, readily available, and does not require further adjustments and *second*, using older standard has enabled immediate audit of the ISMS without further delays. Finally, current architecture of the information infrastructure and services would not significantly benefit from usage of the newer standard, so it was decided early on to use the old one.

V IMPLEMENTATION OF THE MODEL IN NATIONAL PARK KRKA, CROATIA

After setting up and modifying the initial information security assessment model, the audit of ISMS was undertaken in National park Krka in Croatia. The first step of the pre-audit was gathering all related information in order to properly identify all information assets. This includes the following:

1. List of ICT assets and services (servers, computers, data storage, payment gateways, ERP and CMS systems, internal and external data processing),
2. List of ICT personnel with role description,
3. Organizational chart,
4. Available network topology blueprints (leased lines, ISPs, wireless and fixed network, optical network),
5. Quantitative ICT data (number of users, desktop and notebook computers, mobile phones, budget),
6. List of computer equipment and service vendors,
7. Ongoing and past contracts for ICT equipment and services, and
8. Available ICT procedures and related/applicable legislation.

After preliminary analysis of the data, audit plan was compiled. An advantage of the described model is the fact that audit can be performed quickly if pre-audit phase was completed thoroughly. All applicable matrixes, situations and controls are anticipated in advance so they were readily available during real field audit of the system. The flash-audit was carried during two days. Before the beginning of the audit and after the audit was completed, there was a brief meeting with the management where the methodology was initially described and findings were presented. Focal point for the audit was IT manager, while regulatory system was discussed with the Legal manager and security of practices and procedures related to ERP and CMS systems was discussed with Administration and Finance manager. The audit plan is shown in Table 1.

Day 1	Activity	Duration	Participants	Note
• Morning	Kick-off meeting	0.5 h	The Management ICT Department Consultant	A brief initial meeting, mandatory presence of the Management
	General overview of the business and the ICT procedures	0.5 h	ICT Department Consultant	
	Situation 1	1.5 h	ICT Department Consultant	
	Situation 2	1.5 h	ICT Department Consultant	
• In the afternoon	Situation 4	2 h	ICT Department Consultant	
	Situation 5	2 h	ICT Department Consultant	
Day 2	**Activity**	**Duration**	**Participants**	**Note**
• Morning	Situation 6	1.5 h	ICT Department Consultant	
	Situation 7	1.5 h	ICT Department Consultant	
• In the afternoon	Situation 8	1.5 h	ICT Department Consultant	
	Situation 9	1.5 h	ICT Department Consultant	
	Personal presentation of the findings	1.5 h	The Management ICT Department Consultant	Presentation of findings, ensure presence of the client's Management

Table 1: Audit plan

During audit, all situations, situational fragments and controls of Annex A that fall under them, and that were previously described, are thoroughly analysed in terms of related information assets, their vulnerabilities and related threats. Special care was given to analyze information security of working personnel (permanently and temporarily employed), remote computing and billing system. The end result of the process is matrix of applicable controls of Annex A and identified aspects of information security. These identified aspects are:

1. Information security of human resources,
2. Physical security, security of services and environment,
3. Technical security,
4. Systems security,
5. Business continuity, and
6. Risk avoidance.

Adherence to identified controls is summed up in Table 2.

Number	Symbol	Meaning	Number of controls	Ponder	SUM
1.	N/A	Information security control is not applicable for the analyzed ISMS			
2.	☺	Information security control is respected		5	
3.	😐	Information security control is partially respected		3	
4.	☹	Information security control is not respected		1	
SUM					
Maximum possible mark			133	5	665
Maximum possible mark corrected for not applicable controls				5	
Estimated average of implementation of controls of Annex A			x,xx on a scale of 1 to 5		

Table 2: Summary table of adherence to identified controls of Annex A

There are four possible statuses of identified controls. Control can be *not applicable*; it can be *respected* or *not respected*. Considering that the proposed audit model is a quick derivative model, authors are not required to adhere to set rules of the ISO 27001 standard, so additional status – information security control is *partially respected* – was introduced in the model. Use of this evaluation of control has to be in practical terms limited to those cases where non-compliance is of lower impact on the ISMS as a whole and when just minor corrections can be quickly implemented to forward the control on to status of "respected". In terms of audit, these controls can be viewed as "minor non-conformities".

Finally, in order to create a synthetic quantitative mark that could give the management a clear overview of ISMS, ponders are introduced. Considering 133 possible controls, maximum sum of 665 points can be reached if ISMS respects all controls. This sum has to be corrected for those controls that are not applicable. Finally, grade "1" is given to control that is not respected, grade "3" to control that is partially respected and grade "5" to control that is fully respected. Dividing corrected sum with number of applicable controls will provide a single quantitative mark in range of 1 to 5, that is easily understandable to the management. However, when evaluating this mark, one has to be very careful to understand that it does not take into consideration possible damage derived from information incidents, as it equally treats all controls regardless of their impact. However, this model may be used as a "*quick and dirty*" litmus test for the state of ISMS. Furthermore, this model will provide easy identification of problem and improvement areas of the system in a quick and very efficient way.

Finally, at the end of the audit, a list of recommendations is produced in order to elevate non-compliance status to partial or full compliance, and elevate partial compliance status to full compliance. This list of recommendations is in line with initial PDCA cycle, because the newly reached state of ISMS can be again audited using the same model, or standard ISO 27001 auditing model. In both cases, it is reasonable that one expects higher level of compliance both with modified model of information security audit and standard ISO 27001 model. In case that the same modified model is used again, using synthetic mark, the management can track the progress in a clear way.

VI CONCLUSION

Overall management and management of information systems of nature parks is connected with various specifics that are derived from the fact that national parks are usually not fully independent, they are usually regionally distributed, their operations are highly seasonal and their operations depend on strict adherence to SLA for ICT services with billing systems being most sensitive and important because lack of their availability translates directly into quantifiable loss of income stream. Management is generally quite wary of information security management audits because they are strange and unknown to them, sometimes they do not see direct benefits but they are usually connected with high cost of information security solutions. Complex information security audits using conventional models just add to all these issues.

In our research, the main hypothesis was that it is possible to propose an audit system that would be better adjusted to specific requirements of ISMS in nature parks. A hybrid model of information system audit is envisaged starting from ISO 27001 standard and modified Brewster-List methodology that replaces standard dilemma "*risk assessment*" vs. "*ICT control adherence*" with a new system that includes ICT security situations, fragments and events. Furthermore, a model was developed that quantifies adherence to identified applicable controls and translates it to a single mark easily measured and understood by the management for further evaluation.

This model was tested on example of the National park Krka in Croatia, where audit was applied in real life situation, final snapshot of the information security system is created and synthetic grade of maturity of the system is produced for further evaluation by the management. End result of the described process is a list of suggestions for improvement, whose implementation would increase synthetic grade.

There are further possibilities of research of this model. For example, it can be modified to suit some other applications, and synthetic grading system can be made more complex, in order to better approximate real situation or to follow some other grading system (applicable grading could range from 1 to 10 or relative percentages can be utilized).

REFERENCES

[1] Oxford Dictionaries – Languages Matter, http://www.oxforddictionaries.com/definition/english/national-park (accessed 14th February 2016.)

[2] Philips, A., Harrison, J. "International Standards in Establishing National Parks and Other Protected Areas", The George Wright Forum, Volume 14, Number 2, 1997.

[3] IUCN, International Union for Conservation of Nature, http://www.iucn.org/about/work/programmes/gpap_home/gpap_quality/gpap_pacategories/gpap_pacategory2/ (accessed 14th February 2016.)

[4] Official gazette of Croatia, NN 8/13

[5] Pelnekar, C. "Planning for and implementing ISP 27001", ISACA Journal, Volume 4, 2011.

[6] ISO survey 2014., http://www.iso.org/iso/home/standards/certification/iso-survey.htm?certificate=ISO%209001&countrycode=JP#countrypick (accessed 14th February 2016.)

[7] Information Security Policy, Yangmingshan National Park, http://www.ymsnp.gov.tw/index.php?option=com_content&view=article&id=549&gp=0&Itemid=568 (accessed 14th February 2016.)

[8] Information Security Policy, Shei-pa National Park, http://www.spnp.gov.tw/v2/Article.aspx?a=jNakCHfRhqw%3D&lang=1 (accessed 14th February 2016.)

[9] Information Security Policy, Taroko National Park, http://www.taroko.gov.tw/English/?mm=0&sm=0&page=5 (accessed 14th February 2016.)

[10] Evans, D. "What is BYOD and why is it important", Techradar, 07. October 2015.

[11] Hagi@Sophia, https://hagia27sophia.wordpress.com/sem-eira-nem-beira/croacia/parque-nacional-krka/ (accessed 14th February 2016.)

[12] "What is a service level agreement", Palo Alto Networks, https://www.paloaltonetworks.com/resources/learning-center/what-is-a-service-level-agreement-sla.html (accessed 14th February 2016.)

[13] National Park Krka, internal organizational chart, 2015.

[14] Brewer, D., Nash, M. "Insights into the ISO/IEC 27001 Annex A", Gamma Secure Systems Limited, 2010.

[15] Brewer, D.F.C., List, W. "Measuring the effectiveness of an internal control system", http://www.gammassl.co.uk/research/time040317.pdf, Gamma Systems Ltd., March 2004. (accessed 9th April 2016.)

Clustering Approach for User Location Data Privacy in Telecommunication Services

Marin Vuković*, Mario Kordić**, Dragan Jevtić*
* University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** Ministry of the Interior, Zagreb, Croatia
marin.vukovic@fer.hr

Abstract - User location has become an important aspect of user's context that may bring valuable insights into user habits and preferences. Various services and applications tend to collect user location data for the purpose of analysis and providing personalized content to the users. This paper examines location data privacy across some aspects of location data processing regarding to European Commission ePrivacy directives. The intent of the new legislation is to strengthen and unify data protection for individuals within the European Union. It is necessary to strike a reasonable balance between the data controllers' business interests and the privacy of data subjects. The Data Protection Directive requires data controllers to observe a number of principles when they process personal data. These principles not only protect the rights of those about whom the data is collected but also reflect good business practices that contribute to reliable and efficient data processing. For this purpose, we propose a new approach for location data processing in which the location data is scaled corresponding to the type of service. Neural processing technique is used for location data clustering. This approach proposes a dedicated server used for location data clustering with adaptive cluster dimensions

I. INTRODUCTION

User privacy has always been an issue in telecommunication services. Unlike traditional services, nowadays services are more ubiquitous then ever due to the fact that most services are delivered via smartphones and/or as smartphone applications. Due to the ubiquity of services and their extension to mobile devices, user privacy is exposed even further.

Payment information, such as credit card numbers and similar, has always been identified as sensitive information that needs to be secured in some manner. However, other user specific parameters are often neglected. Perhaps the most exposed parameter is user location, since precise user location is available from mobile device integrated GPS (Global Positioning System) unit and various services tend to use it in order to provide location profiled data to end users [1].

When examining services, it can be determined that most of them actually do not require high precision location data. In this sense, we propose location proxy that can be used for generalizing high precision location data for telecommunication services. By using the proposed location proxy, users do not have to disclose their precise location and thus do not have to completely deny their privacy, at least regarding location data.

The rest of the paper is organized as follows. Next section comments European Union electronic privacy act and guidelines provided in context of location data privacy. Third section proposes service classes regarding location precision and describes the location proxy functionalities. Finally, last section concludes the paper and gives guidelines for future work.

II. USER PRIVACY IN TELECOMMUNICATION SERVICES

European Union regulates user privacy in the current e- Privacy Directive [2]. However, the current directive is challenged by new technologies that bring new aspects of possible privacy breach. For example, the most media attention to this matter was brought by Google Street View, when people stated that their privacy was offended when Google took pictures for their popular street view service. In order to protect user privacy in the context of emerging technologies and services, new guidelines are proposed within General Data Protection Regulation [3] that should resolve current issues.

Lawfulness usage of location data for LBS, considering e-Privacy Directive Article 9 (subsections 1, 2, 3) entails that location data may only be processed with the permission of the users to the extent and for the duration necessary for the provision of services, or while they are formulated anonymously.

The same directive stated that location data processing not only depends on the location data and on the type of service to be provided, but also on whether it is possible (technically and procedurally) to distinguish the individual using an service from the legal or natural person having subscribed to it.

Moreover, service provider must inform the user (prior to requesting permission) about the type of location data that will be processed, the objectives and duration of the processing, and whether the location data will be transferred to a third party. Users or subscribers shall be given the possibility to withdraw their consent for the processing of location data other than traffic data at any time.

On the 15 December 2015, the European Parliament, the Council and the Commission reached agreement on

the new data protection rules, establishing a modern and harmonized data protection framework across the EU. The European Parliament's Civil Liberties committee and the Permanent Representatives Committee (Coreper) of the Council then approved the agreements with very large majorities. The agreements were also welcomed by the European Council of 17-18 December as a major step forward in the implementation of the Digital Single Market Strategy. Once the Regulation and the Directive receive formal adoption from the European Parliament and Council, the official texts will be published in the Official Journal of the European Union in all official languages. The new rules will become applicable two years thereafter.

The new rules will ensure that user receives clear and understandable information when his personal data is processed. Whenever his consent is required, it will have to be given explicitly before a company could process his personal data. The European Commission will also strengthen individuals' right to be forgotten, which means that if user no longer wants his data to be processed, and there is no legitimate reason for a company to keep it, the data shall be deleted.

To follow lawful use of location data there are several activities identified by e-Privacy Directive that should be considered separately and each one with particular attention.

The new data protection framework also includes rules to ensure a common level of data protection in the area of cooperation between police and justice authorities across Europe, which will facilitate exchanges of data to prevent and investigate cross-border crime. This will increase citizens' security; hand-in-hand with a guarantee that the protection of their personal data is safeguarded. Data Protection Authorities will work more closely together in the future, especially through the one-stop shop mechanism to solve cross-border data protection cases.

Service provider must inform the users about the risk of a breach of the security of the network and, where the risk lies outside the scope of the actions to be taken by the service provider, i.e. obligation to inform the users.

For normal purposes duration of location data processing may continue only for the duration necessary for the service. On the other hand, selected data need to be stored by service and network providers for special purposes (investigation, detection and prosecution of serious crime, etc.) depending on national laws. Regardless of that ostensible conflict, for ordinary user location data retention is not allowed longer than it is necessary for the service.

To provide data security, service providers need to protect location data in location systems, particularly location data corresponding to individuals such as users. Service provider must take technical operations and organizational procedures to protect the security of its services in conjunction with the provider of the public communications network.

Further, location data processing restriction is limited to personnel of the provider of the public network or service provider, and processing must be limited to what is strictly necessary to provide the service. This restriction for location data processing prohibits to the provider of a service to process location data for promotion of its services.



Figure 1. Example location data types

From the user point of the view location information has a variable meaning. Meaning of location for a user depends on the actual situation and user intentions. For certain services it is important to discover the precise location data or direction of movement, while for others this precision is not significant.

For example, the precise location is required when someone need assistance in danger and as soon as possible. On the other hand, there are situations and services for which the accuracy of location is not important or even not tolerable. For example, the location of a town, province, region or state, and it is important to know the change - user comes in or user moves toward or leave of town, or some other events will notify the user if he is in that town, otherwise not.

Therefore, the applicable solution that meets previously mentioned e-Privacy directive points out an intelligent component allocated for each specific service. The purpose of that component is formatting of location data in form that is requested only for known type of the service. That form excludes the possibility that anybody could find exact user location.

III. LOCATION PROXY

The purpose of the proposed location proxy is to transform more precise location data into less precise location data appropriate for the given service.

The precision of initial data that is transformed depends on technology that provides it. User location data can be provided by user's mobile device (GPS, cell identification, WLAN triangulation) or by mobile network infrastructure (cell identification, cell segmentation, triangulation). Mobile network location is usually available by open network interfaces such as Parlay and similar. However, the growth of smartphone market in recent years makes device centric positioning more interesting and more accessible to application providers/developers and is much more common today.

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

When using mobile device positioning, native application installed on mobile device reads location and uses it within the application or forwards it to service on the Internet for processing. Prior to location data usage in this manner, the user must give permissions for usage location data to the application. However, once the application has these permissions, user's privacy may be at risk, especially if location data is sent toward Internet based service. Once positioned in the Internet, private location data may become a subject of various attacks, whether during transport or while being processed and possibly in storage at service application server.

In order to make this data less alluring for attackers and appropriate enough for the service, initial precision of location data can be transformed to less precision. Since different services require different location precision, we propose several classes of precision. Classes are enumerated from 1 to *9* where 1 represents highest precision, in our context GPS, and *9* lowest precision.

A. Proposed precision classes

Example classification is shown in table 1. Class 1 has the highest precision and uses direct GPS information obtained from the device. As such, it is suitable for routing and navigation applications which request highest location precision. In scope of privacy, most routing and navigation applications on the market keep their executive logic on the user device and do not send location information from the device, so privacy is preserved in this case. Class two can be used for various on site marketing services, e.g. discount offers when user is near retail store, and usually relies on indoor positioning techniques (Bluetooth, sensors, RFID, etc.). Nevertheless, this can also pose privacy threat if user proximity to some beacon is transferred over the Internet and used for purposes other than marketing. Class 3 precision is similar to precision required by today emergency services based on mobile network positioning (e.g. E-911 in United States). However, such services tend to work with more accurate location data in the future with use of newer technologies such as GPS (e.g. eCall [4]).

Class 4 corresponds to cell identification location precision. Since cell identification is the most common mobile network positioning method, there are lot of existing services that use this method and may use the proposed precision class. Various traffic related information, nearest services, stores and similar all rely on cell identification. Furthermore, cell identification is often used by various native applications on mobile devices for positioning users when there is no GPS signal available, e.g. indoors and similar. However, the positioning is not done by mobile network in this case; rather than that, the user is localized based on online GIS databases with geocoded mobile network cell tower positions. Similar to class 4, class 5 also may provide slightly less precise location information and is used for similar applications. Besides that, class 5 precision often corresponds to cell identification in sub-urban, less populated areas.

TABLE 1 – PROPOSED PRECISSION CLASSES

Class	Precision	Example technology (where applicable)	Example application
1	< 10 m	GPS	routing, navigation
2	10 - 50 m	cell tower triangulation, WLAN localization	on-site marketing
3	50 - 100 m	cell segmentation	emergency positioning
4	100 - 200 m (urban areas)	cell identification	traffic information
5	200 - 500 m	BTS cell cluster	"near me"
6	> 500 m	location area	
7	city	-	offers, marketing
8	region	-	weather information
9	country	-	language localization

Classes from 6 to 9 represent rather low location precision that may be used for various location based marketing offers (e.g. within a city), weather reports (e.g. within a region) or language localization based on country identification. Although not as precise as other classes, these classes still enable a high degree of user profiling and should not be ignored. For instance, if a user wants to buy cinema ticket it would be appropriate to present him with an offer of cinemas in his town. Similarly, it would not make sense to market a product that can be bought in different city or even region, etc.

B. Architecture

In order to explain how services acquire user location through the proposed location proxy, system architecture is shown on figure 2. Location proxy mediates between end-users and services available on the Internet. End user location can be read either by native application installed on mobile device or by the mobile network. Regardless of the method of acquisition of user location, if such location is transferred to the service available on the Internet it may be subject to privacy denial. On the other side, service that uses such location information may be implemented either as a cloud service, which is much more common today, or as a standalone service on the web.

Once the user registers for specific service or installs application on his/her device a registration request to the location proxy should be initialized as well. The proxy serves as a central point for location information. Therefore location requests from services and responses (provided by user device application or mobile network) should be handled by the location proxy. In this sense, the proxy should identify services that may request user location and corresponding users for each service. User identification within the proxy is completely anonymous, in a sense that each user (for certain service) is identified either by locally installed application random seed or specific cryptic value derived from user device IMEI (*International Mobile Equipment Identity*) when using mobile network positioning. Location proxy should be able to map anonymous user identification parameters

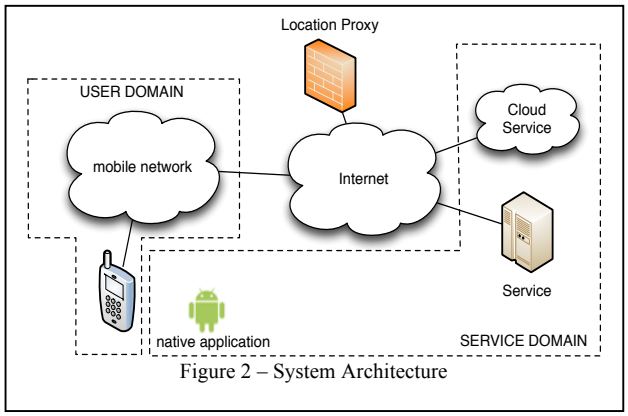


Figure 2 – System Architecture

known by outer service and real user identification, provided either as unique application identifier or by IMEI. Since sensitive data is transferred from user mobile device or mobile network to the location proxy, this connection should be protected with suitable encryption mechanisms and protocols like TLS (*Transport Layer Security*).

The cycle of location acquisition is initiated by the service that sends a location request for certain anonymous user. It needs to be emphasized that user and service communicate other service related data as usual, but only the location is transferred through the proposed proxy. This is important because usually it is the user that triggers service request and the service must know when to request user location data. Once the user initiates a general service request, the service detects that it needs user location and initiates a request to the location proxy with anonymous user identification. Furthermore, the service may know the real user identity, which depends on the service and it's functionalities, but the location proxy should be addressed with anonymous data at all times.

Once location request is made by the service the proxy *identifies* the user and the location class needed for the given service. It then initiates location request either to the mobile network or to the application installed on the user device. The communication between proxy and mobile network can be resolved through open network interfaces by using frameworks such as Parlay X or OneAPI. On the other hand, communication between proxy and mobile device application can be done as push service (where available, depending on the mobile platform) or by standard web service connection through HTTPS.

Once the location is acquired, the proxy transforms it according to the required class of precision and forwards the transformed location to the service that initiated the request.

C. *Location scaling*

Location proxy's main purpose is transforming high precision locations to locations of lesser precision appropriate enough for the service that initiated the request. Such transformation can be done by various methods, ranging from conservative straightforward scaling to machine learning approach that may introduce a degree of intelligence in location scaling process.

The proposed proxy uses self-organizing maps (SOM) for the purpose of location scaling. SOM is a neural network architecture introduced by T. Kohonen [5] that can be used for clustering similar data based on it's features. The data is used as an input to the SOM and should be presented as an array, where each element corresponds to a single neuron in input layer. The value of the location data corresponds to activation of a neuron in the input layer. The problem that emerges in this case is how to represent different locations, i.e. locations of different precisions acquired by different technologies. For example, GPS data is identified by pairs that mark longitude and latitude of certain point. Mobile network cell is uniquely identified by it's network ID, but the question is how to represent that ID to the SOM.

Due to different location representations depending on technology the only option is to train individual SOM network for each specific class, primarily based on the type of location that is used as an input. In this sense the network that uses GPS coordinates as an input has two input neurons, one reflecting latitudes and the other longitudes. For larger geographical areas there should exist additional input neurons in order to precisely differentiate one location from another, as explained in [6].

Regarding other location input types, such as cells, clusters, location areas and similar, the only thing that is obligatory is that each location representation is unique within location proxy and that similar locations have similar input representation. In this sense, each specific location obtained either from the network or from the user mobile device via application should be uniquely presented to SOM used for scaling location for selected type and precision. The representation itself can be done in various ways by dividing unique location identifier into an array of elements that reflect real locations. In order to do so and prior to learning the SOM with location identifiers, the real geographical location layout should be known. Once real layout of locations is known, their representation to the SOM can be adjusted in such a way that locations that are nearby have similar input patterns. It is necessary to emphasize that the learning of the SOM should be done once before the application of scaling. In case of location or topology change, the SOM should be retrained.

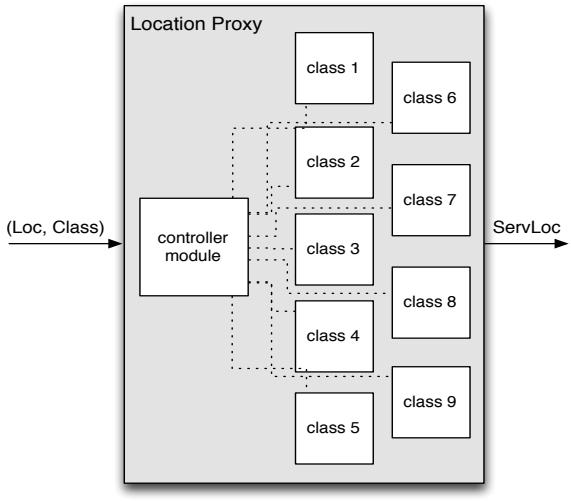


Figure 3 - Location Proxy

Figure 3 shows architecture of the proposed location proxy. Class modules on the right hand side represent self-organizing maps used for actual location scaling. Controller module receives request for location scaling that contains location representation (coordinates, cell identification, WLAN AP or similar) and class of precision required for the service that initiated the request. According to the required class, controller module makes additional formatting of the input data and forwards it to the appropriate class module. Formatting of input data consists of transforming the data to SOM input parameter, as discussed previously. The output of the location proxy is service location identifier *ServLoc* if the input location matched any of the predefined service locations (e.g. airport, bus station, Zagreb center, Zagreb region, Croatia…). If not, the output is marked as not available which means that the user is not located at any locations that may be interesting to the service in question.

IV. CONCLUSION

Privacy in general is an important issue in today interconnected world of various competing services that are trying to attract users by contextualizing service content. In order to do so, services have no option but to invade user privacy. Among other aspects of user privacy that are well addressed, user location often remains neglected at this point because it has not been available before as it is today due to new mobile devices with GPS technology.

The paper presents how telecommunication services handle user locations and questions whether most of the services actually need the location precision they are using. In order to limit user location data usage and thus preserve user privacy to some extent, a location proxy for location generalization is proposed. The proxy's main purpose is to obfuscate more precise user locations in service scenarios where such location precision is not necessary. It serves as a location related central point between end-users and/or their mobile devices and services available on the Internet.

Proposed location proxy uses self-organizing maps for location generalization. These neural network architectures are often used for similar purposes due to their ability of clustering similar data. However, this approach may prove to be too complex if used on large geographical area and simpler yet effective methods will be examined as future work. There is also an option of combining SOMs with other methods in a sense that SOMs can be used for some classes that require more granular differentiation between locations, while conventional methods could be used for other, less demanding classes.

REFERENCES

[1] Ashbrook, D., Starner, T. (2002) Learning significant locations and predicting user movement with GPS. *Proc. of 6 IEEE Intl. Symp. on Wearable Computers*, Seattle, WA.

[2] "Directive on Privacy and Electronic Communications", 2002/58/EC, European Parlament and Council, 2002.

[3] "Proposal for the EU General Data Protection Regulation", European Commission. 25 January 2012.

[4] "Commition recommendation on support for an EU-wide eCall service in electronic communication networks for the transmission of in-vehicle emergency calls based on 112 ('eCalls')", European Commision, 9/2011.

[5] *Self-Organizing Maps* (3rd ed.). T. Kohonen, M. R. Schroeder, and T. S. Huang (Eds.). Springer-Verlag New York, Inc., Secaucus, NJ, USA., 2001.

[6] Vukovic, M.; Illes, S.; Jezic, G.: "Predicting user movement with intelligent GPS location scaling" // 1st GNSS Vulnerabilities and Solutions Conference Conference proceedings, 2008.

Analiza sigurnosnih ranjivosti inteligentnih sučelja za upravljanje podatkovnim centrom

Milan Ramljak, mag.ing.
Ericsson Nikola Tesla (FER), Split, Hrvatska
milan.ramljak@ericsson.com

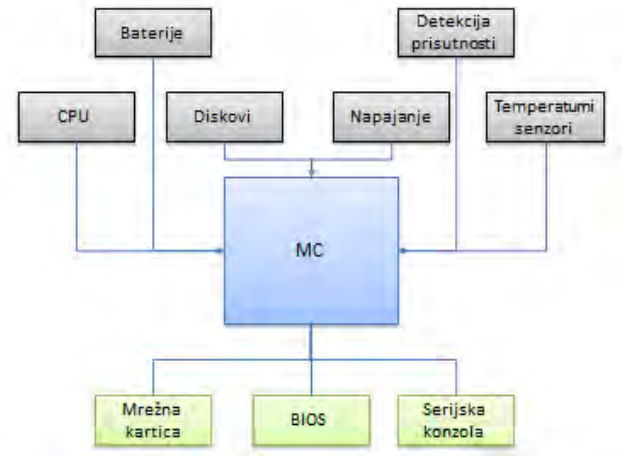
Sažetak - Inteligentno sučelje za upravljanje platformom (IPMI) iako razvijeno od strane i danas velikih kompanija, predvodnika u informacijskom svijetu, ne bi se trebao ponositi svojim imenom. Razvijen prije gotovo dvadeset godina, protokol i danas sadrži neke od velikih sigurnosnih propusta koji će biti analizirani dalje u članku. IPMI protokol standardizira komunikaciju između servera različitih proizvođača na način da komuniciraju na isti način bez obzira na samog proizvođača. Sama činjenica i razlog razvoja protokola je vidljiva iz perspektive velikih informacijski centara gdje razna oprema cijelo vrijeme mora biti dostupna i upravljiva. Upravo u tome i leži glavna veličina sigurnosnih nedostataka IPMI protokola. Iako razne kompanije troše velike količine sredstava da bi podaci i servisi na kojima počiva njihovo poslovno djelovanje bili sigurni, dalje u članku će se vidjeti koliko je ta sigurnost neosnovana i krhka. Kao temelj ovog članka će se prikazati rezultati analize većine sigurnosnih nedostataka IPMI implementacije na modernom hardveru danas dostupnom na tržištu kao i usporedbe rezultata među proizvođačima. Iako moderni sustavi, vrijedni stotine tisuća dolara, rezultati su poražavajući.

I. UVOD

Inteligentno sučelje za upravljanje platformom (Intelligent Platform Management Interface - IPMI) razvijeno je kao rješenje koje pruža mogućnost unificiranog upravljanja poslužiteljima na daljinu, bez potrebe za fizičkim pristupom sklopovlju. IPMI arhitektura opisuje skup sučelja autonomnog računalnog podsustava, tzv. upravljačkog kontrolera (Base Management Controller - MC), koji omogućava kontrolu i praćenje stanja sustava neovisno o matičnom procesoru računala i operacijskom sustavu (OS) koji se vrti na njemu [1]. Upravljački kontroler je inteligentni dio arhitekture, a predstavlja mikrokontroler koji je ugrađen na matičnoj ploči računala te ima potpuno odvojen processor i memoriju. Mikrokontroler i matični procesor su povezani unutarnjim serijskim sabirnicama pa je upravo zbog toga IPMI sustav u mogućnosti potpuno upravljati kompletnim računalom [2]. Mnogo je razloga zašto administratori poslužitelja koristite IPMI a neki od njih su:

- praćenje statusa platforme (temperatura sustava, napon, struja, brzina ventilatora)
- pregled sklopovskih zapisa (kritične temperature)
- obavljanje procedure za oporavak (engl. disaster recovery procedures)

Slika 1 prikazuje razne senzore temperature, prisutnosti diskova, mjerače struje i napona koji su povezani na MC. Često MC ima i serijski konzolni priključak koji je jako važan kod instalacije nove verzije firmvera.



Slika 1. Blok shema elemenata sustava upravljačkog kontrolera

Glavna odlika IPMI sustava je činjenica da promatrani sustav može biti ugašen a istodobno upravljiv sve dok je spojen na izvor napajanja i lokalnu mrežu[3]. IPMI funkcionira bez obzira u kojem je stanju matični procesor:

- prije pokretanja operacijskog sustava
- kod pada sustava
- kada je matični procesor ugašen

Sve prethodno navedene funkcionalnosti razlog su korištenja IPMI sučelja u svim modernim distribuiranim informacijskim sustavima. Međutim, s velikim mogućnostima dolaze i velike odgovornosti što se u informacijskim sustavima manifestira kroz sigurnosne aspekte. Od uvođenja prve IPMI specifikacije, prije nešto više od petnaest godina, IPMI je doživio više revizija [4] koje su osim dodavanja funkcionalnosti velikim dijelom bile i dorade sigurnosne funkcionalnosti.

Vrlo je često mišljenje da upravljački kontroleri imaju jako malo mogućnosti, što i je istina za prve implementacije IPMI specifikacije, međutim današnji sustavi su vrlo jaki što proizvođačima daje mnogo prostora za proširenjem funkcionalnosti MC-a. Upravo ova činjenica povećava potrebu za sviješću koliko je sigurnosni model IPMI specifikacije vrlo slab.

II. SIGURNOSNI MODEL

IPMI specifikacija definira pristup temeljen na ulogama (engl. roles) administrator, operator i korisnik (engl. user) [5]. Pojedina uloga ograničava način pristupa pojedinim resursima sustava. Administratorska uloga predstavlja najvišu sigurnosnu ulogu i ima potpunu kontrolu nad upravljačkim kontrolerom.

U tablici I je prikazana matrica različitih sigurnosnih postavki u odnosu na pojedinu korisničku ulogu a prikazuje što pojedina uloga može raditi. Vidljivo je da korisnik sa administratorskom ulogom ima mogućnost konfiguriranja svih sigurnosnih postavki upravljačkog kontrolera. IPMI verzije 2.0 ima mogućnost autentifikacije korisnika putem LDAP imenika, ali postavke vezane uz LDAP može mijenjati samo administrator dok ostali korisnici mogu samo čitati. Isto pravilo vrijedi i za SSL postavke koje imaju MC-ovi sa Web sučeljem.

TABLICA I. MATRICA ULOGA I RESURSA

Resurs	administrator	operator	korisnik
Korisničke postavke	RW[a]	R	X
SSL postavke	RW	R	X
LDAP postavke	RW	R	X
BIOS	RW	R	R

a. R=Čitanje, W=Pisanje, X=nema prava

Promatrajući mogućnosti definirane u tablici I, postavlja se pitanje kako ovakav sustav može biti nesigurniji od bilo kojeg drugog sustava koji definira pristup temeljen na ulogama. Odgovor se krije u samoj povijesti razvoja protokola IPMI. U vremenu definicije protokola, resursi računalnih sustavi su bili relativno ograničeni pa samim time sigurnosni aspekti su većim dijelom zanemareni. Brz razvoj snage računala zadnjih godina i težnja servisima u oblaku (engl. Cloud Services) još više stvara potrebu za mogućnostima koje pruža IPMI što posebice dolazi do izražaju kod podatkovnih centara (engl. Data Centers).

Razni proizvođači računala i platformi za podatkovne centre baziraju svoja rješenja na specifikaciji protokola IPMI, što rezultira sa velikim brojem implementacija. Neki od primjera IPMI implementacija su:

- HP iLO [5]
- DELL iDRAC [6]
- AMI MegaRac [7]
- IBM IMM2 [8]

A. IPMI implementacije

Prema specifikaciji postoje tri dostupne verzije IPMI protokola, 1.0, 1.5 i 2.0 od kojih je samo verzija 1.0 gotovo uklonjena sa većih računalnih mreža. Ozbiljnija sigurnosna proširenja su započela sa verzijom 2.0 koja je doživjela pet revizija od kojih je zadnja bila prije nešto više od godinu dana.

IPMI verzija 1.5 osim sažimanja (engl. hash) lozinke, koje je opcionalno ne i obavezno, nema nikakve kriptografske zaštite komunikacije klijenta i upravljačkog kontrolera. Specifikacija definira protokol RMCP (engl. Remote Management Control Protocol) baziran na protokolu UDP [9]. Kao rezultat toga komunikacija je izložena većini sigurnosnih napada, kao što su otimanje veze (engl. Session Hijacking) i napad čovjeka u sredini (engl. Man-in-the-middle). Isto tako kada se postavlja ili mijenja korisnička lozinka nova lozinka se šalje preko mreže u čistom (nezaštićenom) obliku. Možda se na prvi pogled činilo kao napredniji protokol, ali slobodno se može reći da sigurnost verzije 1.5 nije ništa bolja od one koju pruža telnet, dakle nema je [10].

U verziji 2.0 podržani su razni algoritmi za zaštititu povjerljivosti, čak petnaest različitih [4], koje može birati udaljeni korisnik prilikom pristupa. Standard definira i proširenje u listi algoritama što omogućava proizvođačima implementaciju vlastitog algoritma za zaštitu povjerljivosti. Tablica II prikazuje neke podržane algoritme. Korisnici nisu u mogućnosti kombinirati algoritme zaštite integriteta i povjerljivosti prilikom pristupa jer su oni definirani specifikacijom, već samo postaviti način zaštite odabirom identifikacijskog broja označenog u prvom stupcu tablice.

TABLICA II. MATRICA ALGORITAMA

id	autentifikacija	integritet	povjerljivost
0	Nema	Nema	Nema
1	HMAC-SHA1	Nema	Nema
2	HMAC-SHA1	HMAC-SHA1-96	Nema
5	HMAC-MD5	HMAC-MD5-128	Nema
17	HMAC-SHA256	HMAC-SHA256-128	AES-CBC-128

IPMI verzija 2.0 uvodi niz kriptografskih protokola koji trebaju poboljšati sigurnost, ali nažalost dodaje nekoliko ozbiljnih sigurnosnih propusta koji rezultiraju sa još više ranjivosti. Iako postoje neke pristojne mogućnosti zaštite u verziji 2.0 većina sustava i dalje koristiti staru verziju 1.5 kao tvorničku postavku.

B. Sigurnosni propusti

Vrlo je teško vjerovati da osim preformansi postoji dobar razlog pružanja mogućnosti odabira autentifikacijskog algoritma bez šifriranja, dakle bez zaštite povjerljivosti uopće. To stvara situacije u kojima poslani podaci imaju zaštitu integriteta, ovjereni su kao nezaštićen tekst, ali bez šifriranja [10].

Razne su mogućnosti koje pridonose poboljšanju sigurnosti IPMI komunikacije, ali svakako glavni problem je u algoritmu koji nema šifriranje.To je u suštini protokol, bez provjere autentičnosti, integriteta i povjerljivosti. Za autentifikaciju korisnika sa metodom bez šifriranja može se koristiti bilo koja lozinka, jer je protokol zanemaruje

[4]. Važeći korisnički račun je potreban za provjeru autentičnosti, ali gotovo svaka lozinka će biti prihvaćena. Većina poslužitelja ima ovu metodu omogućenu prema tvorničkim postavkama, a neki proizvođači čak ne dopuštaju njeno deaktiviranje.

Još je jedan veliki sigurnosni problem predstavljen u verziji 2.0 a rezultat je pokušaja uvođenja autentifikacije djeljenjem ključeva (engl. negotiation) kod uspostavljanja sigurne veze. Uvodi se RAKP (RMCP+ Key Exchange Protocol), koji IPMI koristi kod pregovaranja pri uspostavi sigurne komunikacije [11]. Međutim RAKP protokol omogućuje da anonimni korisnik (daljinski) dobije sažetak lozinke upravljačkog kontolera (MC). Za mnoge je to ništa drugo već nevjerojatno loš dizajn, jer omogućuje napadaču lokalno pogađanje lozinke (engl. offline) i upad u sustav. Prema IPMI specifikaciji ovaj slučaj se ne identificira kao nedostatak već kao funkcionalnost koja otvara vrata vrlo jednostavnom napadu za koji ne postoji rješenje [12]. Naime, IPMI implementacija u slučaju neuspjele autentifikacije vraća korisniku (ili napadaču) identifikacijski broj samog MC-a poznatiji pod nazivom globalno jedinstven identifikacijski broj (engl. Globally Unique Identifier - GUID). GUID se koristi u podatkovnim centrima za identifikaciju elementa velikog sustava.

Spomenuti problem napadači koriste tako da šalju upite MC-u i dobivene odgovore koriste u programima za pogađanje šifri. Da bi ovaj napad bio izvediv potrebno je da napadač zna ili pogodi korisničko ime koje postoji na sustavu. Ovo nas uvodi u treći sigurnosni propust koji nije vezan uz samu IPMI specifikaciju već uz način uobičajnog korištenja sustava. Većina proizvođača prilikom proizvodnje IPMI sustava definira tvorničkog korisnika i njemu pridružuju tvorničku lozinku. Takav korisnik ima najviše privilegije. Većina proizvođača zadržava iste vrijednosti i kod potpuno nove verzije svojih proizvoda. Pretragom na Internetu [13] lako je pronaći listu nekih od tvorničkih lozinki za pojedine proizvođače. Tablica III sadrži listu najčešćih lozinki korištenih na raznim IPMI implementacijama dostupnim na tržištu.

TABLICA III. ČESTO KORIŠTENE LOZINKE

	lozinka	**lozinka**
0	ADMIN	computer
1	admin	changeme
2	PASSWoRD	4rfv$RFV
3	calvin	password
4	root	123456
5	superuser	

Osnovna IPMI autentifikacija se obavlja preko malog skupa (obično od 10-16 , iako specifikacija definira čak 63) lokalnih IPMI korisničkih računa, kojima se može dati lozinka do 16 (verzija 1.5) ili 20 (verzija 2.0) znakova. Kod izrade lokalnih korisničkih računa ne obavlja se provjera ulaznih podataka što rezultira da korisničko ime može biti i prazan string (engl. null).

C. Komunikacija sa upravljačkim kontrolerom

Četiri su glavna načina pristupa modernim upravljačkim kontrolerima:

- interaktivni terminal (SSH, telnet)
- kernel sučelje (operacijski sustav matičnog procesora putem kernel-a može komunicirati sa upravljačkim kontrolerom izravno preko sistemskih sabirnica)
- Web sučelje (HTTP, HTTPS)
- razne aplikacije (ipmitool, open ipmi)

Rijetki poslužitelji dolaze sa isključenim IPMI sustavom prema tvorničkim postavkama. U suprotnom sustave je potrebno preko nekog od drugih sučelja kao što je BIOS uključiti što ne zahtjeva dodatnu konfiguraciju ili instalaciju. Mnogi poslužitelji sa IPMI podrškom imaju odvojeni mrežni priključak (ethernet) koji se koristi za fizičko odvajanje od ostatka računala, ali postoje i rješenja koja dijele isti mrežni priključak.

IPMI standard specificira više sučelja (medija) kojima se može komunicirati kao što su LAN mrežno sučelje, sučelje unutarnje sabirnice, serijske linije, VLAN. Osim sučelja definirani su i kanali koji mogu usmjeravati komunikaciju između pojedinih sučelja. Bitna karakteristika svih kanala je da razmjenjuju informacije vezane uz autentifikaciju, konfiguracije i uloge korisnika. Kanali se identificiraju jedinstvenim brojem, koji se koristi kada se vrši interno usmjeravanje komunikacije između sučelja. IPMI verzija 1.5 ima samo 9 kanala, dok verzija 2.0 ima 14. Svaki kanal je potpuno neovisan od drugih i može raditi preko istog ili različitih medija.

Protokol IPMI koristi samo UDP port 623 za komunikaciju, ali MC općenito pruža niz drugih mrežnih usluga na različitim portovima. Među uobičajenim uslugama tu je:

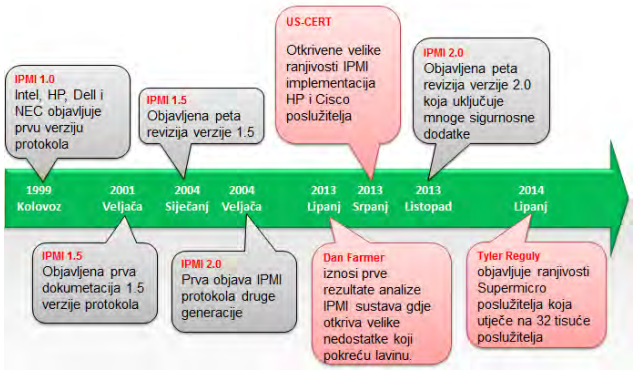
- web (HTTP i HTTPS)
- SSH, telnet
- SMTP, SNMP, DHCP
- KVM (engl. Keyboard Video Mouse)

Kod većine proizvođača udaljena komunikacija s MC-om nije tvornički omogućena, te je potrebno eksplicitno omogućiti promjenom konfiguracije. MC kao glavna kontrolna točka računala može otkriti sva mrežna sučelja na računalu, ali općenito je onemogućeno slušanje prometa poslužitelja (promet koji generira operacijski sustav).

D. Pregled objavljenih ranjivosti

U ovom poglavlju se prikazuju neke od objavljenih ranjivosti poslužitelja raznih proizvođača. Slika 3 prikazuje vremenski raspon ključnih trenutaka u razvoju i analizi IPMI standarda. Iako je IPMI standard definiram pred 15-ak godina, većina sigurnosnih nedostataka je otkrivena u vrlo kratkom vremenu prije dvije godine (Lipanj i Srpanj 2013. godine). Glavnu zaslugu u otkrivanju većine problema ima Američki sigurnosni

stručnjak Dan Farmer [14] koji je ujedno i bio pionir u otkrivanju nedostataka IPMI standarda.



Slika 2. Pregled IPMI ranjivosti kroz povijest

U srpnju 2013. godine Američka vlada je otkrila niz problema IPMI implementacije Cisco UCS poslužitelja serije B i C [15]. Ranjivosti su objavljene u Američkoj bazi računalnih ranjivosti US-CERT pod identifikacijskim brojem TA13-207A [16]. Neki od nedostataka su vezani izravno na IPMI standard, kao što je ranije opisana mogućnost otkrivanja sažetka lozinke.

Druge ranjivosti vezane uz Cisco implementaciju:

- nedostatak mehanizma za stvaranje zapisa kod rušenja elemenata stoga (engl. stack dump) koji sadržavaju IPMI korisničke podatke u nezaštićenom obliku
- otkrivanje jedne IPMI lozinke omogućava pristup svim poslužiteljima unutar iste domene
- na Cisco MC-ovima je otkriven veliki broj servisa bez ikakve zaštite pristupa (HTTP, telnet, KVM)

U istom mjesecu otkrivena je i ranjivost HP poslužitelja, točnije HP-ova implementacija MC-a (Integrated Lights Out - iLO) koja dopušta napadaču izvršavanje IPMI komandi bez kontrole pristupa. Ova ranjivost je označena kao sigurnosni problem najvećeg mogućeg prioriteta, jer omogućava potpuni pristup poslužiteljima bez ikakve kontrole pristupa. Ranjivost je evidentirana u bazi zajedničkih ranjivosti i sigurnosnih nedostataka (engl. Common Vulnerabilities and Exposures) pod identifikacijskim brojem CVE-2013-4784 [17].

U lipnju 2014. godine Tyler Reguly objavljuje ranjivost vrlo visokog prioriteta IPMI implementacije Supermicro poslužitelja [18]. Supermicro je Američki brend poslužitelja koji se vrlo često koristi zbog svoje dostupnosti i niske cijene. Ranjivost se odnosila na propust u implementaciji Web servisa koji je omogućavao neautentificiranom napadaču pristup datoteci PSBlock putem jednostavnog HTTP upita. Rezultat upita je prikazao sve IPMI korisničke lozinke u nezaštićenom obliku. Daljnjom analizom je utvrđeno da se na Internetu nalazi više od 31 tisuću Supermicro poslužitelja izloženih opisanoj ranjivosti [19] čemu je pridonijela činjenica velike korištenosti Supermicro poslužitelja.

III. ANALIZA RANJIVOSTI MODELA PODATKOVNOG CENTRA

U ovome poglavlju će se na modelu prijetnje prikazati rezutati analize sigurnosnih propusta IPMI implementacije na podatkovnom centru sastavljenom od poslužitelja triju velikih kompanija, HP, Dell i American Megatrends Inc (AMI).

A. Identifikacija resursa podatkovnog centra

U svrhu izrade modela prijetnje korišten je podatkovni centar sastavljen od više različitih varijanti poslužitelja. U praksi nije rijedak slučaj povezivanja poslužitelja različitih proizvođača unutar istog segmenta podatkovnog centra te je u svrhu izrade analize takav podatkovni centar konfiguriran. Osnovne specifikacije korištenih poslužitelja su prikazane u tablici IV.

TABLICA IV. SPECIFIKACIJE POSLUŽITELJA

Vrsta	Datum proizvodnje	Firmver verzija	CPU
HP DL360 [20]	03/2013	iLo G7	Intel Xeon X5690
Dell R630 [21]	06/2015	iDrac 8	Intel Xeon E5 2600
Dell R730 [22]	05/2015	iDrac 8	Intel Xeon E5 2680
AMI MegaRAC [23]	05/2014	0.4	ugradbeni procesor

Promatrajući specifikacije navedenih poslužitelja lako je vidljivo da se radi o jako snažnim računalima koji se koriste kao gradivi elementi današnjih podatkovnih centara. Svaki poslužitelj je zasebna jedinica i može se prikazati kao minimalni podatkovni centar, međutim takav pristup se rijetko koristi zbog teškog upravljanja velikim brojem zasebnih jedinica.

U praksi se više poslužitelja povezuje u zajednički sustav koji se preko korisničkih upravljačkih programa višim slojevima prikazuje kao virtualni podatkovni centar. Upravljački programi se uvijek izvršavaju na zasebnom poslužitelju koji potpuno kontrolira svaki od poslužitelja podatkovnog centra. Upravljački program se često izvršava kao Web servis koji omogućava konfiguriranje i upravljanje elemenata podatkovnog sustava.

B. Arhitektura podatkovnog centra

Glavni cilj ovog koraka je identifikacija arhitekture, tehnologija i funkcija podatkovnog centra koji je korišten za analizu ranjivosti.

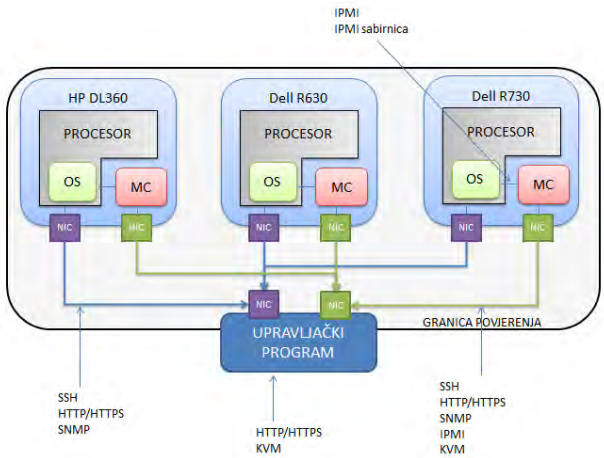
Slika 4 prikazuje blok shemu podatkovnog centra. Podatkovni centar se sastoji od tri poslužitelja fizički razmještena u istom kabinetu. Svaki poslužitelj je prikazan kao zasebna jedinica povezana putem upravljačkog programa u jedinstvenu cjelinu. Upravljački program predstavlja granicu povjerenja u kojoj se nalaze sva tri poslužitelja koja si međusobno vjeruju.

Upravljački program se razlikuje kod svakog pružatelja usluga, ali u načelu svi rade istu funkciju povezivanja i segmentiranja jedinica podatkovnih centara. Tipična implementacija upravljačkih programa je u obliku Web servisa (HTTP programsko sučelje - API) koje vlasnicima podatkovnih centara daje jednostavan način za proširenje umrežavanjem dodatnih poslužitelja. Vrlo je važno pravilno konfigurirati poslužitelj koji izvršava upravljački program jer u protivnom granica povjerenja se potpuno gubi.

U analiziranom sustavu su identificirane četiri vrste protoka podataka kroz podatkovni centar. Prva vrsta protoka podataka je preko IPMI sabirnica koje interno povezuju MC i matični procesor. Ova komunikacija je potpuno nezaštićena, ali zahtjeva fizički pristup samom procesoru ukoliko se pokuša iskoristiti te se u načelu smatra internom granicom povjerenja procesora i MC-a. Komunikacija putem IPMI sabirnice je korištena kada matični procesor, odnosno operacijski sustav, želi komunicirati sa MC-om putem IPMI kernel sučelja. Kod nekih jeftinijih poslužitelja ovo je jedina moguća veza sa MC-om a koristi se u rješenjima koja nemaju zasebnu mrežnu karticu na MC-u.

Drugi protok podataka se obavlja preko operacijskog sustava koji se izvodi na matičnom procesoru, a predstavlja uobičajan skup protokola koji se koriste kao SSH, SNMP i HTTP.

Treći protok podataka prolazi kroz mrežnu karticu MC-a. Osim često korištenih sučelja vrlo je važno navesti tipkovnica, video i miš (engl. Keyboard Video Mouse - KVM) podršku koju pruža MC. KVM sučelje se u svojim začetcima vezalo uz sklopovlje koje je omogućavalo upravljanje poslužiteljima kod procedura oporavljanja, međutim razvojem podatkovnih centara pojavila se potreba za virtualizacijom KVM sklopovlja. Bitno je napomenuti da KVM sučelja stvaraju vrlo velike izazove kod izrade sigurnih podatkovnih centara čemu pridonosi činjenica što je virtualni KVM često izveden kao Java program.



Slika 3. Blok shema arhitekture podatkovnog centra

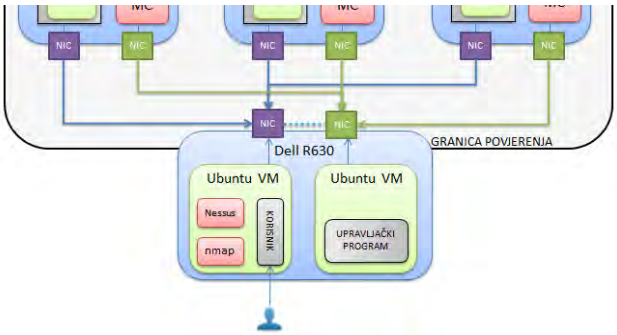
Konačno, posljednji protok podataka je onaj koji generira upravljački program ili operater podatkovnog centra te ujedno predstavlja i ulaznu točku sustava. U idealnim slučajevima protokoli koje pruža upravljački program su jedini koje vanjski korisnik može otkriti skeniranjem. Svaki pokušaj pristupa nekom od poslužitelja unutar granice povjerenja ide isključivo preko upravljačkog programa, naravno ako se napadač ne nalazi u domeni poslužitelja koji izvršava taj program. Ulazna točka mora biti osigurana čuvarima (engl. gatekeepers) koji osiguravaju autorizaciju i provjeru valjanosti.

Upravljački programi se u praksi izvršavaju na zasebnom poslužitelju koji ima samo tu ulogu a funkcionalno djeluje kao vatrozid (engl. firewall). Glavna odlika takvog poslužitelja je vrlo visoka razina dostupnosti (engl. availability).

Funkcionalnost upravljačkog programa se razlikuje između proizvođača kao i tehnologija u kojem je razvijen. Često korištene tehnologije su Java i Python koje se koriste kao podloga za izradu grafičkih i programskih Web sučelja.

C. *Identifikacija prijetnji*

Poslužitelj korišten za analizu je podijeljen na dvije virtualne jedinice sa zasebnim operacijskim sustavima Linux Ubuntu 14.04 [27] što je prikazano na slici 5. Jedna virtualna jedinica izvršava upravljački program, dok druga predstavlja ulaznu točku za napadača. U daljnjoj analizi se kreće od pretpostavke da je napadač nekom od tehnika došao do neovlaštenog pristupa korisničkoj virtualnoj jedinici.



Slika 4. Detalj podatkovnog centra sa aspekta napadača

Napadač započinje prikupljanje informacija s ciljem razrade stabla napada. U tu svrhu se koriste mnogi alati koji su dostupni na većini Linux distribucija:

- nmap (alat za skeniranje mreže)
- tcpdump (alat za monitoriranje prometa na mrežnim karticama)
- ping (alat za provjeru dostupnosti računala)

Od ostalih alata korišten je Nessus, komercijalni alat za skeniranje ranjivosti sustava na temelju baze poznatih ranjivosti. Ovaj alat je praktički postao standard u industriji te se često koristi kao zadnji korak u analizi ranjivosti sustava, ali prije nego sustav odlazi u funkciju. Kod sustava čije se funkcionalnosti mjenjaju vrlo je važno vršiti skeniranje periodički kako bi se pratilo da li nova funkcionalnost umanjuje sigurnost sustava.

Tablica V predstavlja važnije rezultate skeniranja poslužitelja podatkovnog centra Nessus alatom. Vidljivo je da su otkrivene neke od ranjivosti visokog prioriteta.

Svi otkriveni problemi trebaju biti uklonjeni kako bi sustav bio što sigurniji. Neki od otkrivenih problema, kao otkrivanje sažetka lozinke, se ne mogu otkloniti jer zahtjevaju intervencije na IPMI 2.0 standardu.

TABLICA V. NESSUS SCAN POSLUŽITELJA

Vrsta	Prioritet	Opis ranjivosti
HP DL360	Visoki 7.8/10	Otkrivanje IPMI 2.0 sažetka lozinke
	Srednji 6.4/10	SSL certifikatu se ne može vjerovati
Dell R630	Visoki 7.8/10	Otkrivanje IPMI 2.0 sažetka lozinke
	Visoki 7.5/10	Tvornički zadani SNMP naziv zajednice (javno)
	Srednji 6.4/10	SSL certifikatu se ne može vjerovati
	Srednji 6.4/10	SSL certifikat samopotpisan
Dell R730	Visoki 7.8/10	Otkrivanje IPMI 2.0 sažetka lozinke
	Visoki 7.5/10	Tvornički zadani SNMP naziv zajednice (javno)
	Srednji 6.4/10	SSL certifikatu se ne može vjerovati
	Srednji 6.4/10	SSL certifikat samopotpisan

U tablici VI su prikazani rezultati nmap skeniranja svih poslužitelja podatkovnog centra. Vrlo je važno uočiti da svaki MC ima niz usluga koje predstavljaju sučelja za upravljanje od kojih je za nastavak članka najvažniji IPMI servis detektiran na portu 623.

TABLICA VI. NMAP SCAN POSLUŽITELJA

Vrsta	Port Protokol	Servis
HP DL360	22 / tcp	ssh
	80 / tcp	http
	443 / tcp	https
	623 / udp	ipmi
Dell R630 Dell R730	22 / tcp	ssh
	80 / tcp	http
	443 / tcp	https
	623 / udp	ipmi
	5900 / tcp	vnc
AMI MegaRAC	22 / tcp	ssh
	80 / tcp	http
	443 / tcp	https
	623 / udp	ipmi
	4000 / tcp	telnet

Svaki MC ima podršku za HTTP i HTTPS što je svakako nepotrebno te je dovoljno i preporučljivo aktivirati samo HTTPS. Od ostalih servisa, da bi se poslužitelj adekvatno osigurao, preporučljivo je isključiti sve servise osim SSH. Inicijalne verzije MC-ova su imale uključen samo port 623 što ujedno predstavlja temelj komunikacije protokola IPMI. Nažalost, i najmoderniji sustavi također moraju imati podršku za stare implementacije (engl. legacy) kako bi se nesmetano moglo obavljati zamjene starih poslužitelja sa novim. Upravo ta činjenica predstavlja osnovu analize ranjivosti ovog članka.

Uspoređujući rezultate provedene skeniranjem podatkovnog centra sa listom poznatih ranjivosti vidljivo je da skeniranje nije otkrilo sve poznate ranjivosti IPMI sustava. Jedina ranjivost koju potkrijepljuje rezultat Nessus skeniranja je mogućnost otkrivanja sažetka lozinke kod sustava koji implementiraju IPMI verziju 2.0.

Iako navedeno, uskraćivanje usluge na sustavima koji implementiraju IPMI se neće posebno analizirati po pojedinom modelu jer se provodi trivijalno, osobito ako je šifriranje uključeno. Uključenjem šifriranje MC-ovi znatno degradiraju performanse a posebno ako se izlože napadu s puno prometa ili procesiranja. Svi testirani sustavi se vrlo lako sruše čak i kada se šalje velik broj ispravnih upita. Kao rezultat napada uskraćivanjem usluge neki udaljeni klijent može dobiti neispravan odgovor zbog kašnjenja paketa odgovora ako se naredbe šalju vrlo visokom frekvencijom. Kao i većini drugih sustava sa znatno boljim performansama, vrlo se teško obraniti od ovakve vrste napada, što naročito dolazi do izražaja kada su sustavi jako ograničeni memorijom i procesorskom snagom.

Napad počinje tako što napadač započinje skeniranjem mreže sa korisničke virtualne jedinice u svrhu otkrivanja servisa koje pojedini poslužitelj pruža. Cilj je otkriti da li neki od skeniranih računala ima otvoren port koji koristi protokol IPMI. Na poslužitelje kod kojih je otkriven IPMI port napadač stavlja fokus kako bi izvršio korake napada treće razine. Na trećoj razini se definira više mogućih puteva koji vode prema cilju, a temelje se na otkrivanju sažetka lozinke IPMI korisničkog profila. Otkriveni sažetak se koristi kao ulazna vrijednost kod alata za otkrivanje lozinke. Većina takvih alata koriste bazu poznatih i često korištenih lozinki kod otkrivanja.

Osim otkrivanja sažetka lozinke IPMI korisnika, stablo napada definira i granu koja opisuje situaciju kada napadač osluškuje mrežu te pokušava otkriti IPMI korisničke podatke. Kako se vidi u prethodnom poglavlju, IPMI specifikacija podržava i autentifikaciju korisnika bez šifriranja što znači da se lozinke šalju mrežom kao čisti tekst. Upravo ovaj propust IPMI specifikacije omogućuje napadaču izvlačenje korisničkih podataka direktno s mreže.

D. Rezultati analize prijetnji

HP poslužitelj nije ranjiv na autentifikaciju bez šifriranja. Tvornička postavka koju sustav ima koristi algoritme koji pružaju autentifikaciju (hmac-sha1), integritet (hmac-sha1-96) i povjerljivost (aes-cbc-128).

Dell R630 iako podržava razne algoritme za osiguranje inegriteta i povjerljivosti je ranjiv na autentifikaciju bez šifriranja. Ovaj problem proizlazi iz tvorničke postavke, koja se može promjeniti.

Dell R730 poslužitelj nije ranjiv na autentifikaciju bez algoritma šifriranja. Tvornička postavka koju sustav ima koristi algoritme kao i HP poslužitelj. Dodatna zanimljivost ovog poslužitelja je što je nulti algoritam potpuno uklonjen sa sustava, a ne samo deaktiviran kao kod HP-a.

AMI MegaRAC nije ranjiv na ovaj problem, jer je opcija tvornički deaktivirana. Rezultati ranjivosti na ovu prijetnju su prikazani u tablici VII.

TABLICA VII. RANJIVOST NULTOG ALGORITMA

Vrsta	Rezultat	Napadačka tehnika
HP DL360	otporan	
Dell R630	ranjiv	Alati za slušanje prometa mrežne kartice
Dell R730	otporan	
AMI MegaRAC	otporan	

Svi navedeni poslužitelji su ranjivi na izvlačenje sažetka lozinke koje se vrlo jednostavno koriste kod alata za pogađanja lozinki (tablica VIII). Naime, ovo nije direktno nedostatak implementacije već samog IPMI standarda. Možda najgori problem s IPMI specifikacijom je činjenica da se lozinke interno spremaju kao nezaštićen tekst, a ne kao sažetak jer su potrebne u procesu računanja sažetka. IPMI u proceduri autentifikacije koristi dinamički generiran sažetak za provjeru valjanosti lozinke:

H(lozinka + privremeni broj + string + lozinka) (1)

Izraz (1) definira argumente funkcije sažetka koja se računa kod provjera autentičnosti korisnika. Vidljivo je da MC ne može izračunati sažetak ako nema pristup lozinci u čistom obliku. Upravo ovaj propust pruža mogućnost izvlačenja lozinke iz sustava, a tome pridonosi činjenica da privremeni broj i string generira klijent koji se želi autentificirati. Jedan od bitnih elemenata koje ovaj napad može znatnije otežati je definiranje lozinki i korisničkih imena kao slučajnih kombinacija, a ne kao tvorničke postavke. Takav primjer slijedi HP-ov poslužitelj.

TABLICA VIII. RANJIVOST OTKRIVANJEM SAŽETKA

Vrsta	Rezultat	Napadačka tehnika
HP DL360	ranjiv	Ipmitool, openipmi alati na Linux okruženju sa detaljnim ispisom
Dell R630	ranjiv	
Dell R730	ranjiv	
AMI MegaRAC	ranjiv	

Tvornički definirane vrijednosti lozinki u mnogo čemu su problematične sa sigurnosnog aspekta, što naročito dolazi do izražaja kod sustava koji su osnova velikih računalnih mreža. Tri od četiri sustava koja testiramo to potvrđuju. HP poslužitelj je onaj koji popravlja sliku na način da se za vrijeme proizvodnje generira slučajna lozinka od maksimalno dvadeset alfanumeričkih znakova i zapisuje se na naljepnici koja se postavlja na sam poslužitelj. S obzirom da je lozinka slučajno generirana, znatno je teže izvršiti pogađanje lozinke koristeći se listom često korištenih lozinki. Ostali proizvođači koriste lozinke koje su dostupne na Internetu, a zajedničke su za sve njihove proizvode. Primjerice Dell koristi kombinaciju root kao korisničko ime i calvin kao lozinku, dok AMI koristi kombinaciju admin/admin. Rezultati su prikazani u tablici IX.

TABLICA IX. RANJIVOST TVORNIČKIH POSTAVKI

Vrsta	Rezultat	Napadačka tehnika
HP DL360	otporan	
Dell R630	ranjiv	Razni alati za otkrivanje lozinki na temelju liste poznatih lozinki
Dell R730	ranjiv	
AMI MegaRAC	ranjiv	

E. Umanjivanje izloženosti prijetnjama

Sigurno najvažnija stvar bez koje se nikako ne može imati siguran IPMI sustav je razdvajanje i ograničavanje mreže MC-a i vanjskog svijeta. MC zbog ograničenih performansi ne pruža naprednije metode zaštite čime se zaštita treba obaviti drugim elementima kao što su usmjernici (engl. router) i preklopnici (engl. switch). Vrlo je važno razdvojiti sva sučelja koja omogućavaju komunikaciju (web, IPMI port 623) na način da se postavi jako sigurno računalo kao vatrozid (engl. firewall). Takvo kontrolno računalo može promatrati promet koji MC obavlja i na taj način otkriti anomalije te preventivno djelovati [29].

Zbog IPMI specifikacije sustavi često definiraju algoritam bez šifriranja u listi podržanih metoda autentifikacije. Kako smo vidjeli na testu, takav algoritam u potpunosti isključuje sve sigurnosne aspekte IPMI sustava. Vrlo je važno isključiti ovu mogućnost, ako ne u vrijeme proizvodnje, onda sigurno prije prve instalacije sustava. Ova ranjivost je uvedena sa IPMI verzijom 2.0 te u nemogućnosti njene deaktivacije, ipak je bolji način forsiranje korištenja IPMI verzije 1.5 koja nije izložena ovoj ranjivosti.

Preporučeno je korištenje lozinki od 16 ili 20 znakova, sa povećanom kompleksnošću. MC lozinke često mogu biti kompromitirane te je na taj način omogućen neovlašten pristup sustavu. Lozinke se trebaju mijenjati periodički, iako često zahtjeva dosta administratorskog posla, ali barem jednom godišnje. Svi korisnički računi moraju imati definiran životni ciklus. Čest je slučaj da korisnici koriste automatizirane aplikacije koje komuniciraju putem protokola IPMI, te na taj način definiraju periodičke cikluse promjene lozinki. U takvom pristupu vrlo je važno izvršiti testiranje nakon promjene kako bi se potvrdila uspješna promjena lozinke. Sve promjene se moraju zapisivati u odvojenom sustavu jer sam IPMI ne pruža mogućnost zapisivanja promjena lozinki. Svi korisnički računi koje tvornički definira proizvođač trebaju biti izbrisani te definirani vlastiti. Obavezno je onemogućiti pristup anonimnom korisničkom računu (engl. anonymous).

MC ne treba slati odgovore na ARP (Address Resolution Protocol) upite kako se sustav ne bih mogao detektirati u lokalnoj mreži. MC treba imati statički definiranu IP adresu, ne definiranu preko dinamičkih protokola.

Iako postoji cijeli niz dodatnih sigurnosnih problema, prethodno navedene stvari, proizašle iz višegodišnje praktične primjene, su dovoljan način osiguranja postojećih IPMI sustava.

IV. ZAKLJUČAK

Prolazeći kroz rezultate analize sigurnosnih propusta koje IPMI specifikacija donosi i sama pomisao na sigurnost sustava i aplikacija koje serviraju je nevjerovatna. Korisnici takvih sustava nisu uglavnom ni svjesni da postoji maleno ugradbeno računalo koje može potpuno narušiti sigurnost cijele organizacije. Rezultati pokazuju da dok god IPMI i MC-ovi imaju veliku količinu strukturnih, ahitekturalnih i proceduralnih ranjivosti vrlo je teško reći da je sustav siguran. Sve prethodno analizirane nedostatke pronalazimo u najmodernijim poslužiteljima koji se koriste kao gradivi elementi računala u oblaku (engl. cloud computing).

Iako problemi i dalje postoje, važno je što više zaštititi ranjivost sustava. Kroz članak je opisano nekoliko načina (engl. best practices) koje se trebaju pratiti kako bi se umanjile ranjivosti IPMI sustava. Jedini način rješenja svih problema je potpunim restrukturiranjem protokola, što je svakako teška odluka zbog velikog broja postojećih sustava.

LITERATURA

[1] J. Hargrave, "An Introduction to Intelligent Platform Management Interface," Dell Powersolutions

[2] Intel, Hewlett Packard, NEC, Dell, "Intelligent Platform Management Interface Specification," vol 2, Rujan 2013.

[3] W. Fischer, "IPMI Basics," Thomas-Krenn, 1963, https://www.thomas-krenn.com/en/wiki/IPMI_Basics, 20.10. 2015

[4] Intel, Hewlett Packard, NEC, Dell, "Intelligent Platform Management Interface Specification," vol 2, Rujan 2013.

[5] HP Integrated Ligts-Out, http://www8.hp.com/us/en/products/servers/ilo/integrated-lights-out.html, 05.03.2016

[6] Dell Remote Access Controller - DRAC / iDRAC, http://en.community.dell.com/techcenter/systems-management/w/wiki/3204.dell-remote-access-controller-drac-idrac, 05.03.2016

[7] AMI MegaRAC, https://en.wikipedia.org/wiki/MegaRAC, 05.03.2016

[8] IBM Integrated Management Module, https://www-947.ibm.com/support/entry/portal/docdisplay?lndocid=MIGR-5079770, 05.03.2016

[9] C. Cline, "Alert Standard Format (ASF) Specification," DMTF, vol. 1.03, Lipanj 2001

[10] D. Farmer, "Sold Down the River," Srpanj 2014.

[11] C. Cline, "Alert Standard Format Specification," DMTF, Intel, vol. 2.0, pp. 29, Travanj 2003

[12] Rapid 7 Comunity, "A Penetration Tester's Guide to IPMI and BMCs," Metasploit, https://community.rapid7.com/community/metasploit/blog/2013/07/02/a-penetration-testers-guide-to-ipmi, 13.01.2016

[13] My Knowledge Base, https://rodolfovaraujo.wordpress.com/2012/05/25/dell-idrac-default-user-and-password/, 13.02.2016

[14] Dan Farmer, https://en.wikipedia.org/wiki/Dan_Farmer, 05.03.2016

[15] IPMI Security VUlnerabilites, http://www.cisco.com/c/en/us/about/security-center/ipmi-vulnerabilities.html, 12.01.2016

[16] Alert TA13-207A, https://www.us-cert.gov/ncas/alerts/TA13-207A, 13.01.2016

[17] CVE-2013-4784, http://www.cvedetails.com/cve/CVE-2013-4784/, 05.03.2016

[18] VERT Alert: Supermicro IPMI/BMC Plaintext Password Disclosure, http://www.tripwire.com/state-of-security/latest-security-news/vert-alert-supermicro-ipmibmc-plaintext-password-disclosure/, 05.03.2016

[19] CARISIRT: Yet Another BMC Vulnerability, http://blog.cari.net/carisirt-yet-another-bmc-vulnerability-and-some-added-extras/, 05.03.2016

[20] HP Proliant DL360 Specifications, http://www.dell.com/learn/us/en/04/shared-content~data-sheets~en/documents~dell_poweredge_r730_spec_sheet.pdf, 13.02.2016

[21] Dell R630 Specifications, http://i.dell.com/sites/doccontent/shared-content/data-sheets/en/Documents/Dell-PowerEdge-R630-Spec-Sheet.pdf, 13.02.2016

[22] Dell R730 Specifications, http://i.dell.com/sites/doccontent/shared-content/data-sheets/en/Documents/Dell-PowerEdge-R730-Spec-Sheet.pdf, 13.02.2016

[23] AMI MegaRac Specifications, https://ami.com/ami_downloads/MegaRAC_DCMI_Spec_Sheet.pdf, 13.02.2016

[24] JSON Web Token, https://jwt.io/introduction/, 10.12.2015

[25] Redfish Scalable Platforms Management API Specification, https://www.dmtf.org/sites/default/files/standards/documents/DSP0266_0.96.0a.pdf, 13.01.2016

[26] Basic Access Authentication, https://en.wikipedia.org/wiki/Basic_access_authentication, 13.01.2016

[27] Linux Ubuntu 14.04, http://releases.ubuntu.com/14.04/, 13.02.2016

[28] D. Laurie, "ipmitool(1) - Linux man page," http://linux.die.net/man/1/ipmitool , 15.01.2015.

[29] D. Famer, "IPMI++ Security Best Practices," vol 1, Travanj 2013

.

miproBIS

International Conference on

BUSINESS INTELLIGENCE SYSTEMS

Steering Committee

Chairs: Boris Vrdoljak, University of Zagreb, Croatia
Mirta Baranović, University of Zagreb, Croatia
Matteo Golfarelli, University of Bologna, Cesena, Italy
Marko Banek, University of Zagreb, Croatia

Members: Bartosz Bebel, Poznan University of Technology, Poland
Ladjel Bellatreche, University of Poitiers, France
Ljiljana Brkić, University of Zagreb, Croatia
Alfredo Cuzzocrea, University of Calabria, Rende, Italy
Dragan Gamberger, Ruđer Bošković Institute, Zagreb, Croatia
Paolo Garza, Polytechnic University of Turin, Italy
Igor Mekterović, University of Zagreb, Croatia
Damir Pintar, University of Zagreb, Croatia
Zoran Skočir, University of Zagreb, Croatia
A Min Tjoa, Vienna University of Technology, Austria
Mladen Varga, University of Zagreb, Croatia

PAPERS

Analyzing air pollution on the urban environment

Elena Baralis, Tania Cerquitelli, Silvia Chiusano, Paolo Garza, and Mohammad Reza Kavoosifar
Department of Control and Computer Engineering, Politecnico di Torino, Torino, Italy
{name.surname}@polito.it

Abstract - Air pollution is one of the most important factor that can affect the quality of citizen life in the urban environment. Consequently, monitoring air pollution is currently a critical issue that needs to be addressed for enhancing the well being of citizens. This paper proposes a data analysis engine, based on business intelligence methodologies and open technologies, to support different targeted analysis on air pollution data. To analyse the problem from different facets, air pollution measurements are enriched with additional information such as meteorological and traffic data, which are collected through sensor networks available in the smart city context. This integrated dataset is periodically analyzed to generate informative dashboards based on a selection of Key Performance Indicators (KPIs). The informative dashboards can provide useful insights about pollutant concentration at different time granularity levels in the urban areas and support a joint evaluation of pollutant concentrations with climate conditions and traffic flow. As a reference use case, open data on air pollution in the urban area of a major Italian city is analyzed to demonstrate the effectiveness of the proposed approach in a real smart city context.

I. INTRODUCTION

Today's there is an the increasing concern of citizens and city administrations on air pollution, because of its possible significant impact on human health. The analysis of air pollution involves the simultaneous study of additional data that may impact on air quality such as traffic flow and meteorology. Consequently, the abundance of information available from different sensor networks deployed in the urban city provides an unprecedented opportunity to understand the quality of life of city dwellers and how it deteriorates over time due to air pollution.

This paper presents the APA (Air Pollution Analysis) system, based on business intelligence methodologies and open technologies, to efficiently support different targeted analyses for increasing user awareness on air quality in the city environment. Data on air pollution has been integrated with weather and traffic data to build a richer data collection and effectively support (i) the evaluation of the *pollutant concentration* in different time periods and urban areas, the analysis of the impact of (ii) *climate conditions* and (iii) *traffic of vehicles* on air pollution.

We defined a set of Key Performance Indicators (KPIs) to monitor the environmental pollution together with weather and traffic data at different spatial and temporal granularity levels. For example, a KPI has been defined to evaluate the concentration of a given air pollutant in a time period, its variation over time, or how many times it was above a given reference threshold in a given time period and spatial area. KPI values are graphically visualized (also on geographical maps) on informative dashboards to support an easy understanding of the results of the analysis and thus enhance user awareness on air quality. This analysis can provide useful information for the definition of medium- and long-term intervention policies aimed at reducing the concentration of pollutants. For example, the proposed informative dashboards can provide useful feedbacks about *when* and *where* pollutants reached critical concentrations. This information may support the local Administration in planning/re-planning the paths used for the distribution of goods in some urban zones or sizing alternative infrastructures to support (green) urban mobility (such as public transport and bike sharing systems) in some parts of the city.

Although the APA system is general and it can be applied to data acquired in different city environments, to assess the proposed approach we consider as use case scenario the analysis of air pollution data monitored in a major Italian city.

This paper is organized as follows. Section II discusses the main research activities addressing the analysis of air pollution. Section IIII describes the main building blocks of the proposed business intelligence system. Section IV summarizes the development and experiments on real data. Section V draws conclusions and presents future developments of this work.

II. RELATED WORK

Today's world cities are negatively affected by pollution that significantly deteriorates the quality of life of dwellers. Air quality can vary over time and across different areas of a city, and it is also influenced by different factors such as weather conditions (e.g., humidity, temperature and atmospheric pressure), human activities (e.g., traffic flows, people's mobility), provided services and the presence of points of interest in urban areas [1]. Monitoring air quality, in terms of pollutant concentrations such as PM2.5, NO2, and SO2 is usually performed by means of a set of fixed stations deployed in the urban environment. Since the costs of designing, developing, deploying and maintaining these stations are high, they are usually deployed in a limited number.

With the evolution of mobile communication protocols and sensing technologies innovative sensors have been developed to monitor a wide range of pollutant, thus evaluating air quality. Authors in [2] proposed to monitor the air quality through the deployment of mobile sensors on the bicycle wheels. Such sensors are able to monitor CO2 and some meteorological data such as temperature and atmospheric pressure. Despite their high potential,

these approaches allow monitoring a limited number of gaseous substances (such as CO2 and CO), while other pollutants (such as PM2.5 and PM10) require devices not easily portable and long periods of monitoring to generate accurate measurements [1].

A parallel effort has been devoted to designing innovative business intelligence and/or data mining solutions to perform different targeted and interesting analyses on pollutant measurements to evaluate air quality. Authors in [3] studied the pollutant concentration in different cities, or in different areas of a city, its variation over time and the correlation degree between concentrations of pollutants and other information such as weather conditions. Pollutant measurements were collected through a network of fixed monitoring stations, integrated with meteorological data and stored in a data warehouse. To monitor air quality different indicators were defined to perform the analyses at a different spatial-temporal granularity. The APA engine addresses the research issue discussed in [3]. However, APA exploits different technological solutions and supports a richer set of analyses because traffic data are also integrated in the system.

Data Mining techniques were used to support the analysis and the prediction of air quality in the urban environment [1, 4]. Authors in [1] proposed to exploit classification techniques to forecast the air quality level in areas without monitoring stations. Historical and real-time measurements on air quality together with additional information such as weather, traffic, and people's mobility have been analyzed as training data. Authors in [4] proposed a jointly analysis of historical data on air quality and weather forecasts to predict future values (in the next time periods) on air quality. In the current implementation of APA the business intelligence methodologies have been only integrated. However, data mining algorithms will enrich the second release of APA to address advanced air quality data analytics.

III. THE APA PLATFORM

The APA engine is designed for the modeling, integration, storage, and analysis of a large amount of heterogeneous data related to air pollution to provide different levels of relevant knowledge on air quality on their urban environment to users. APA supports different targeted analyses for different users, i.e., citizens and staff of the public administration.

To address air quality analyses, different data types provided by ad-hoc sensor networks deployed in the urban areas are integrated in APA. Specifically, data on *pollutant concentrations* were integrated with *climate data* and *traffic data* related to traffic of vehicles. All preprocessed data are stored into a unique data repository (a *data warehouse*). Different *key performance indicators* (KPIs) have been defined to monitor the concentration of various pollutants and to analyze their trend over time together with meteorological and traffic data. APA also includes *informative dashboards* to present the results of the analysis (e.g., trend of KPIs over time, correlations between two KPIs) on air quality in an informative fashion, using simple and easily understandable data representations. Each component of APA is detailed in the next sections.

A. Data sources

The APA engine is currently integrating three different types of data. In addition to measurements on pollutant concentration, APA also considers climate data and data on traffic of vehicles, since both aspects may have a significant impact on the air quality value. The considered data categories are briefly described below.

(i) Measurements on *pollutant concentrations* in the urban areas. Different pollutant are currently analysed in the APA engine including particulate matter (PM10 and PM2.5), carbon monoxide (CO), ozone (O3), nitrogen dioxide (NO2), benzene (C6H6), sulfur dioxide (SO2).

(ii) *Climate conditions* in the urban areas, including air temperature, relative humidity, precipitation level, wind speed and atmospheric pressure.

(iii) *Vehicle traffic data* as the number of vehicles entering in a given urban area at a given time granularity (e.g., hourly). To characterize traffic under different perspectives, we collected traffic data for different categories of vehicles. Specifically, we categorized vehicles based on their fuel type (e.g., petrol, diesel) and their use (e.g., bus, private vehicles or transport of goods).

For each of the above data types, measurements in the urban environment are usually collected at a given time granularity through ad-hoc *geo-referenced sensor networks* deployed in the city area.

Concentration measurements for each pollutant were collected through sensors deployed in monitoring stations located in different areas of the city. Each *pollution monitoring station* (PolMS) is characterized by the geo-coordinates (i.e., latitude and longitude) of its location. It includes different sensors to monitor the concentrations of various pollutants at hourly and daily time granularity.

For monitoring climate conditions, the Weather Underground web service[1] has been considered. It gathers data from a geo-referenced network of Personal Weather Stations (PWS) registered by users. For many cities a large number of stations are distributed throughout the territory. Although the measurement frequency can be easily set by the user for each PWS (and can vary over time), the average value for the ones we considered was about 15 minutes.

Traffic data were collected through geo-referenced sensor networks measuring the number of vehicles (for each of the above considered vehicle types) entering, with hourly time granularity, in different urban areas.

B. Data preparation and modeling

To efficiently support different targeted analyses, data on pollutant concentrations are integrated with climate and traffic data. These data are additional measurements describing the environmental context in which the pollutant concentrations were monitored.

[1] http://www.wunderground.com/

Available climate and traffic data are pre-processed before the data integration phase because of their (possible) different time and space granularities with respect to the pollutant timeline and the monitored areas. Since we are focusing on pollution analysis, the spatial-temporal granularity of the sensor network monitoring pollutant concentrations has been considered as a reference. Then, data coming from climate and traffic monitoring networks have been integrated using the same spatial-temporal granularity of the reference network.

For the weather Underground web service, a large number of Personal Weather Stations (PWS) are distributed throughout the territory for many cities, and the measurement frequency can be easily set by the user for each PWS. However, sensors networks monitoring pollutants and climate data may have different geo-locations and sampling rate in collecting measurements. For example, weather data may be unavailable for a specific pollution monitoring station (PolMS) located in a given urban area or at a given instant of time, while the pollutant measurements monitored in the station are instead available. To deal with these issues, in the data integration phase, weather data associated with a given pollution station are computed as a distance-based weighted mean of the values provided by the three nearest PWSs. The weight is inversely proportional to the distance from the three PWSs to the PolMS. Hence, three equally distant PWSs would have the same weight in determining the weather values of a given city zone. Weather data timestamps were aligned to the closest timestamp available for the given pollution monitoring station (PolMS) through an approximate join.

For traffic data readings, the number of entering vehicles in each area has been associated to all the sensors deployed in the area. Traffic data were timely integrated through an approximate join similar to that adopted for climate data integration.

Air quality data together with weather and traffic data are then stored into a unique data repository (a *data warehouse*), whose conceptual representation, according to the Dimensional Fact Model (DFM) [5], is reported Figure 1. The *fact table* consists of a main measure, i.e., the hourly pollutant concentration per sensor, and some additional measures describing the context in which the pollutant concentration was monitored. These additional measures are the climate conditions (e.g., hourly temperature, humidity, and wind speed per sensor) coming from outdoor PWSs, and traffic data per sensor (e.g., number of entering vehicles for each type of vehicle). Three *dimension hierarchies* are defined to analyse the spatial and temporal distribution of the air quality and the other contextual information with different granularity levels. Specifically, we defined a *location-related dimension hierarchy* linked to the sensor device monitoring the concentration of a given pollutant, and *two temporal-related dimension hierarchies* linked to the time for pollutant concentration values reported in the fact table. Dimensions are detailed below.

The *temporal-related dimension hierarchies* provide many different blends of time spans. In one hierarchy, time spans from hours to time slots. The hour is aggregated into different intervals (2-hours, 8-hours), and the corresponding daily time slot (morning, afternoon, evening, or night). In the other hierarchy, time spans from dates to years. Moreover, each date is classified as working or high day and as week day (e.g., Monday or Tuesday).

The *location-based dimension hierarchy* starts from physical sensors and builds up to the whole city area, with the type of sensor (i.e., the measured pollutant) as related feature included in the dimension. To analyze the spatial distribution of the air quality, higher-level space granularities are also considered beyond the geographical coordinate of the sensor from which the measurements refer to. Each sensor is mapped to the corresponding address and city area. While the geographical coordinate is recorded for the considered sensor, the address and the area name are added as additional contextual features to perform interesting analysis.

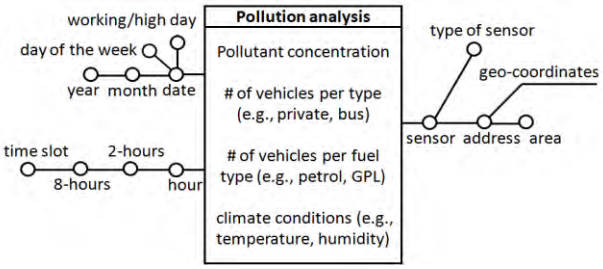


Figure 1 - Data warehouse conceptual model

C. Data analysis

The prepared data are analyzed to gain insights into the air pollution condition on the urban area. The aim of the analysis is to produce useful feedbacks to the end-users by generating informative dashboards using Key performance Indicators (KPIs).

We identified two operational roles representing users of the APA engine: (i) *staff of the local administration* and (ii) *citizens* of the urban environment. (i) The *staff of the local administration* is mainly interested in understanding the main causes of pollution affecting the city with the aim of guaranteeing a good level of air quality. These users are interested in monitoring air pollution in the urban area to point out *when* and *where* the pollutant concentrations became critical. They are also interested in assessing contextual aspects that may contribute to air pollution as the climate condition and the traffic of vehicles in the urban areas. It follows that these users need to analyze the complete streams of collected data (also at different spatial and temporal granularities), to observe and understand the observed phenomenon, assess the different components and identify possible causes. (ii) *Citizens* are interested in locally assessing the air quality in some specific urban areas, as for example where they live or work. These users need the visualization of a few indicators at different spatial and temporal granularities, but possibly provided in an easily understandable and intuitive way.

APA includes informative dashboards to present the results of the analysis (e.g., trend of KPIs over time, correlations between two KPIs) on air quality in an informative fashion, using simple and easily understandable comparisons among KPI values. In our

context, KPIs are quantitative indicators of the monitored air pollution condition. Moreover, additional KPIs are used to quantify the climate and vehicle traffic conditions aimed at enriching the analysis on pollutant concentration. The following KPIs have been defined.

(A) *Spatial-temporal pollutant concentration KPIs*. They report the pollutant concentration per sensor, at different time granularity levels in a user-specified date. The following three KPIs are defined. (A.1) Daily average pollutant concentration per sensor. (A.2) Average pollutant concentration per daily time slot (morning, afternoon, evening, night) per sensor. (A.3) Average hourly pollutant concentration per sensor.

(B) *Critical pollutant concentration KPI*. This KPI analyses how frequently the daily pollutant concentration per sensor is critical. Specifically, it reports the percentage of days within a given time period in which the pollutant concentration was critical according to a user-specified threshold.

(C) *Climate KPIs*. They report the hourly average values in a given date for various climate data as temperature, humidity, and pressure.

(D) *Vehicle traffic KPIs*. To characterize traffic under different perspectives, vehicles have been categorized based on the type of fuel (e.g., diesel, petrol, or methane) and the use of the vehicle (e.g., bus or private). KPIs report the total number of vehicle per day for each of the above categories, within a user-specified time period in a given urban area.

(E) *Statistical correlation analysis KPI*. Pollutant concentrations can be affected by meteorological conditions. To deepen the evaluation of the impact of climate conditions on air pollution, we analysed the statistical correlation between pollutant concentration and climate data. The established Pearson correlation coefficient [6] has been currently adopted for KPI computation over the sequence of measurements collected within a given time window. The Pearson correlation coefficient assumes values in the range [-1, +1]. Values higher than 0 represent positive correlations, while values lower than 0 represent negative correlations. The higher the absolute coefficient value the stronger the correlation.

APA includes four types of informative dashboards each one reporting one or more of the above KPIs. Dashboards have been designed to address the following targeted analyses: the analysis of (1) the *monitoring station characteristics*, (2) the *spatio-temporal pollutant concentration*, (3) the *impact of climate conditions on pollutant concentration,* and (4) the *impact of vehicle traffic on the pollutant concentration*. In the dashboards, we combined an interactive map to easily visualize the distribution of KPI values in geographical areas, with charts reporting more in detail the trend of KPI values over time. Dashboards are briefly described below.

(1) The *monitoring station characteristics* dashboard shows on a map the position of the pollution monitoring stations, and for each of them the types of measured pollutants. This dashboard provides useful information to understand the available hardware infrastructure, to analyze the statistical significance of the achieved results, and to support the planning of further extensions of the monitoring station network.

(2) The *spatial-temporal pollutant concentration* dashboards allow characterizing the pollution in urban areas over time. The temporal/spatial-based analysis allows understanding if some time periods/urban areas are more subject to air pollution than others. To support data analysis at different temporal granularity levels, three different dashboards have been defined based on KPIs on spatial-temporal pollutant concentration (KPIs (A)). An additional dashboard points out the number of days with pollutant concentration critical with respect to a user-specified threshold (KPI (B)).

(3) The *impact of climate condition on pollutant concentration* dashboards allow analyzing which weather conditions strongly affect the air quality in different urban areas and over time. A first dashboard includes the spatial-temporal analysis of the climate KPI trend jointly with pollutant concentration KPIs (i.e., KPIs (A) and (C)). A second dashboard reports the statistical correlation between climate conditions and pollutant concentration (i.e., KPI (E)).

(4) The *impact of vehicle traffic on pollutant concentration* dashboards allows analyzing how the circulating vehicles in urban areas impact on the concentration of the pollutants. To this end the vehicle traffic KPIs jointly with spatial-temporal pollutant concentration KPIs are computed and their trends (over time) are shown in easy readable charts. (i.e., KPIs (A) and (D)).

IV. DEVELOPMENT AND RESULTS

This section presents the current implementation of the APA system and its validation on real open data collected in Milan, a major city in the north of Italy. Some examples of interesting analysis scenarios are also discussed.

A. The APA development

The APA architecture is based on a multi-tier structure for supporting flexible, easily reusable and customizable applications. APA has been developed using open source technologies. Since all data are stored in the data warehouse based on the (fixed) schema reported in Figure 1, the technological solution adopted in APA for data storage exploits a relational DBMS. Specifically, MySQL is the RDBMS currently selected for empowering APA analytics. To visualize the results of the analysis, a secure responsive web-based application has been developed. The application supports various devices with a different screen size like mobile phones, tablets, and desktop computers. The application is based on HTML 5 and CSS 3 and it has been developed using the PHP programming language. To provide a more interactive navigation for the end user, the KPI visualization on geographical maps is based on a multiple-layer representation, where more geographical layers are overlapped (as the layer representing different geographical areas). Google Fusion Tables are used because of their potential in mapping KML files for creating layers on Google Maps.

In this study the assessment of the APA system was performed by deploying it on a dual-core 2.50 GHz Intel(R) Xeon(R) workstation with 4 GBs of RAM, running Ubuntu Linux 12.04 LTS.

B. *Datasets*

This section presents the three data sources considered in this study. Data from these sources have been collected for year 2013, and then integrated in APA as discussed in Section III.B.

The *ARPA Lombardia Pollutant dataset*[2] includes the concentration values for a set of pollutants (e.g, PM10, PM2.5, CO, O3) gathered through some monitoring stations located in the Lombardia Region. Each station is equipped with a set of sensors, each one measuring a given pollutant. We focused our analysis on the monitoring stations located in the city of Milan. The provided data consists of a set of hourly or daily readings, depending on the type of pollutant. Each reading is characterized by the monitoring station identifier, the sensor identifier, the name of the measured pollutant, the measured concentration of the pollutant, and the date and hour of the reading.

The *Weather Underground dataset* [3] includes the meteorological measurements collected through all Personal Weather Station (PWSs) registers by users. Since our scenario analysis is related to the city of Milan, we selected three PWSs near to the considered urban environment. For each selected PWS, meteorological measurements considered in APA (e.g., humidity, and atmospheric pressure) have been downloaded for the considered time period. Each record includes the identifier of the PWS, the timestamp of the measurement and for each weather measure the corresponding measured meteorological value. As discussed in Section III.B, we properly aggregated the weather data collected through different PWSs to obtain a representative value at the hourly granularity associated to each sensor of the pollutant monitoring stations.

The *Municipality of Milan's Traffic dataset*[4] provides an aggregated information about the number of vehicles entering in the central area of Milan per hour. The dataset contains the total number of entering vehicles per hour and the number of vehicles separately for each vehicle category. Vehicles are classified according to the fuel type (e.g., petrol, diesel, gasoline) and to the vehicle type (e.g., private, bus, goods). As discussed in Section II.B, traffic data has been associated to sensors monitoring pollutant concentrations in the central area of Milan.

C. *Analysis scenario*

Here we discuss four representative examples of interesting analyses that can be performed using APA. These analyses have been carried out on data related to the city of Milan, selected as a reference case.

Example (A): Analysis of daily pollutant concentration. The "Daily pollutant concentration" dashboard allows the user to select a date of interest and a pollutant to analyze. For the selected pollutant, the dashboard visualizes the following information. (i) The average concentration in the selected date in Milan areas, through colored markers placed on a map (one marker for each monitoring sensor). Markers are colored based on some predefined thresholds to visually represent a "good", "non-critical", "warning", "alert", or "alarm" pollutant concentration. (ii) The concentration trend over a ten days period including the selected date and the nine preceding dates. This analysis can support in identifying critical daily pollutant concentrations, but also in evaluating increasing/decreasing trends in pollutant concentration over a time period. As an example, Figure 2 reports the average daily PM10 concentration for April 11, 2013. In the upper part of Figure 2, the yellow markers on the map highlight a potential critical PM10 concentration on April 11, 2013 in three different areas of Milan. Moreover, the chart in the bottom part of Figure 2 shows an increasing PM10 concentration for all the three areas in dates from April 9 to 11, 2013, with the highest value on April 11 (bars on the right-end side of the chart). In the considered time period, a great event took place in Milan that may be one of the factors contributing to the increment of pollutant concentration, because a larger number of vehicles may have accessed the Milan areas.

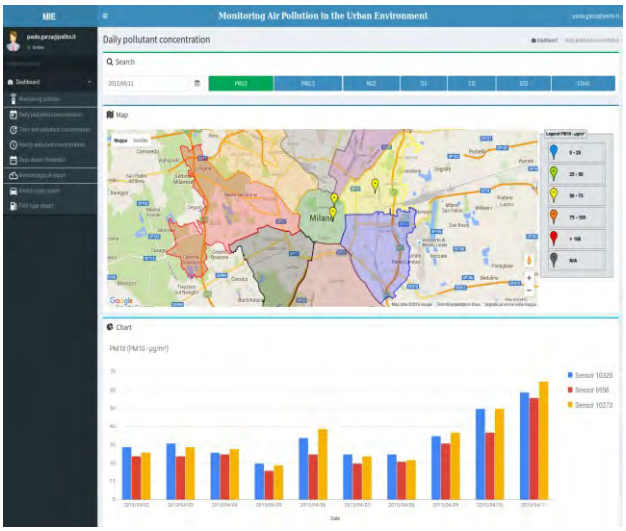


Figure 2 - Daily pollutant concentration dashboard

Example (B): Days above threshold. Using APA as discussed in Example (A), the user can analyse daily pollutant concentrations and discover critical days for pollutant concentration. Once a critical date, or a critical time period, has been pointed out, the user can deepen the analysis through the "Days above threshold" dashboard. This dashboard allows analyzing the percentage of days for which the pollutant concentration was above a user-specified reference threshold. This analysis can reveal the persistence of high pollutant concentrations over a long time period, which is critical for the wellness of citizens. As an example, we analysed the PM10 concentration from April 9 to 14, 2013, by setting 50 µg/m³ as a threshold. The results show that in the analyzed time period in two out of the three sensors the percentage of days above

[2] http://www2.arpalombardia.it/
[3] http://www.wunderground.com/
[4] http://dati.comune.milano.it/

threshold is higher than 33% and in the third sensor is 16%.

Example (C): Impact of vehicle traffic on pollutant concentration. The number and type of circulating vehicles can significantly affect the air quality. APA includes the "Impact of mobility on air pollution" dashboard to analyse the correlation between the number of vehicles in a given area and the pollutant concentration in the area, as well as to point out the most impacting types of vehicles on pollutant concentration. In the dashboard, first the user selects a time period of interest and a pollutant to analyze. Then, a chart in the upper part of the dashboard reports the number of vehicles entering in the central area of Milan each day in the time period. Both the total number of vehicles, and the number of vehicles per type, are reported. In the bottom part, for the same time period, a chart reports the daily concentration for the selected pollutant. As an example, Figure 3 reports the status of the dashboard for the PM10 pollutant in the time period from June 1 to 15, 2013. These results show that, for the analyzed area, (i) the majority of the traffic is related to private vehicles (approximately 80%) and (ii) the PM10 concentration has a trend similar to that of the number of private vehicles. For reducing the number of private vehicles, staff of the public administration can plan campaigns in favor of alternative (green) transportation vehicles (e.g., bicycles, electric cars, underground) for private users.

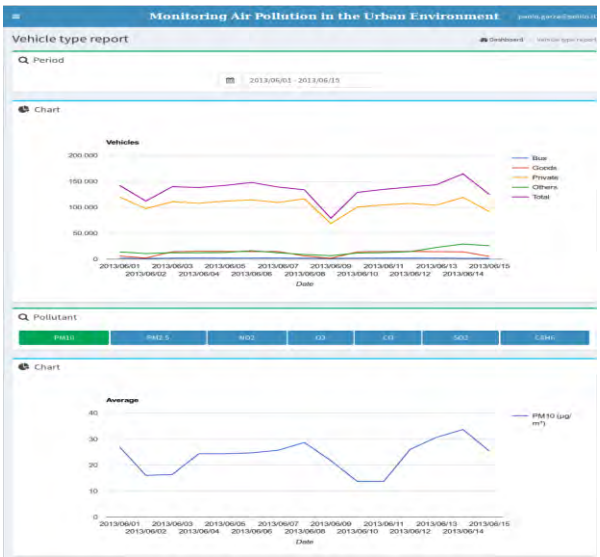


Figure 3 - Vehicle type dashboard

Example (D): Correlation between climate conditions and pollutant concentrations. This dashboard allows analyzing the statistical correlation between meteorological conditions and pollutant concentrations over a time period. Given a pollutant and a meteorological feature (e.g., temperature), the dashboard plots their statistical correlation (based on KPI (E)) per month within a selected time period. For example, we analyzed the correlation between temperature and ozone from January, 2013 to December, 2013. The chart (not reported here for sake of space) showed a positive correlation over all time period (correlation value always greater than 0) and a significant positive correlation from June to September (correlation value greater than 0.5) when the temperature is usually higher (i.e., in the summer season). This result is consistent with previous studies discussing the impact of temperature on ozone concentration [7].

V. Conclusions and Future Works

This paper presented APA, a business intelligence engine to analyze air pollution from different perspectives (e.g., pollutant concentration, meteorological data, traffic data) for the purpose to make citizens aware of air quality and to understand how people activities impact on air quality. As a case study, APA has been evaluated on open data on air pollution in the urban area of a major Italian city. There is still room for improvements for our system. For example, APA may be enriched with advanced data mining algorithms: (i) to analyze the correlations hidden in the air pollution-related data at different abstraction levels [8] and (ii) to forecast fine and coarse grained air pollution data [9].

VI. Acknowledgment

The research leading to these results has received funding from the Italian Ministry of Research (MIUR) under the "Cluster Tecnologie Smart Communities Progetto MIE – Mobilità Intelligente Ecosostenibile".

References

[1] Y. Zheng, F. Liu, and H. Hsieh, "U-air: when urban air quality inference meets big data," in the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, pp. 1436–1444, 2013.

[2] S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, and B. Nath, "Real-time air quality monitoring through mobile sensing in metropolitan areas," in Proceedings of the 2Nd ACM SIGKDD International Workshop on Urban Computing, pp. 15:1–15:8, 2013.

[3] Y. Xiao and C. Ji, "Management of air quality monitor data with data warehouse and GIS," in CSIE 2009, 2009 WRI World Congress on Computer Science and Information Engineering, pp. 148–152, 2009.

[4] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2267–2276, 2015.

[5] M. Golfarelli, D. Maio, and S. Rizzi, "The dimensional fact model: A conceptual model for data warehouses," in International Journal of Cooperative Information Systems, vol. 7, pp. 215–247, 1998.

[6] J. Han, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers Inc., 2005.

[7] E. Stathopoulou, G. Mihalakakou, M. Santamouris, and H. S. Bagiorgas, "On the impact of temperature on tropospheric ozone concentration levels in urban environments," in Journal of Earth System Science, vol. 117, issue 3, pp. 227-236, 2008.

[8] D. Antonelli, E. Baralis, G. Bruno, L. Cagliero, T. Cerquitelli, S. Chiusano, P. Garza, and N. A. Mahoto, "Meta: Characterization of medical treatments at different abstraction levels," ACM TIST, vol. 6, no. 4, p. 57, 2015.

[9] E. Baralis, T. Cerquitelli, S. Chiusano, A. Giordano, A. Mezzani, D. Susta, and X. Xiao, "Predicting cardiopulmonary response to incremental exercise test," in 28th IEEE International Symposium on Computer-Based Medical Systems, CBMS 2015, pp. 135–140, 2015.

Application of Model Driven Architecture for Development of Data Consolidation Web-system

A.A. Korobko* and L.F. Nozhenkova*
* Institute of Computational Modeling of the Siberian Branch of the RAS (ICM SB RAS), Russia, Krasnoyarsk
agok@icm.krasn.ru

The current state of information technology and Model Driven Development capability allow authors to suggest original approach to development and maintenance of model driven systems, without programming and using external software for domain engineering. Advantages and disadvantages of Model Driven Development usage for data consolidation are described in this article. Modification of model driven architecture in order to provide dynamic reconfiguration of the platform-independent model and content of the model driven system is proposed. The proposed approach is embodied in Scientific Activity Management System for ICM SB RAS and Data Consolidation Web-system for the Center of Emergency Monitoring and Prediction of Krasnoyarsk region.

I. INTRODUCTION

Due to development of information technologies, there appear new ways to automatize consolidation and processing of large amount of data that presently is accumulated and transferred in the form of scheduled records and reports. Consolidation of such data is important for different monitoring systems, for example, for scientific activity management, emergency monitoring and for many other applied fields.

The peculiarity of data collection in such areas is that the forms are often changed: new ones are introduced, the old ones are deleted and attributes are added, deleted and re-grouped. An information system must promptly respond to the form changes as well as to provide access to the collected data which is especially important for emergency prediction.

We can solve the problems related to new attributes adding and editing of the existing ones with the help of creation of a tuned data consolidation system, adaptable to the changing user requirements. The system must provide entry, storage and analytical data processing, as well as dynamic extension of the thematic content. The most promising approach allowing to achieve the required level of adaptability of the information system is based on the Model-Driven Development (MDD) [1, 2] technology. However, this technology has a number of features that prevent its direct usage for building of a data consolidation system. Firstly, changing of abstract models requires the work of a highly-skill specialist with additional knowledge in the area of domain modeling and UML-notations. Secondly, the result of all the stages of MDD is a source code of the system, which means that the finishing of the system will include the necessary stage of programming. Thirdly, modification of abstract models includes the necessity to match the old and new schemes of data storage to provide their continuity [3].

It is important to create technological basis for development of model-driven data consolidation system within the original approach to MDD application. Model-driven system is a system where the thematic content can be changed using the tools of the system itself by rebuilding of the control model.

This article presents approbation of the classical approach to MDD for data consolidation issue, and describes the problems related to its application. Authors suggest the approach to development of model-driven data consolidation systems based on original application of MDD technology taking into account the requirements of dynamic reconfiguration of the thematic content of a system.

II. MODEL-DRIVEN APPROACH TO SOFTWARE DEVELOPMENT

The technology of system development using the Model-Driven Architecture (MDA) is based on constructing an abstract control metamodel, metadata (model) exchange and setting of the ways for its transformation into the supported software techniques [1, 2, 3]. The central part of MDA is the four-layer architecture (Fig. 1) that has a number of standards and notations for each of its layers.

Classical MDA consists of two basic types of modeling: creation of abstract models (i.e. platform-independent models) and creation of applied models (i.e. platform-dependent models). The first type includes construction of a meta-metamodel (layer M3) which presents modeling language [4] and construction of meta-model (layer M2) which presents some different specific application domains. The second type includes construction of applied model (layer M1) and formation of the instances of concepts (layer M0) determined on layer M1 [5]. The result of modeling is a source code that partially implements the system functions and it is used by programmers to get the working system.

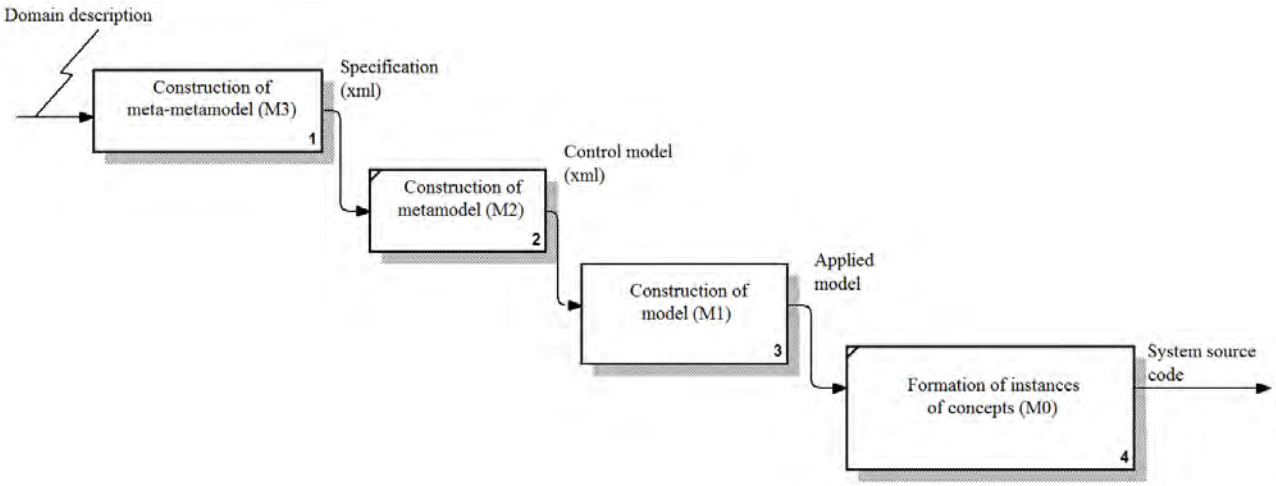


Figure 1. Diagram of classical MDA-based system development

III. CLASSICAL APPROACH TO DATA CONSOLIDATION WEB-SYSTEM DEVELOPMENT BASED ON MODEL-DRIVEN ARCHITECTURE

Basing on the classical approach to MDA application the authors have constructed a model of system development taking into account the actually pursued task and domain description (fig.2).

This diagram, as well as its prototype (fig.1), contains the processes of platform-independent and platform-dependent models construction. At the abstract level in the UML notation, firstly the meta-metamodel is constructed as a model of classes which describes the set of domain objects and the order of their interaction (process 1 "Construction of meta-metamodel", fig. 2). This model is very rarely changed and it can be created by a specialist skilled in MDD. Then, according to developed specification together with a user the metamodel is constructed (xml-format) as an object model which describes important aspects of the domain (process 2 "Construction of metamodel", fig.2) [6].

At the platform-dependent level using specification and system implementation requirements a relational scheme of metadata storage is created (process 3, fig. 2) [6]. Decomposition of the process "Creation of metadata storage relational scheme" is shown in fig. 3.

In the process 3 "Creation of metadata storage relational scheme", the tables for storage of the objects are created in accordance with classes defined in specification of the meta-metamodel. In addition, the tables include foreign keys to store relations between classes. Also, tables are created to record the multi-multivalue relations. The final scheme allows to store full information about the metamodel objects, but it's not enough for system operation. In the process 3.4 "Extension of metadata structure", based on the data about system implementation (e.g. software language, the library of components), the

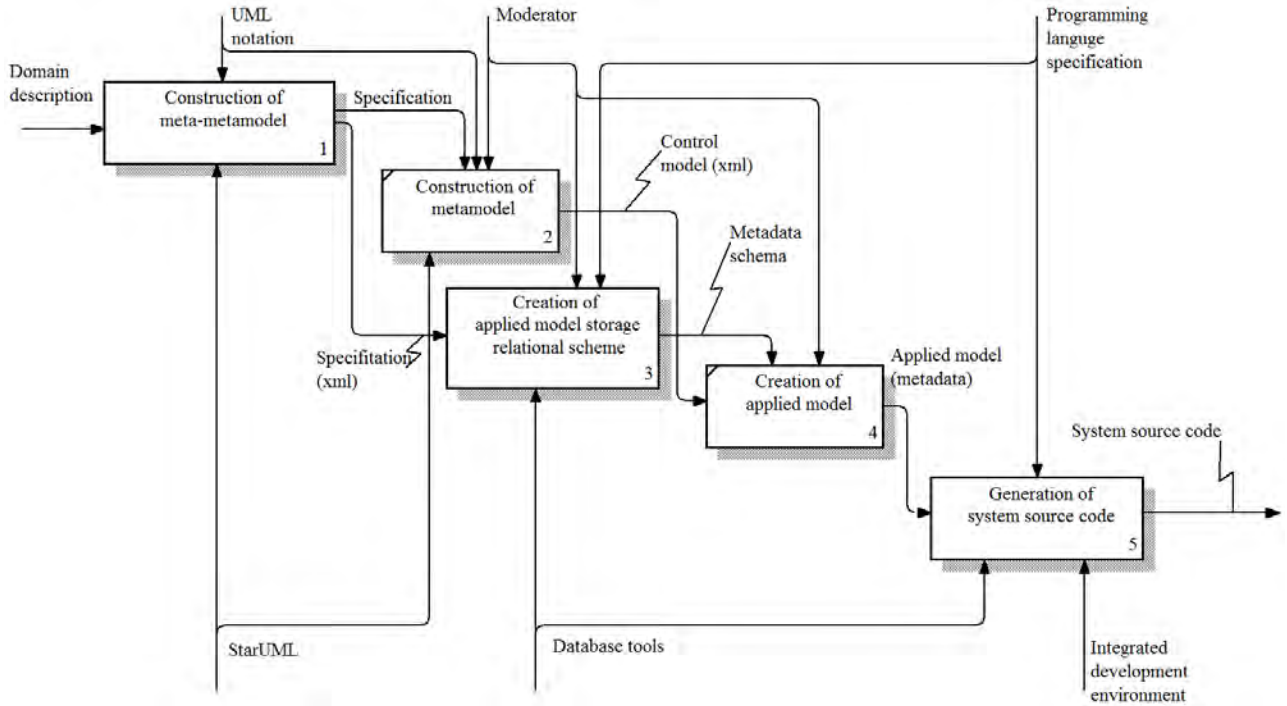


Figure 2. Diagram of MDA-based system development

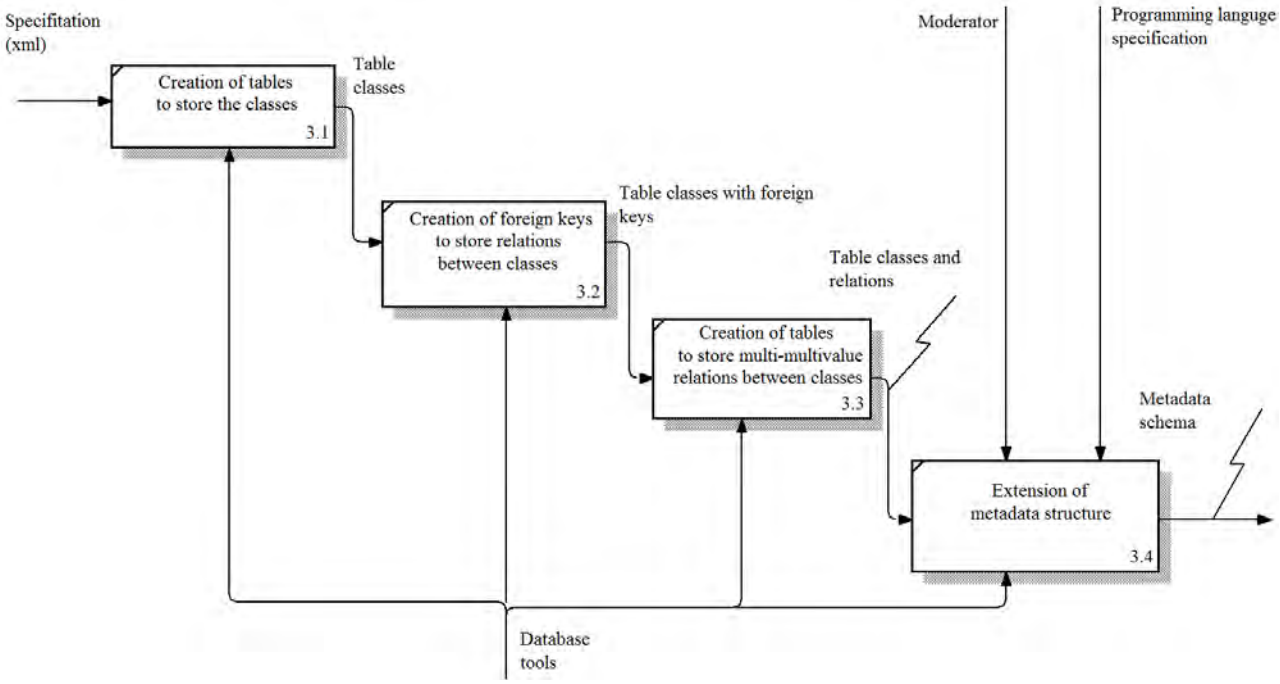


Figure 3. Creation of metadata storage relational scheme

metadata structure is extended.

In the process "Creation of Applied model" (process 4, fig. 2), metadata is filled in accordance with the created metamodel. At the first stage of this process, xml-file containing information about metamodel is processed and most of the tables are filled. At the second stage, the user fills the table fields that are used for system implementation. As a result, we get all required information to create the source code of a system (process 5 "Generation of the system source code", fig. 2).

The classical approach to application of MDA allows to simplify the system reconstruction in accordance with the requirements. However, it requires the participation of a programmer to get the working system that decreases significantly the reaction to models changes. Also, there is a problem with modification of the control model. Within the chosen approach, creation of the control model is completed using external software StarUML that provides the possibility to construct different UML-diagrams, including the required class models and object models. Application of this type of software for building the control model produces the number of positive aspects, the most important of which is the possibility to use the acquired result in similar domains. But, at the same time, the control model is transferred for further processing in the xml-file format and in order to create the applied model we have to merge the existing and the imported models. In the case if the system was built using the existing model and the data was collected, the change of data structure can lead to its loss. Such external software also creates problems for users because it requires additional knowledge in modeling and UML-notation.

It is important to make it possible to create a system automatically based on the acquired source code (without programmer engagement) and to integrate the control model into the system that allow to track model changes and take proper measures, and users will be able to change the control model using the system interface.

IV. ORIGINAL APPROACH TO SYSTEM DEVELOPMENT BASED ON MODEL-DRIVEN ARCHITECTURE

In order to integrate the control model into the system, it is necessary to delegate the process of metamodel creation (fig. 2) to the system or to distribute the tasks solved by this process between the processes of the system. Also, it is necessary to determine the control model storage method.

As long as the system uses a database for its function, it is logically to store information about the control model there as well. This task has been already solved at the stage of "Creation of metadata storage relational scheme" where we have got the structure containing full information about the objects of metamodel (fig. 3).

We suggest to combine the "Metamodel design" and "Applied model creation" processes of classical approach in a new process "Creation of control and applied models" (fig.4). In the process 3 the control and applied models are created by system tools instead of external tools. Models are presented in the metadata forms. Transformation from the control model to the applied model is completed by adding the attributes related to specific system implementation.

The process "Creation of the control and applied models" allows to export the control model created with the help of a model-driven system for re-usage. The result of this process is used in data consolidation web-system for the Center of Emergency Monitoring and Prediction of Krasnoyarsk region. Also, it is planned to be used in mobile applications taking into account the specifics of Android, iOS and Windows Mobile platforms.

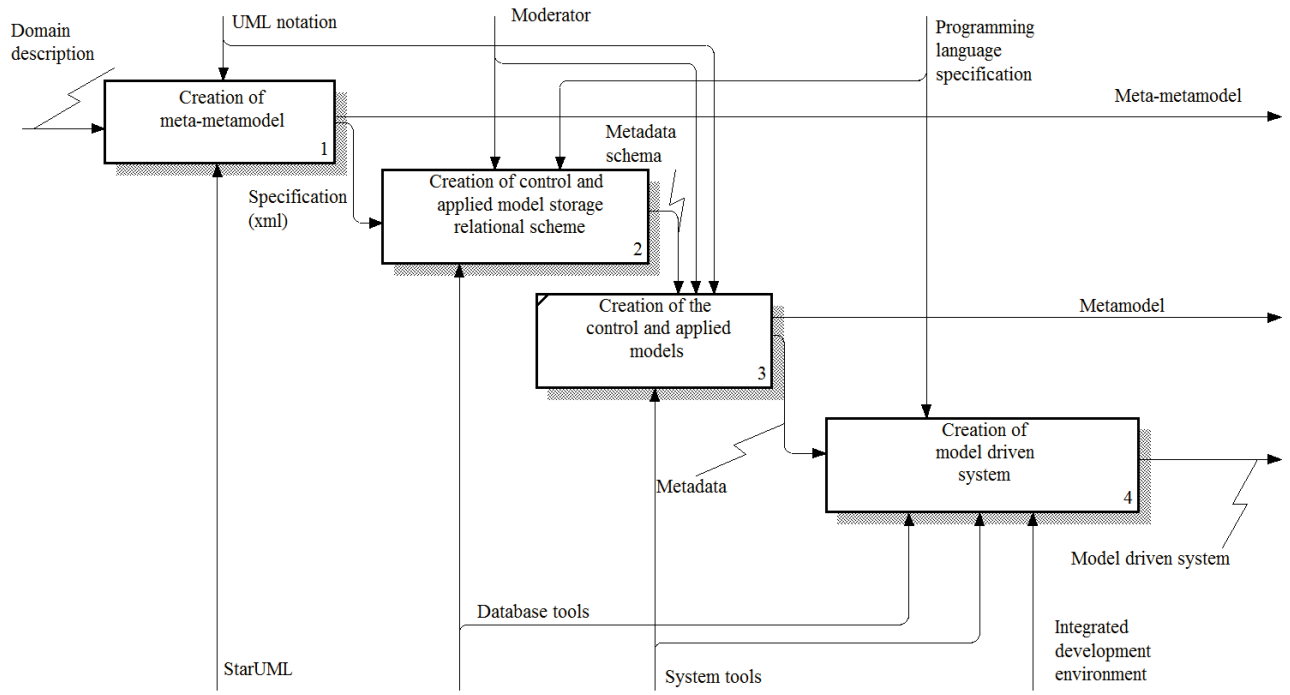


Figure 4. Diagram of modified MDA-based system development

The control model data storage in the form of metadata allows user to reconstruct the model-driven system without a programmer. Details of this process is shown in fig. 5. Model-driven system consists of server and client parts. Server part provides actuality of the database structure, analysis of metadata and creation of program components. Client part is responsible for creation of the user interface in accordance with the thematic content. Communication between server and client parts occurs in JSON format because this format is optimal for the serialization of complex structures describing the control model.

In the process "Database maintenance in accordance with the metadata", the tables for accumulated data are created or updated. The work of this process is very responsible, because while creating new tables one should consider cross-table relationships, create foreign and primary keys, and it is important to keep the previously collected data while editing the existing tables.

In the process "Generation of object model for program components", user interface description is created in accordance with the control and applied models. The result is sent to the client part in a JSON format.

Generation of user interface takes place in the client part of the system. At the first stage, validation of the acquired data is completed. Then, the formation of the system interface components based on the obtained JSON-array and placing of the components in screen forms takes place. The result is passed to the next process for the final

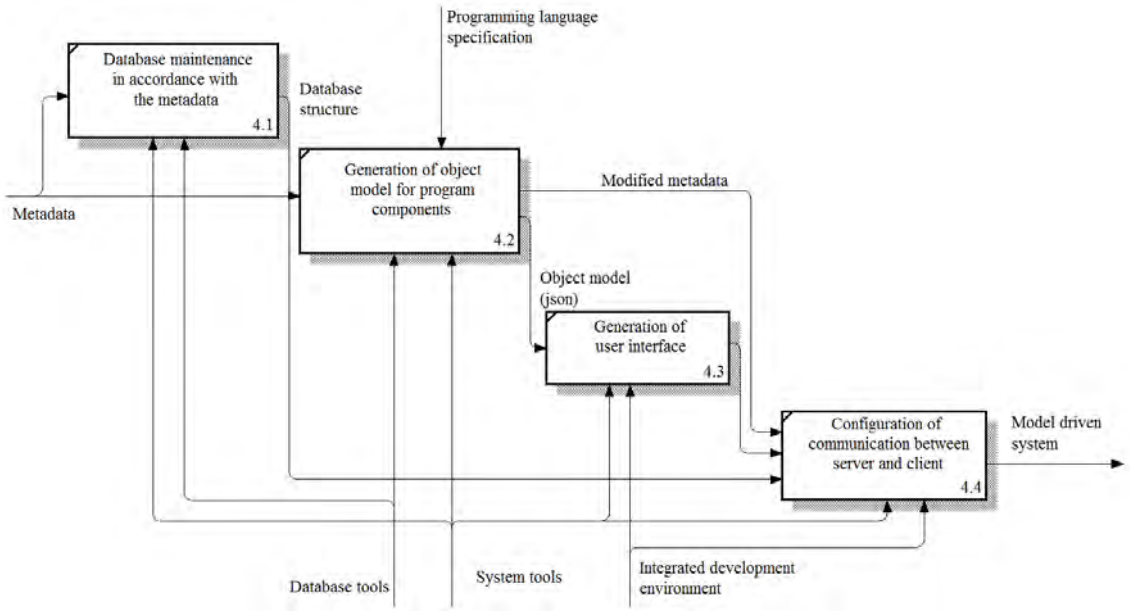


Figure 5. Creation of model-driven system

tuning of the system.

The process "Configuration of communication between server and client" produces the objects responsible for data exchange between the server and interface components, as well as sets the connection between them. The process output is a user-controlled model-driven system, which varies in accordance with the requirements of the domain area.

This approach to system development based on MDA allows to adopt to function condition changes due to integration of the control model into the system. Interface expansion based on specific control model tools allows to delegate the creation of system thematic content to moderators that increases the speed of the system respond to user's requirements.

V. EXPAMPLES OF APPLICATION

This original implementation of the model-driven approach to software development has been tested in ICM SB RAS: the data consolidation web-system for the Center of Emergency Monitoring and Prediction of Krasnoyarsk region and the system of scientific activity management have been created.

A. The data consolidation web-system for the Center of Emergency Monitoring and Prediction of Krasnoyarsk region

Figure 6 demonstrates the editor tables in data consolidation web-system that presents the functions of the process "Creation of the control and applied models" and result of the process "Model-driven system design". This interface is produced dynamically based on the applied model and it allows to collect information about emergencies.

The table editor is a mean of modification for control and applied models. Figure 6 shows that the editor contains information about the objects and attributes, the data describes the control and applied models. The editor is completed in one interface with the operator's and administrator's workplace. This approach significantly simplifies the process of system change. Operator's workplace is made dynamically on the basis of the applied model. It includes two parts, the left one presents the list of the objects available to the user, that are grouped as a tree. The right part is formed dynamically by choosing a table from the tree, each attribute of the chosen object matches with a certain column. The table form contains the previously collected information available for reading and editing. The editing form is dynamically changed when choosing the table in the tree.

The peculiarity of the emergency situation monitoring data consolidation web-system is the possibility to collect periodic information. The period is set for each form, it can be: one year, quarter, month or one day. When one chooses the form according to the current date, the filling period is calculated, and if there is information, it is shown to the user, providing the possibility to change it or add. Within this system, there also exists the module of synchronization with the monitoring center's base, where the collected data is analyzed, and the reports are provided to the managers in charge.

The data consolidation web-system is one of the most important information subsystems of the integrated system of data consolidation and analysis for emergency monitoring [7].

B. The system of scientific activity management

The system of scientific activity management has been developed with the use of the suggested approach. The system contains the editor of the control and applied models and operator's workplace. Besides, in order to simplify the work of operators, the system possesses the tools for data importing from external sources. For example, the information about publications is downloaded from library data bases (Russian Science Citation Index, Web of Science, etc.) The system implements the tools for analytical processing of data and report generating. The reports can be obtained for the whole institute, one of the departments or for a specific researcher.

The system of scientific activity management is

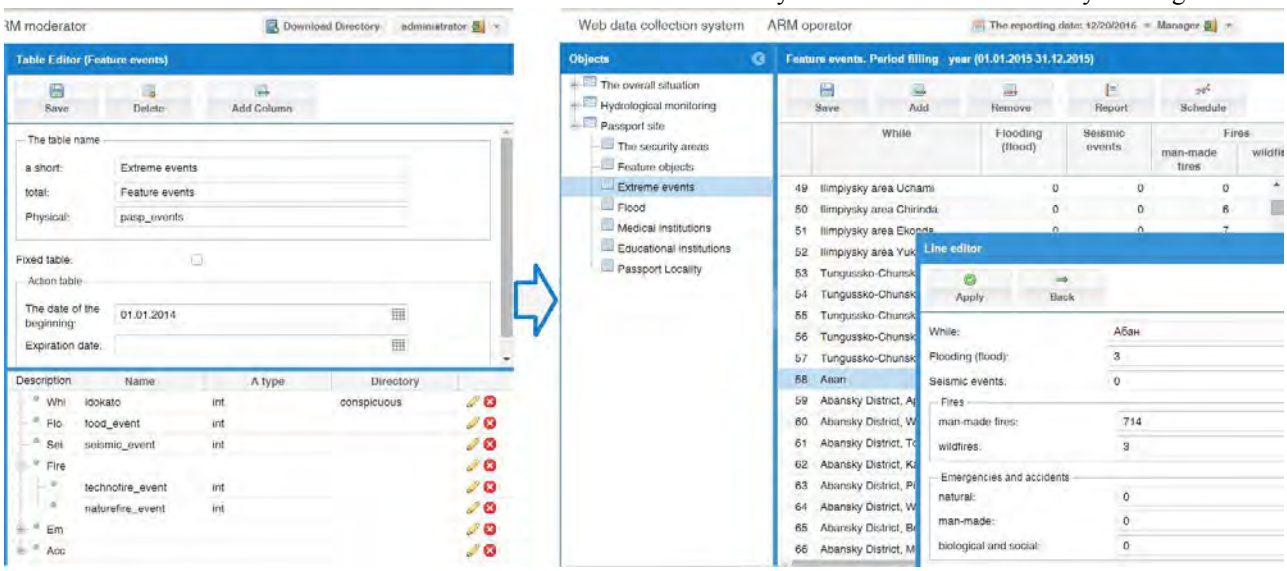


Figure 6. Creation of control and applied models (left) and dynamically produced interface (right)

successfully operated in the Institute of Computational Modeling of the Russian Academy of Science since 2011. The database of this system contains information about almost 2000 scientific objects grouped by basic subject sections: projects, grants, scientific event, patent activity, teaching activity, publication, partnership, etc.

VI. CONLUSION

This article presents an original approach to system development using MDA that allows user to change the thematic content in the model-driven systems. Application of the proposed approach provides a high level of adaptability of information system in accordance with user requirements.

This approach has been implemented in the Institute of Computational Modeling for creation of the scientific activity management system [8] and in the Center of Emergency Monitoring and Prediction of Krasnoyarsk region for creating the data consolidation web-system.

REFERENCES

[1] Object Management Group Model Driven Architecture (MDA), MDA Guide rev. 2.0, OMG Document ormsc/2014-06-01, http://www.omg.org/cgi-bin/doc?ormsc/14-06-01.

[2] OMG Meta Object Facility (MOF) Core Specification, ver. 2.5, OMG Document formal/2015-06-05, http://www.omg.org/spec/MOF/2.5/.

[3] Anneke Kleppe, Jos Warmer, Wim Bast, MDA Explained, The Model Driven Architecture: Practise and Promise, 2003, Addison-Wesley.

[4] Seidewitz E., What Models Mean, IEEE Software, 2003, vol. 20, no. 5, pp..26-32.

[5] Atkinson C., Kühne T., Model-Driven Development: A Metamodeling Foundation (Spec. issue on Model-Driven Development), IEEE Software, 2003, vol. 20, no. 5, pp..36-41.

[6] A.A. Korobko, A.V. Korobko, Model Driving for Scientific Activity Management System, Informatization and communication, 2015, no. 2, pp. 42-46.

[7] Aleksei A. Korobko, Anna V. Korobko, Constructing Domain model for Scientific Activity Management System, Modeling of Artificial Intelligence, 2015, vol. 6, is. 2, pp. 82-89.

[8] Penkova T.G., Korobko A.V., Nicheporchuk V.V., Nozhenkova L.F., On-line modelling and assessment of the state of technosphere and environment objects based on monitoring data, Procedia Computer Science, vol. 35, Elsevier, 2014, pp. 156-165.

Business Process Management Systems Selection Guiedelines: Theory and Practice

Vesna Bosilj Vukšić*, Ljiljana Brkić **, Mirta Baranović**
* Faculty of Economics and Business, University of Zagreb
Department of Informatics, Trg J.F.Kennedya 6, Zagreb, Croatia
** Faculty of Electrical Engineering and Computing, University of Zagreb
Department of Applied Computing, Unska 3, Zagreb, Croatia
vesna.bosilj@efzg.hr
ljiljana.brkic@fer.hr
mirta.baranovic@fer.hr

Abstract – Nowadays many business process management systems (BPMSs) are available on the market and it is quite difficult to select the appropriate one. This paper presents a set of criteria and proposes guidelines for BPMS selection. The proposal has been developed on the basis of literature overview and the feedback from practical experience in BPM projects. The results of this research are valuable for both academic and business community.

I. INTRODUCTION

In recent decades, business process management (BPM) is a research topic for which researchers have been showing increasing interest. Business process management systems (BPMSs) are integrated tools that enable business to perform the required steps in BPM initiatives. According to Khan [1] and Boots [2] in the literature, BPMS functionalities are systematized in a number of ways. The authors distinguish several categories, such as: BPM software, BPM suite, BPM system and BPM tool [3] [4] [5]. Obviously, an unambiguous definition and description is difficult to be expected, but the authors mostly refer to a variety of following description: BPMSs are software platforms that support the definition, execution, and tracking of business processes. For the purpose of this paper the term BPMS is used consistently to describe the previously mentioned categories (software, suite, system and tool).

BPM initially focused on IT aspects in business processes. Latter the holistic approach to BPM is introduced. According to this approach, BPM incorporates an IT focus into the strategic perspective of the business; moreover IT enables innovation instead of automatization [6] [7] [8] [9] [10]. However, selecting the BPMS is not any easy task since BPMS have different capabilities and each organization has specific requirements against the BPMS. Cardwell [11] lists several examples of the mismatches between what the business process require and what BPMS has provided: some BPM tools are too complicated and lack support [12]; a clear definition of requirements if very difficult for organization because of the communication gap between business and IT [13]; people work in highly complex ways that differ from person to person, and these individual work patterns cannot be easily modeled or captured by BPMS [14]. Obviously the problems exist and it is important to find the way how to avoid the obstacles in a case of BPMSs implementation.

A goal of this research is to propose BPMS selection guidelines and to illustrate the applicability of the proposal by establishing and by testing the BPMS evaluation criteria. Due to the complexity of this problem, three BPMS are evaluated and compared according to the previously selected and narrowed set of criteria. This work has been fully supported by Croatian Science Foundation under the project PROSPER - Process and Business Intelligence for Business Performance (IP-2014-09-3729). One of the objectives of the PROSPER project is to systemize, analyze and evaluate BPM commercial and open-source software functionalities in order to develop a BPMS selection guidelines.

The paper has been organized as follows. After the introduction, a preliminary overview of academic articles and professional reports in the field of BPMSs is presented and the theoretical background is given (Section II). The process of the analysis of the research results is presented in the Section III. The fourth part of the paper brings the theoretical background on the BPMSs. Three BPMSs are evaluated and the initial proposal of BPMS selection criteria is given (Section V). Finally, a short conclusion with the limitations and plans for future research is given.

II. THEORETICAL BACKGROUND

According to researchers [15] [4] the knowledge about BPMSs is still premature. Besides, the implementation of BPMSs if very often regarded as a standard software development project [16], while the critical success factors (CSFs) of BPMSs implementation are mainly neglected [17]. Many research studies have defined IT as one of many CSFs of BPM adoption [18] [10] [19]. Trkman [10] notices that organizations should consider different CSFs and not just focus on some of them. Therefore future research can be done to explore the field. During the last two decades, hundreds of vendors rushed to create BPMS products, and large companies are making major investments in BPM initiatives. According to the latest forecast from Gartner, Inc. worldwide spending on BPMS is set to grow 4.4

percent to reach $2.7 billion in 2015 [20]. Since nowadays there is a large number of BPMSs available on the market and each has its own specific objectives and characteristics, a systematic proposal or criteria or a guideline for selecting BPMS considering the needs of a given organization is required [21].

III. LITERATURE RESEARCH AND ANALYSIS

The development of BPM essentials framework in this paper have been based on a survey of academic literature. For this purpose, in February 2016 we checked Scopus and Web of Science (WoS) databases and focused on peer-reviewed articles in journal and conference publications. We searched for keywords 'business process management system' OR 'business process management software' OR 'business process management suite' OR 'business process management tool' in the abstract and in the keywords of the articles and all together received only 102 hits (76 in Scopus and 26 in WoS), among which 97 different articles. Our search strategy is presented in TABLE I.

TABLE I SCOPUS AND WOS SEARCH STRATEGY

Database	Search Strategy
Scopus	TITLE-ABS-KEY ("business process management system") OR TITLE-ABS-KEY ("business process management software") OR TITLE-ABS-KEY ("business process management suite") OR TITLE-ABS-KEY ("business process management tool") AND SUBJAREA (mult OR arts OR busi OR deci OR econ OR psyc OR soci) AND (EXCLUDE (DOCTYPE , "cr")) AND (EXCLUDE (SUBJAREA , "MATH")) AND (EXCLUDE (DOCTYPE , "ch") OR EXCLUDE (DOCTYPE , "bk"))
Web of Science (WoS)	TOPIC ("business process management system") OR TOPIC ("business process management software") OR TOPIC ("business process management suite") OR TOPIC ("business process management tool") AND Web of Science Categories (COMPUTER SCIENCE INFORMATION SYSTEMS OR COMPUTER SCIENCE THEORY METHODS OR COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE OF COMPUTER SCIENCE ARTIFICIAL ENGINEERING OR COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS) AND DOCUMENT TYPE (ARTICLE)

For each article the domain of research was registered, 4 types of research domains are used to categorize the articles. "BPM essentials" domain covers the issues that influence the success of BPMS implementation and therefore must be considered by BPM experts, top-managers and other participants of BPM initiatives. These are: BPMS capabilities (BPMS elements and functionalities, or BPMS architecture), a method that is used to evaluate and select a specific BPMS on the current BPMSs market and a set of CSFs that are relevant when implementing a BPMS in a specific organization. Research domains 2 (Methods, techniques and languages for BPMS development) and 3 (BPMS integration frameworks and IT for extension of BPMSs) are not in a focus of this research. These issues are of high relevance for the further evolution and expansion of the BPMS and therefore these domains are usually investigated and explored by academics, BPMS developers, software vendors, and IT integration specialist. The fourth domain (BPMS case studies) comprises "the success stories" of BPMS implementation in a specific companies, industries or countries. The findings of case studies could be of interest for the further research in order to determine whether CSFs are different depending on the type of organization (e.g. in specific sectors).

After reviewing the abstracts of all 97 articles we eliminated 3 articles, which were not relevant for this research. Article was considered relevant, if it specifically covers one of four BPMS domains. Accordingly, from the 94 remaining articles we identified the most important research results and gathered them into four domains (TABLE II). According to the discussion given above, the papers which were classified under the first domain (BPMS essentials) are further assigned into one of three subdomains (1.1 – 1.3).

The list of 19 articles that examine BPMS essentials is presented in APPENDIX I. The results of research about BPMS capabilities are found in 8 articles, there are 8 articles about BPMS evaluation and/or selection methods and guidelines, while CSFs of BPMS implementation are discussed in only 3 articles. 75 articles were not relevant for this research and thus are not referred here. Generally, a small number of articles about BPMS that have been found confirmed the previous assumption about the need for further insights into the field.

IV. BPMS ESSENTIALS: ANALYSIS AND SYSTEMATIZATION

In BPM capabilities subdomain (1.1) the most comprehensive work is done by Margherita [22] and Shaw et al. [23]. According to Margherita [22] a BPMS consists of four subsystems: (1) process strategy; (2) process (3) process execution; and (4) process performance. The process strategy subsystem provides the link among the core processes and the overall business goals of organization. It includes market and industry value drivers; business capabilities; culture values and beliefs; and company balanced scorecard.

TABLE II SCOPUS AND WOS SEARCH RESULTS

No.	Domain		No. of hits: Scopus	No. of hits: Wos	No. of different hits
1	BPMS essentials	1.1. BPMS capabilities	6	3	8
		1.2. BPMS evaluation and/or selection methods and guidelines	8	1	8
		1.3. BPMS implementation CSFs	3	0	3
		Number of articles in domain 1	17	4	19
2	Methods, techniques and languages for BPMS development		21	7	26
3	BPMS integration frameworks and IT for extension of BPMSs		22	9	30
4	BPMS case studies		14	5	19
	Number of relevant articles		74	25	94
	Number of eliminated articles		2	1	3
	Total		76	26	97

This subsystem includes value chain model; process models and simulation; stakeholders and process roles; and business rules and governance. The process execution subsystem provides the effective and efficient collaboration/communication; workflows coordination; IT application; and process automation. The process performance subsystem includes key performance indicator (KPI) and performance monitoring; process control dashboard; flows and networks analysis; and business performance links. On the other hand, Shaw et al. [23] theoretically explained and empirically illustrated BPMS architecture's core technologies that span numerous disciplines. They proposed a pyramid architecture of BPMS that is based on two pillars. The first is the subject that is modeled (e.g. the actual business processes managed by BPMS). It consists of formal model constructs; formal modeling notation; ontology-based modelling grammar; and model abstraction capability. The second pillar is the relating IT, it comprises several capabilities, such as: software application; technical infrastructure; software language; software notation and software grammar; and software formalism. At the top of both pillars are "enactable models". These are models that can be executed and thus become active. According to Shaw et al. [23] only an enactable business process model gives a BPMS the ability to automatically manage business processes. These two articles propose very sophisticated and theoretically based frameworks for BPMS evaluation that can be used to propose a prototype of BPMS selection guidelines. The analysis of articles within the sub domain 1.2 (BPMS evaluation and/or selection methods and guidelines) shows that the authors propose different evaluation methods, frameworks and taxonomies to select a suitable BPMS. Besides, some authors demonstrate different methodologies that may be of interest to evaluate specific capabilities of BPMS. Most of them emphasize that evaluation and selection of BPMSs is complicated, cost and time consuming process, while the success of the selection depends on the specific needs of each organization. Among the CSFs for BPMS implementation the most often mentioned are situational factors, such as a cultural background, BPM maturity level and process orientation of a company. The authors of the articles within the subdomain 1.3 (BPMS implementation CSFs) agree that "a one-size fits all" approach cannot ensure the success of BPMS implementation.

It is also important to take into account a feedback from business practice. Boots [2] compiles BPMS selection criteria that are provided by the analysts and consultants at Gartner, Forester and BPTrends. According to Boots [2] a starting point list for BPMS evaluation comprises 14 criteria: vendor basics; .NET of Java coding; network and server requirements; database technology; scalability and performance; installation and configuration; functionality scope; administrator role; solution developer role; user interface and solution usability; marketplace success; software subscription model options and costs; pilot project options and vendor support. Nowadays the basic components of the "traditional" BPM software (process modeler; data modelling environment; visual forms editor; process engine; and rules engine) are enhanced. Gartner [24] introduced a new term - the intelligent business process management suite (iBPMS). This so-called "next generation of BPMS" supports advanced analytics (big data, real-time intelligence and predictive analytics); real-time decision management; human collaboration support and integration with social media; cloud and mobile access to processes. Therefore the set of additional features must be taken into account while evaluating BPMSs.

V. BPMS SELECTION GUIDELINES: PROPOSAL AND EVALUATION

The aim of this section is to provide guidelines for BPMS selection. A set of evaluation criteria is based on the findings about BPMS capabilities that are showed and discussed in Section IV. A process of selection consists of three steps: (1) BPMS market analysis and evaluation of BPMS vendors; (2) BPMS capabilities evaluation; and (3) identification and evaluation of BPMS implementation CSFs considering the needs of a given organization. The last step is out of the scope of this research and will not be further discussed. In order to illustrate the applicability of the proposal, three BPMSs are evaluated and ranked.

At the first step, the Gartner report [24] is used to identify the candidates from 3 categories: Leaders, Visionars and Niche players (TABLE III). The fourth category ("Challengers") in this Gartner report remained empty. All evaluated BPMS vendors also appear in Forester's report [25] which makes them appropriate for evaluation.

TABLE III EVALUATED BPMSs

Vendor	Prooduct evaluated	Gartner category [3]
IBM	IBM Business Process Manager	Leader
Software AG	webMethods BPM	Visionar
K2	K2 blackpearl	Niche player

The criteria for selecting BPMS vendors is defined as follow: (1) **maturity** – the vendor's maturity and the maturity of BPMS must be satisfactory relative to the maturity of the overall IT area; (2) **geographic presence** – a vendor must have a solid customer portfolio in BPMS domain across America, Europe and Asia; (3) **revenue** – a vendor must have a significant BPMS-related revenue (according to [25] each evaluated vendor has revenue over $1 billion); and (4) **applicability in different business branches** – vendor must have BPMS-related implementations in at least tree business branches (i.e. sectors). Since the selected BPMS vendors satisfy the initial requirements, their BPM products are further analyzed.

IBM offers many features in the sense of BPM and process oriented application development. IBM Business Process Manager [26] includes process design, process execution, discrete event simulation, business activity monitoring and process optimization functionality. It provides process oriented applications development and cloud-based deployment. Also includes process portal web interface for executable business processes and collaboration with other process users and experts. Compared to other evaluated BPM suites, IBM Business Process Manager is tough to configure and deploy. This is because applications usually depend upon several different products and the setting of many parameters and options.

The situation is caused by the overlapping of IBM products in business rule processing, business activity monitoring, portal and other capabilities. IBM Business Process Manager requires high level of skills, and people having those skills are expensive and hard to find.

Software AG has a various components for covering the diverse requirements of modern BPMS. Vendor incorporates numerous technologies and products coming from companies acquired in a period of last few years. Similar to IBM, some products overlap and make these difficult for implementation. The process modelling and analysis environment supports various process modelling standards and notations, including Even-driven Process Chain (EPC) and Business Process Modelling Notation (BPMN). This evaluation is focused on Agile Process Platform consisting of webMethods BPM and webMethods AgileApps [27]. Software AG's Agile Process Platform is a unified offering designed to create process-driven software applications based on end-to-end business processes or workflows. webMethods BPM ARIS is used to model business process that can be subsequently executed with webMethods AgileApps. WebMethods AgileApps is an environment for developing process oriented applications that works with all mobile, desktop and tablet devices.

K2 provides platform for quickly building process applications through visual drag-and-drop design environments. Thanks to low-code principle, K2 is known for its fast development and minimal training required to start building and deploying solutions. In the past K2 has been tightly associated with Microsoft SharePoint but now it is widely used across different scenarios and platforms. According to customer's experience, K2 is well suited for small departmental workflow but is not so god for system workflows that involve data transformation, manipulation and exchange between systems. It can be integrated with multiple systems such as Exchange, Dynamics CRM, Microsoft SharePoint, SAP, Salesforce and any custom in house application with supported database layer. For the purpose of this research K2 blackpearl product [28] is examined. It has a pretty steep learning curve, meaning it is suitable for people who do not have a technical or programming background.

All selected vendors are able to produce software applications suitable for desktop, laptop, smartphone and tablet devices; supports cloud development and installation practices; and to some extent facilitate the development of software applications in terms of low-code. TABLE IV shows some basic technical characteristics of examined BPMSs.

TABLE IV BASIC TECHNICAL CHARACTERISTICS

Product	Modeling notation	Platform/OS	Out-of-the-box supported RDMS
IBM Business Process Manager	BPMN	Windows , Linux	DB2, SQL Server, Oracle JDBC + ODBC-
K2 blackpearl	Share Point workflow	Windows	SQL Server, Oracle
webMethods	EPC, BPMN	Windows (and Linux, Mac unofficially)	Oracle

At the second step, an initial (a narrower) set of criteria for the evaluation of BPMS capabilities is established.

Evaluation categories are listed in TABLE V as column's labels. **Process modelling and design**: refers to the ability to design organization's current (or "as-is") process and improved (or "to-be") process version; possibility of collaborative modelling as well as process flow design. **Business rules**: refers to the degree of complexity of business rules that can be supported in the design phase of business process; defining business rules in the design phase (using graphical interface) of a business process centralizes business process logic, eliminates multiple definitions of the same rule and separates the rule's logic from other application code; a rule engine is usually embedded in BPM suite or tightly integrated, or sometimes sourced from another vendor. **User interface and user experience**: this category includes two aspects: (1) the ease of use of the suite's components for modelling, simulation, testing and executing of business processes; and (2) the level of complexity and ease of use of the user interface that can be developed within the suite. **Mobile and tablet**: the degree of supported mobile and tablet features. **Cloud capabilities**: the possibility of cloud-based development and deployment. For example, one of the evaluated suites (IBM) has cloud-based environment that supports complete model-to-execution cycle. **Low-code development**: this principle is typically, to some extent, embedded in the BPM suites; low-code platforms minimize hand-coding and speed up software development by providing visual tools as a substitution to hand-coding; generally, low-code platforms require less-skilled developers which subsequently makes solutions cheaper. **Reporting and analytics**: the ability to create advanced reports and analyses based on data collected during the simulation and process runtime; BPM suite can gather data for reporting through process oriented applications (from form fields, validation mechanisms, etc.) and process definitions (human and system interactions, responses, deadlines, etc.); BPM suites can offer predefined out-of-the-box reports or ad-hoc reporting; depending on the BPM product, data collected within the suite, can be combined with enterprise's other system's data to create e.g. drillable dashboards. **Ease of implementation**: the overall impression of simplicity of implementation including the level of skills required and learning curve.

To evaluate the vendors and their products against this set of criteria, the details of BPMSs capabilities are gathered through a combination of lab evaluations, demos, product documentation and reports of eminent market research and advisory firms (emphasizing the Gartner's [29], Forrester's [25] and Ovum's [30] reports). The authors did not conduct any actual product testing.

Considering the complexity of the BPMS market, the comparisons suggesting that some suites are the best while others are "undesirable" are avoided. Instead, each product is evaluated in terms of established set of criteria. Besides, the "ease of adoption" is estimated for each of these products (TABLE V). The ratings are expressed through three descriptive categories: Low, Medium and High. As can been seen in TABLE V, Low rating does not appear in any category, which is expected since evaluated BPMSs are highly positioned in the relevant reports [25] [29] [30].

TABLE V BPM SUITE'S SCORE

Vendor	Process modelling and design	Business rules	User interface and user experience	Mobile & tablet	Cloud capabilities	Low-code development	Reporting & analytics	Ease of implementation
IBM	High	High	Medium	Medium	High	Medium	High	Medium
K2	Medium	Medium	High	Medium	High	High	Medium	High
Software AG	High	High	Medium	High	Medium	High	High	Medium

High scores in "process modelling and design" category shows that Software AG and IBM have made efforts to enhance the user experience of their process modelling and design environments. An easy and intuitive development environment makes the products more usable to nontechnical business analysts. These two vendors have shown to be superior in terms of business rules expressiveness also. K2 also provides modestly possibilities of reporting and analyzing the data collected during the execution process. Its comparative advantage over the other two suites is simpler and therefore faster implementation.

VI. CONCLUSION

The results of this research support the premise that a complex set of criteria must be taken into account while evaluating BPMSs. Without the theoretical foundation companies and their managers are often left with no guidance while making decisions on BPMS acquisition. The proposed set of selection criteria is intentionally quite narrow and general, so it can be applied to different organizations from different industries. On the other hand, the literature overview showed that the success of BPMS selection process originates in identifying specific CSFs that may vary from case to case. Therefore specific selection criteria must be added in order to assure the applicability of guidelines in specific cases. The paper has several limitations and the guidelines should be improved through further research. The set of selection criteria must be recognized and agreed upon by the business users, managers and BPM vendors. An extensive quantitative validation is needed that should be done on a broad scale. Also in-depth studies are needed to determine the factors that depend on the type of organization and its environment.

REFERENCES

[1] R. N. Khan, "Business Process Management: a practical guide," in *Business Process Management: a practical guide*, Tampa: Meghan-Kiffer Press, 2003.

[2] J. Boots, BPM Boots on the Ground: How to Implement Strategic Business Process Management: Lessons Learned from one of the World's Largest Organizations, Morgan-Kiffer Press, 2012.

[3] J. N. Gilot, he Complete Guide to Business Process Management: Business process transformation or a way of aligning the strategic objectives of the company and the information system through the processes, Joel-Noel Gillot, 2008.

[4] P. Harmon and T. Davenport, Business Process Change A Guide for Business Managers and BPM and Six Sigma Professionals, Morgan Kaufmann, 2007.

[5] M. Dumas, M. La Rosa, J. Mendling and H. A. Reijers, Fundamentals of Business Process Management, Springer, 2013.

[6] I. DeToro and T. McCabe, How to stay flexible and elude fads, Quality Progress,, 1997, pp. 55-60.

[7] Gartner, "Delivering IT's contribution: The 2005 CIO agenda," Stamford, CT: Gartner, 2005.

[8] A. Groznik, A. Kovačić and P. Trkman, "The Role of Business Renovation and Informatization in Egovernment," *Journal of Computer Information Systems,* vol. 49, no. 1, pp. 80-88, 2008.

[9] P. Trkman, M. Indihar Štemberger, J. Jaklič and A. Groznik, "Process approach to supply chain integration," *Supply Chain Management - An International Journal,* vol. 12, no. 2, pp. 116-128, 2007.

[10] P. Trkman, "The critical success factors of business process management," *International Journal of Information Management,* vol. 30, no. 2, pp. 125-134, 2010.

[11] G. Cardwell, "The influence of Enterprise Architecture and process hierarchies on company success," *Total Quality Management & Business Excellence,* vol. 19, no. 1-2, pp. 47-55, 2008.

[12] J. Sinur, "The power vendors need to demonstrate credibility, and that will only come as the market matures," Gartner Inc, in: Robert Westervelt, News Editor, 2006.

[13] DM Review Staff, "New survey reveals chasm between perception and reality when it comes to software requirements management," DM review online, 2006.

[14] R. Khan, "BPM: A global view – it's all about the person in the process, stupid," BPTrends, Newsletter, 2006.

[15] P. Ravesteyn and R. Batenburg, "Surveying critical success factors of BPM-systems implementation," *Business Process Management Journal,* vol. 16, no. 3, pp. 492-507, 2010.

[16] D. Krafzig, K. Banke and D. Slama, Enterprise SOA: Service Oriented Architecture Best Practices, Prentice-Hall PTR, 2004.

[17] P. Ravesteyn and J. Versendaal, "Success Factors of Business Process Management Systems Implementation," in *Proceedings of 18th Australasian Conference on Information Systems*, Toowoomba, 2007.

[18] W. Bandara, A. Alibabaei and M. Aghdasi, "Means of achieving business process management success factors," in *Proceedings of the 4th Mediterranean conference on information systems*, Athens, 2009.

[19] K. J. Ringim, M. R. Razalli and N. Hasnan, "Critical success factors for business process management for small and medium banks in Nigeria," *Business and Management Review,* pp. 83-91, 2012.

[20] Gartner, "Gartner Says Spending on Business Process Management Suites to Reach $2.7 Billion in 2015 as Organizations Digitalize Processes," 2015. [Online]. Available: http://www.gartner.com/newsroom/id/3064717.

[21] L. Camara Silva, T. Poleto, V. D. Heuer de Carvalho and A. P. Cabral Seixa Costa , "Selection of a Business Process Management System: An Analysis based on a Multicriteria Problem," in *2014 IEEE International Conference on Systems, Man, and Cybernetics*, San Diego, CA, USA, 2014.

[22] A. Margherita, "Business process management system and activities: Two integrative definitions to build an operational body of knowledge," *Business Process Management Journal,* vol. 20, no. 5, pp. 642-662, 2014.

[23] D. R. Shaw, C. P. Holland, P. Kawalek, B. Snowdon and B. Warboys, "Elements of a business process management system:

Theory and practice," *Business Process Management Journal,* vol. 13, no. 1, pp. 91-107, 2007.

[24] Gartner, "Gartner Says Intelligent Business Operations Is the Next Step for BPM Progams," March 2012. [Online]. Available: http://www.gartner.com/newsroom/id/1943514.

[25] C. Richardson, "The Forrester Wave™: BPM Platforms For Digital Business, Q4 2015," Forrester, November 18, 2015.

[26] IMB, "IBM Business Process Manager V8.5.6 documentation," [Online]. Available: http://www-01.ibm.com/support/knowledgecenter/SSFPJS_8.5.6/com.ibm.wbpm.main.doc/kc-homepage-bpm.html.

[27] Software AG, "webMethods BPM," Software AG, [Online]. Available: http://www.softwareag.com/corporate/products/webmethods_process/bpm/overview/default.asp.

[28] K2, "K2 blackbeapl," K2, [Online]. Available: http://www.k2.com/products-blackpearl-appit.

[29] R. Dunie, R. Schutle, M. Cantara and M. Kerremans, "Magic Quadrant for intelligent Business Process Management Suites," Gartner, 2015, 18 March.

[30] G. Barnet and M. Holt, "Ovum Decision Matrix: Selecting a Business Process Management Solution," 2014.

APPENDIX I LIST OF SELECTED ARTICLES WITHIN THE DOMAIN 1 (BPMS ESSENTIALS)

No	BPMS capabilities	Database
1	Junginger, S.,Kuhn, H., Strobl, R., Karagiannis, D., A next-generation business process management tool - ADONIS: concepts and applications, WIRTSCHAFTSINFORMATIK Volume: 42 Issue: 5 Pages: 392, 2000.	WoS
2	Margherita, A., Business process management system and activities: Two integrative definitions to build an operational body of knowledge, Business Process Management Journal, 20 (5), pp. 642-662, 2014.	Scopus
3	Pacico, L., Ravarini, A.M.. Pigni, FBusiness process modelling within the cycle of continuous improvement, Information Systems: People, Organizations, Institutions, and Technologies - ItAIS: The Italian Association for Information Systems, pp. 91-99, 2010.	Scopus
4	Schulte, S., Janiesch, C., Venugopal, S., Weber, I., Hoenisch, P., Elastic Business Process Management: State of the art and open challenges for BPM in the cloud, Future Generation Computer Systems - The International Journal Of Grid Computing And E-science, Volume: 46 Pages: 36-50, 2015.	WoS
5	Shaw, D.R., Holland, C.P., Kawalek, P., Snowdon, B., Warboys, B., Elements of a business process management system: Theory and practice, Business Process Management Journal, 13 (1), pp. 91-107, 2007.	Scopus
6	Tian, G.-S., Quan, L., An improved framework of business process management system which integrating the strategy management, 2008 International Conference on Management Science and Engineering 15th Annual Conference Proceedings, ICMSE, art. no. 4668925, pp. 256-261, 2008.	Scopus
7	Vergidis, K., Turner, C.J., Tiwari, A., Business process perspectives: Theoretical developments vs. real-world practice, International Journal of Production Economics, 114 (1), pp. 91-104, 2008.	Scopus and WoS
8	Wang, M., Kumar, K.Developing flexible business process management systems using modular computing, technologies, Global Journal of Flexible Systems Management, 10 (1), pp. 1-10, 2009.	Scopus
	BPMS evaluation and/or selection methods and guidelines	
1	Bernroider, E.W.N., Bernroider, M., A comparative study of business process management tools based on open source software and a commercial reference, IMETI 2008 - International Multi-Conference on Engineering and Technological Innovation, Proceedings, 1, pp. 225-230, 2008.	Scopus
2	Blumberg, E.J., Kelly, A.C., Olmstead, R., Youmans, R.J, Usability analysis of a business process management software platform, Proceedings of the Human Factors and Ergonomics Society, pp. 2071-2075, 2013.	Scopus
3	Cingil, I., Ozturan, M., Erdem, A.S., A Decision Support System for evaluation of Business Process Management systems, Information, 15 (2), pp. 537-556, 2012.	Scopus
4	Delgado, Calegari, D., Milanese, P., Falcon, R., García, E., A systematic approach for evaluating BPM systems: Case studies on open source and proprietary tools, IFIP Advances in Information and Communication Technology, 451, pp. 81-90, 2015.	Scopus
5	Mejri, A., Ghanouchi, S.A., Evaluation of paradigms enabling flexibility: BPMSs comparative study, ICEIS 2015 - 17th International Conference on Enterprise Information Systems, Proceedings, 3, pp. 291-298, 2015.	Scopus
6	Poelmans, S., Reijers, H.A., Recker, J., Investigating the success of operational business process management systems, Information Technology and Management, 14 (4), pp. 295-314, 2013.	Scopus
7	Ravasan, A.Z., Rouhani, S., Hamidi, H., A practical framework for business process management suites selection using fuzzy TOPSIS approach, ICEIS 2014 - Proceedings of the 16th International Conference on Enterprise Information Systems, 3, pp. 295-302, 2014.	Scopus
8	Štemberger, M.I., Bosilj-Vukšić, V., Jaklič, J., Business process management software selection - Two case studies, Ekonomska Istrazivanja, 22 (4), pp. 84-99, 2009.	Scopus and WoS
	BPMS implementation CSFs	
1	Ravesteyn, P., Batenburg, R., Cultural differences in implementing business process management systems, 16th Americas Conference on Information Systems 2010, AMCIS 2010, 6, pp. 4106-4115, 2010.	Scopus
2	Ravesteyn, P., Jansen, S., A situational implementation method for business process management systems, 15th Americas Conference on Information Systems 2009, AMCIS 2009, 8, pp. 5451-5459, 2009.	Scopus
3	Reijers, H.A., Implementing BPM systems: The role of process orientation, Business Process Management Journal, 12 (4), pp. 389-409, 2006.	Scopus

Organization of tax data warehouse for legal entities

Miodrag Sretenović* Božidar Kovačić** Vladan Jovanović***
* MI&DA d.o.o., Karlovac, Hrvatska, m.sretenovic@mida.hr
** Odjel za informatiku, Sveučilište u Rijeci, Hrvatska, bkovacic@inf.uniri.hr
*** Georgia Southern University, Statesboro, USA, vladan@georgiasouthern.edu

Abstract - Legal entities information systems for tracking tax changes belong to the group of business office information systems, which provide answers to the user requirements on the monitoring of tax changes. By its structure, such information system represents only one segment of the overall legal entities information system, but despite this it has a central role in reporting to the Tax authorities and it represents the link between legal entities and the Tax authorities. This paper describes a model for storing tax data that has been implemented in practical business applications. Furthermore it describes the process of building and organization of the data warehouse and implemented ETL process for the extraction, cleansing and transformation of data in a multidimensional model of tax data warehouses, which is based on the metadata model and model of the set of legal rules. In addition it states a concept and basic methods used in the development of business intelligence systems for the construction and use of a warehouse and it gives guidelines for future development. Model warehouse tax information forms the basis for the execution of various analysis for decision making. Formalizing the process performance analysis directly affects the quality of business intelligence.

I. INTRODUCTION

Laws in the field of tax policy regulate the relationship between taxpayers and tax authorities implementing tax regulations. Due to increasing needs of public finance, tax laws are subject to frequent change. But despite these changes, such laws also require storing of historical tax data for several years.

A. Business Intelligence

Business intelligence is the process of collecting data and information from internal and external business environments and their transformation into business knowledge used for making business decisions[1].

B. Data Warehouse

Data warehouse is a segment of business intelligence, provided that the data within the warehouse are filtered. There are several ways of using filtered data. For example, by making inquiries that cannot be made in transactional databases because data are unstructured or semi-structured, or because inquiries could slow down the usual operation of databases. Data warehouse represents a central source of previously cleaned and transformed data, which are used for providing support in making business decisions[2].

Data warehouse is a set of procedures, technologies and tools that enable end users to perform data analyses, which help in the decision-making process and improve information resources[3].

Tax data warehouse represents a source of tax data that are filtered from the database by applying rules and processes based on tax legislation in a form suitable for analysis and creation of tax reports.

C. Data Mart

Data mart is a data warehouse that covers one segment of business, and includes data related to a single topic of business[4]. One of the leading experts in the field of application and implementation of data marts is Ralph Kimball, who founded Red Brick Systems in 1986[5].

Kimball's model [6] was used as a model for designing the model of tax data warehouse. In the dimensional approach of warehouse architecture, adopted by Kimball's model [6], data are divided into fact table, which is the largest one with respect to the amount of data; and this table is connected to one or more dimension tables, which contain attributes, and with the application of adequate tools, enables multidimensional view to the user, so we can call such structures multidimensional or "hypercube". Within the tax data warehouse, dimensions are independent by their structure, so they can be shown as an n-dimensional coordinate system.

By applying the relational structure of databases, access to data warehouse has been optimized, which results in a significantly improved performance with respect to the former technological level of databases.

D. Data Warehouse Architecture

Data warehouse architecture includes the process of defining data warehouse elements, methods for creating and connecting warehouse elements in accordance with set requirements. Data warehouse architecture represents a formal description of a data warehouse system[7].

The reasons that contributed to the architecture and development of data warehouses was the inability of databases to support complex requests for processing in a reliable and safe manner, and an increasing need for analysing a large amount of data and information.

This paper describes the development of tax data warehouses in chapter two, and presents the concept of tax data warehouses in chapter three. Chapter four contains description of a data warehouse model, the model implementation is presented in chapter five, and conclusion in chapter six.

II. DEVELOPMENT OF TAX DATA WAREHOUSE

Tax data warehouse must be able to quickly adapt to potential changes in legislation, so agile methodology is imperative when it comes to the development of data warehouse. Agile methodology applied in software development enables fast adjustment to changes in the environment[8].

Therefore, modelling requires:

- Conceptual modelling of tax data,
- Business process modelling according to ETL (Extraction, Transformation and Loading) process elements:
  - data extraction process,
  - data cleansing process,
  - data transformation process,
- Modelling of data marts within tax data warehouses,
- Enabling the presentation of data on prescribed tax forms, and enabling flexible access to all relevant data, in order to allow system users to further analyse tax changes.

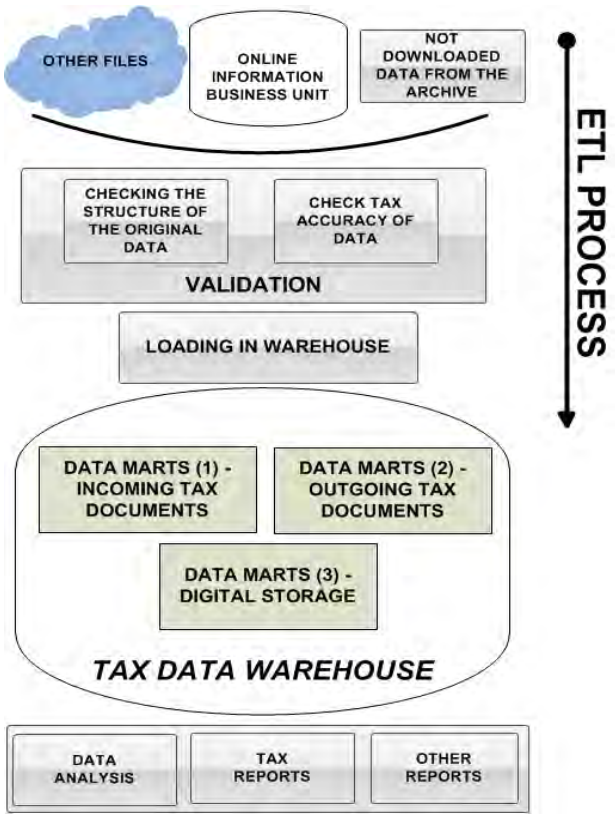


Figure 1. Shows the process of building a data warehouse

A. Conceptual Modelling of Tax Data

Conceptual modelling of data prescribed by tax legislation is a highly abstracted process of data modelling. The structure of data within the information system for tracking tax changes must reflect the structure of prescribed abstract reports.

The conceptual data model is a complete, consistent and concise description of data within an information system[9].

B. ETL Process Modelling

The process of building a tax data warehouse is shown in Figure 1 and starts by defining ETS processes (Extract, Transform, Load).

- The first step includes extracting the data from data sources. The most common data sources are relational databases, with the exception of tax data stored in various file formats generated from incompatible applications. In the process of data extraction, it is crucial to trace the source of data, date and time of their entry and data on entry operators, in order to enable complete insight to auditors. Data Vault concept emphasises the need to record where the data came from, as well as the time they were stored[10]. At the moment ETL tools pull data from relational databases, they pull metadata about their structure. Data structure in relational databases must reflect the final, abstract and prescribed form of report. Therefore, we may say that the tax data warehouse is based on the model of metadata and the model of prescribed legal rules.
- Transform means the transformation of data from the source to the warehouse. The preparations for loading data include an abstracted process of preparation and transformation of format and structure, validating the accuracy and quality of source data. The abstracted process within the tax data warehouse is conditioned primarily by regulatory nomenclature. Therefore, it requires an automatic process of cleansing and removing useless, incomplete or inaccurate data and returning them to the source for reprocessing.
- Load is a process of loading transformed data into the warehouse. These data will be suitable for further processing and enabling access and creation of prescribed tax documents, which is the ultimate aim of the user.

Legal regulations regarding the manner and structure of reporting to tax authorities directly prescribe the obligation to periodically load the data into the tax data warehouse, in intervals of at least 30 days.

C. Modelling of Data Marts within Tax Data Warehouse

Data marts within the tax data warehouse have specific structure, and they are optimized for the field of regulatory requirements for users. They are defined as abstracted reports and do not allow users much creativity or design.

Each tax data warehouse contains multiple data marts, which are coordinated and synchronised, and can be classified as supermarts according to their structure.

Supermarts consist of multiple mutually coordinated and synchronised tables. Table synchronisation is performed using foreign keys coming from conformed dimension. The structure of reporting the course of tax changes to tax authorities requires designing of at least three data marts, which are mutually connected by common dimensions and dimension attributes. Based on this fact, we have modelled the following:

1. data mart containing outgoing tax data or tax liabilities. Outgoing tax documents are invoices issued to customers for performed services or delivered goods.
2. data mart containing incoming tax data or tax receivables. Incoming tax documents are invoices received by suppliers for received services or goods.
3. data mart containing digital tax documents.Data mart containing digital tax documents includes: digital incoming invoices, digital outgoing invoices, digital documents by tax authorities, digital documents submitted to tax authorities, as well as any other tax-related digital documents.

Data marts have shared dimensions (Taxable period, Business units, Type of tax document, Entry operators, Invoice items, Tax partners, Date) and their own dimension attributes. Each dimension table has its surrogate key as a unique identifier of history of dimension data.

D. Modelling of Reports Within Tax Data Warehouse

Tax reports are abstracted reports that must contain elements prescribed by law. Since this is the ultimate aim of the tax data warehouse, it is necessary to perform optimization of all processes and thoroughly examine sources of data that will be used for modelling reports.

III. CONCEPT OF TAX DATA WAREHOUSE

Each change in tax regulations results in a change in the structure of databases at the source. Such changes increase requirements towards the data warehouse with respect to storing data and all previous changes, including the structure of data and metadata, covering a longer period of time. Figures 2, 3 and 4 show Entity-Relationship Diagram (ERD) of data marts, according to Martin Entity-Relationship Diagram[11].

Data Marts are used for the following purposes:

1. Data mart input tax data for the storage of incoming tax documents.
2. Data mart output tax data for the storage of output tax documents.
3. Data mart digital archive for the storage of digitized tax documents.

Entity is anything we wish to collect and store information about, entity-relationship method is a diagram of mutually connected groups of data within a system in question [12].

A. Defining the Structure of Data Marts Containing Incoming Tax Data

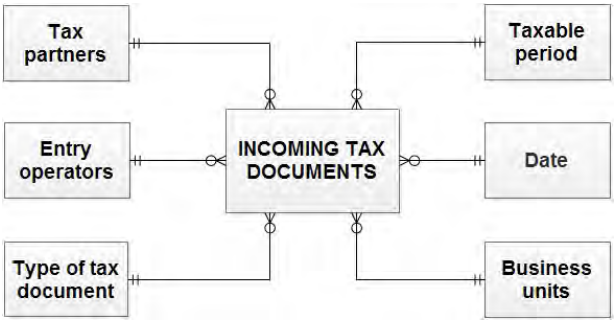


Figure 2. Data mart containing incoming tax data

Figure 2 shows the structure of a data mart containing incoming tax data. The structure of tax reports for the purpose of reporting the amounts and structure of tax receivables to tax authorities is defined by the data mart containing incoming tax data. Data mart structure diagram is shown in figure 2, where the entity INCOMING TAX DOCUMENTS is connected to other entities, which are partly shared with other data marts.

B. Defining the Structure of Data Marts Containing Outgoing Tax Data

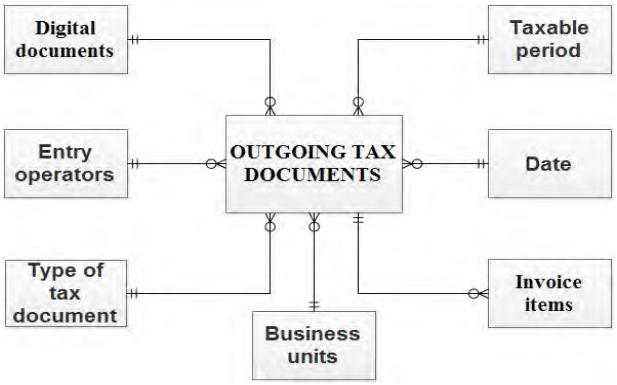


Figure 3. Data mart containing outgoing tax data

Figure 3 shows the structure of a data mart containing outgoing tax data. The structure of tax reports for the purpose of reporting the amounts and structure of tax liabilities to tax authorities is defined by the data mart containing outgoing tax data. Data mart structure diagram is shown in figure 3, where the entity OUTGOING TAX DOCUMENTS is connected to other entities, which are partly shared with other data marts.

C. Defining the Structure of Digital Storage Data Mart

Figure 4 shows the structure of digital storage data mart. Designing a digital storage data mart for storing tax documents in digital form enables fast and accurate access to digital tax documents. Figure 4 shows data mart structure diagram, where the entity DIGITAL STORAGE is connected to other entities, which are partly shared with other data marts.

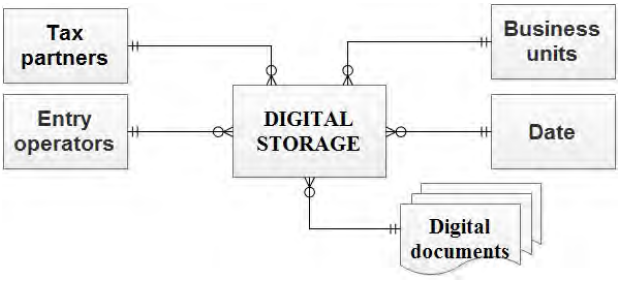


Figure 4. Digital storage data mart

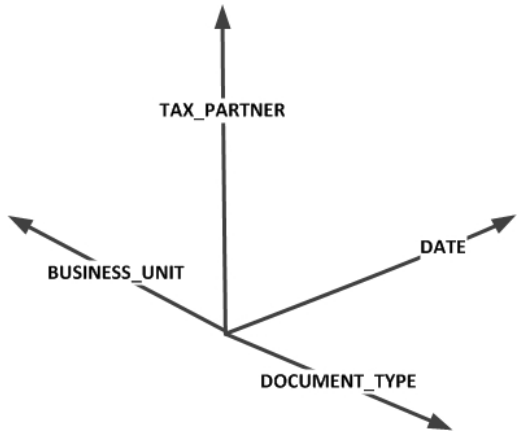


Figure 5. Four dimensional digital archives

The model allows you to perform analysis based on various criteria, for example in Figure 5.

The purpose of verification of tax data. The analysis presents a logical sequence of tax digitized documents in the preparation of binding documents. An analysis enables quality decision-making in the domain of tax activities, and also decided upon in terms of objection to the issued tax solutions.

IV. MODEL OF TAX DATA WAREHOUSE

Before building data marts implemented the ETL process modeling. The process of extraction is carried out from relational databases, with the exception of tax data stored in various file formats generated from incompatible applications (xml, txt, xlsx). Validation is performed by checking the data structures and check the correctness of the data. After validation was performed import data into the warehouse. Data warehouse consists of three data marts, and data marts are mutually connected by using shared dimension tables. Figure 6 shows the organisational scheme of tax data warehouse, with fact tables and dimension tables visible.

A. Defining Dimension Tables

Dimension tables used in the tax data warehouse, visible in figures 2, 3, 4 and 6 are as follows:

- *Taxable period* - data on taxable periods established by legislation, taxable period is usually one calendar month, three calendar months or a calendar year, depending on the status of tax payer and in accordance with the decision of tax authorities.
- *Type of tax documents* – data on defined types of tax records and reports in line with regulations,
- *Tax partners* – data on tax partners and their tax statuses, these are usually customers, suppliers and tax authorities,
- *Business units* – data on business units that created or received a document,
- *Date* – data about dates,
- *Entry operator* –data on operators authorized to enter, update or view data,
- *Invoice items* - data about items on issued invoices with the following elements-data: products or services, unit price, unit price in currency, official currency code, approved discount (rebate) and tax rate,

By using more dimensions with fact tables, we can model subspecialist views and have the opportunity to analyse business processes in detail.

B. Fact Tables OUTGOING TAX DOCUMENTS

This is a fact table with data about invoices issued to customers for performed services or delivered goods, as well as data on tax liabilities, which contains the following attributes:

- *Invoice_number* – Number of issued invoice,
- *Taxable_period_code* – Code of taxable period,
- *Tax_document_type_code* – Code determining the type of tax records,
- *Operator_code* – Code of the operator generating the data,
- *Business_unit_code* – Business unit,
- *Business_partner_tax_number* – Code of business partner,
- *Date_code* – Date,
- *Customer_invoice_amount* – Amount of receivables from a customer,
- *VAT_amount_5%_rate* – Amount of tax liability at a 5% rate,
- *VAT_amount_13%_rate* – Amount of tax liability at a 13% rate,
- *VAT_amount_25%_rate* – Amount of tax liability at a 25% rate,
- *Local_reverse_charge* – Amount of local reverse charge,
- *Delivery_of_goods_in_other_EU_members* – Amount of delivery of goods in other EU member states,
- *Delivery_of_goods_within_EU*– Amount of delivery of goods within EU,
- *Delivery_of_services_within_EU*– Amount of delivery of services within EU,
- *Services_provided_to_persons_without_residence_in_RoC* – Amount of services provided to persons without residence in the Republic of Croatia,
- *Goods_installed_or_assembled_in_other_EU_members* – Amount of goods installed and assembled in another EU member state,
- *Not_subject_to_taxation_within_EU* - Amount that is not subject to taxation in EU,
- *Not_subject_to_domestic_taxation* - Amount that is not subject to taxation in the Republic of Croatia,
- *Export_deliveries* - Amount of deliveries outside EU,
- *Other_exemptions* – Amount of other exemptions,

C. Fact Tables INCOMING TAX DOCUMENTS

This is a fact table with data on invoices received from suppliers for received goods or services, as well as data on tax receivables, which contains the following attributes:

- *Invoice_number* – Internal number of received invoice,
- *Taxable_period_code* – Code of taxable period,

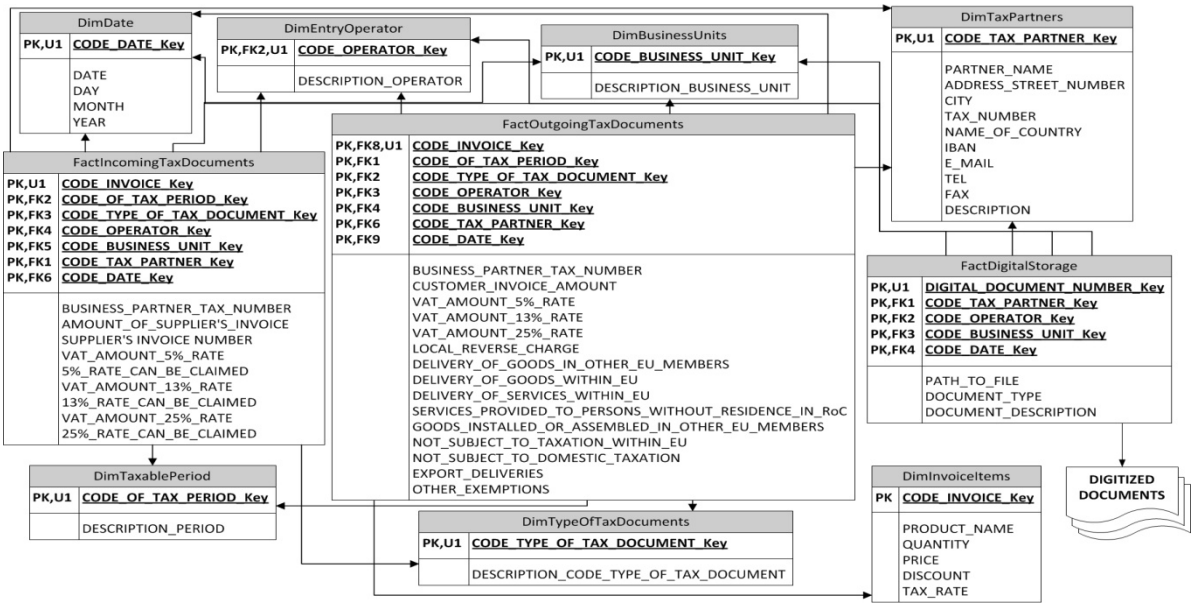


Figure 6. The organization of the warehouse tax data

- *Tax_document_type_code* – Code determining the type of tax records,
- *Operator_code* – code of the operator generating the data,
- *Business_unit_code* – Business unit,
- *Supplier's_invoice_number* – Invoice number of the supplier,
- *VAT_amount_5%_rate* – Amount of total tax shown on the invoice at a 5% rate,
- *5%_rate_can_be_claimed* - Amount of tax that is deductible and can be claimed at a 5% rate,
- *VAT_amount_13%_rate* – Amount of total tax shown on the invoice at a 13% rate,
- *13%_rate_can_be_claimed* - Amount of tax that is deductible and can be claimed at a 13% rate,
- *VAT_amount_25%_rate* – Amount of total tax shown on the invoice at a 25% rate,
- *25%_rate_can_be_claimed* - Amount of tax that is deductible and can be claimed at a 25% rate,

D. Fact Tables DIGITAL STORAGE

This is a fact table with data about digital tax documents, containing the following attributes:

- *Digital_document_number* – Represents the internal number of digital document,
- *Business_partner_tax_number* – Code of business partner
- *Operator_code* – Code of the operator generating the data,
- *Business_unit_code* – Business unit,
- *Date_code* – Date,
- *Path_to_file* – Represents the path to the location in the warehouse where the file with corresponding name of digital document is located,
- *Document_type* – Type of digital document (for example, incoming invoices, outgoing invoices, decision, appeal, tax return),
- *Document_description* – Short description of digital document.
- *Business_partner_tax_number* – Code of business partner,
- *Date_code* – Date,
- *Amount_of_supplier's_invoice* – Amount due to the supplier,

V. IMPLEMENTATION

The application for legal entity tax data warehousing has been created in Embarcadero studio architect technology using programming language Delphi XE5, and Embarcadero InterBase XE5 database technology.

Figure 7 shows one of the fundamental bookkeeping reports, outgoing invoices register which records all the invoices issued to customers, from which the following can be read:

- date of bookkeeping change that has occurred,
- data on customers receiving goods or services,
- data on the amount of receivables from customers,
- data on tax liabilities analytically shown according to tax rates, including tax bases shown according to the analytics of tax rates,
- data on tax exemptions analytically shown according to the type of tax exemption,

The report is created within a taxable period and is a result of data search from the fact table OUTGOING TAX DOCUMENTS, as well as corresponding dimension tables, which is shown in figures 3 and 6.

The incoming invoices register in figure 8 is a result of data search within the taxable period of the fact table INCOMING TAX DOCUMENTS, as well as corresponding dimension tables, which is shown in figures 2 and 6. Input accounts records all invoices from suppliers for goods or services, consisting of:

- date of bookkeeping change that has occurred,
- data on suppliers of goods or services,
- data on invoice number of the supplier,
- data on the amount due to the supplier,
- data on total taxes analytically shown according to tax rates, including tax bases shown according to the analytics of tax rates,
- data on deductible amounts of tax, and on tax that is not deductible, shown analytically according to tax rates,



Porezni obveznik: PRIMJER d.o.o.
Adresa: mjesto primjer,ulica primjer
Šifra djelatnosti: D2042
OIB: 12345678901

KNJIGA IZDANIH (IZLAZNIH RAČUNA)
01.09.2015 30.09.2015

Stranica: 1
Obrazac I-RA
- iznos u lunama i lipama -

RAČUN			KUPAC (PRIMATELJ DOBARA ILI USLUGA)		NEPODLIJEŽE OPOREZIVANJU OSLOBOĐENO PDV-a										OPOREZIVO					
															5%		13%		25%	
Rb	Broj	Datum	Naziv	OIB	IZNOS (s PDV-om)	TUZEMNI (NALJEPNICE) POREZNE OBVEZE	ISPORUKA DOBARA U DRUGIM ČLANICAMA EU	ISPORUKA DOBARA UNUTAR EU	OBAVLJENE USLUGE UNUTAR EU	OBAVLJENE USL. OSOBAMA BEZ SJEDIŠTA U RH	SASTAVLJANJE DOBARA U DRUGOJ DRŽAVI ČLANICI	ISPORUKE NPSU EU	U TUZEMSTVU	IZVOZNE ISPORUKE	OSTALA OSLOBOĐENJA	Osnovica	Porez	Osnovica	Porez	Osnovica	Porez
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	315	18.09.2015	INA d.d.	[illegible]	1.650,00																1.320,00	330,00
2	315	20.09.2015	HAC d.o.o.	[illegible]	1.800,00												200,00	10,00	300,00	39,00	840,80	210,20
3	415	22.09.2015	APPLE inc.	[illegible]	750,00					750,00												
			UKUPNO		4.000,00	0,00	0,00	0,00	0,00	750,00	0,00	0,00	0,00	0,00	0,00	0,00	200,00	10,00	300,00	39,00	2.160,80	540,20

Figure 7. The book of outgoing invoices



Porezni obveznik: PRIMJER d.o.o.
Adresa: mjesto primjer,ulica primjer
Šifra djelatnosti: D2042
OIB: 12345678901

Knjiga PRIMLJENIH(ULAZNIH RAČUNA)
01.08.2015 31.08.2015

Stranica: 1
Obrazac U-RA
- iznosi u lunama i lipama -

RAČUN			DOBAVLJAČ (ISPORUČITELJ DOBARA I USLUGA)		POREZNA OSNOVICA			Ukupni iznos računa	Ukupno	PRETPOREZ					
										5%		13%		25%	
Rdbr.	Broj	Datum	Naziv dobavljača	OIB	5%	13%	25%			Može se odbiti	Ne može se odbiti	Može se odbiti	Ne može se odbiti	Može se odbiti	Ne može se odbiti
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	17-25-3	08.08.2015	HAC d.o.o.	57500462912	200,00	500,00	780,00	1.750,00	270,00	10,00		65,00		195,00	
2	15-11	12.08.2015	INA d.d.	27759560625			400,00	500,00	100,00						100,00
3	1668/15	16.08.2015	HP d.d.	87311810356			280,00	350,00	70,00					49,00	21,00
			UKUPNO		200,00	500,00	1.460,00	2.600,00	440,00	10,00	0,00	65,00	0,00	244,00	121,00

Figure 8. The book incoming invoice

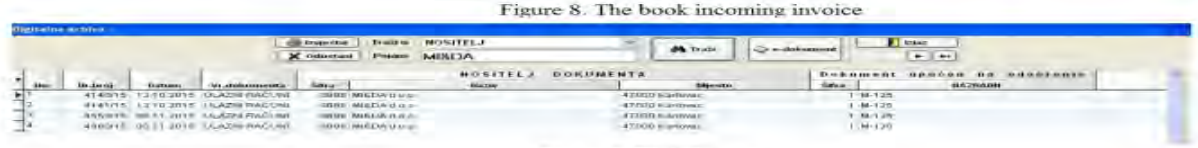


Figure 9. Digital storage

Digital storage in figure 9 represents the latest method of storing business documentation and is a result of fact table DIGITAL STORAGE. Its advantages are:

- fast access and search of digital incoming and outgoing invoices, as well as all tax documents delivered to or received from competent tax authorities,
- simultaneous access by multiple operators,
- complete protection from unauthorized access,
- reduction of costs with respect to storage of paper documents.

The application enables the overview of all prescribed reports, as well as of the reports providing a broader image of the tax payer regarding the flow of tax liabilities and receivables. These are not described in this paper due to space limitations.

VI. CONCLUSION

This paper describes the model for tax data warehousing implemented in practical business applications. With the implementation of tax support system, applying the ETL model to the model of tax data warehouse, prerequisites for achieving a quality control of digital tax storage have been met, as well as for managing and monitoring tax data in the system of profit-oriented economy.

The aim of this paper is to help the management of legal entities decide on the creation of tax data warehouse, and to propose guidelines and a model for building a warehouse to designers, based on own experience and applied model.

By forming a tax data warehouse and adopting a model that separates the transactional part of information business system from the tax data warehouse itself, we primarily increase security and enable transparency in the tax segment and in the achievement of tax secrecy instrument. We also reduce pressure on operational databases in the sense of operation, amount of data and transaction conditions.

In circumstances of compliance with legislation which provides for high fines in case of lack of or possession of inaccurate tax records or their possession on non-compliant forms (reports), and in circumstances in which the security of information systems requires increasing amounts of human resources, as well as financial resources, the need for creating tax data warehouses for legal entities has become imperative.

REFERENCES

[1] Javorović, B., Bilandžić, M.: Poslovne informacije i business intelligence, Golden marketing – Tehnička knjiga, Zagreb 2007.

[2] O'Brien J. A., Marakas G. M., Management Information Systems, McGraw-Hill, New York, 2006

[3] Data Warehouse Design: Modern Principles and Methodologies, Matteo Golfarelli and Stefano Rizzi, McGraw-Hill, 2009.

[4] Golfarelli M., Rizzi S., Saltarelli E.: WAND: A CASE Tool for Workload-Based Design of a Data Mart. SEBD 2002

[5] http://www.kimballgroup.com/html/about.html

[6] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Wiley Computer Publishing, 2002

[7] Huang X., Data Warehouse Architecture: Practices and Trends, IGI Global, 2010.

[8] R. Huges, Agile Data Warehousing for the Enterprise- A Guide for Solution Architects and Project Leaders, MK 2015.

[9] Mladen Varga, Baze podataka, Društvo za razvoj informacijske pismenosti (DRIP) 1994,

[10] D.Linstedt, Olschimke, Building a Scalable Data Warehouse with Data Vault 2.0, MK 2015

[11] Martin, J. Principles of Object Oriented Analysis and Design Prentice Hall, Englewood Cliffs, 1993

[12] Mile Pavlić, Oblikovanje baza podataka, Rijeka: Odjel za informatiku Sveučilišta u Rijeci 2011,

Predictive analytics in Big Data platforms – comparison and strategies

M. Zekić-Sušac*, A. Has**
University of Josip Juraj Strossmayer in Osijek, Faculty of Economics in Osijek, Osijek, Croatia
* marijana@efos.hr
** adela.has@efos.hr

Abstract – Predictive analytics include methods that are able to model data for prediction, classification, clustering, and association purposes. In the context of Big Data paradigm, the tools for performing those methods need certain adjustments to address the Big Data volume, velocity, variety, veracity, and value. The aim of this paper is first to compare the characteristics of several biggest tools for predictive analytics according to their flexibility and adjustability to Big Data platforms such as Hadoop and Spark. The second aim is to suggest strategies for efficient exploatation of business analytics in decision making within a business intelligent system. Both predictive analytics tools and business intelligence systems need adjustments to address new trends in data collection and analyses. The paper also discusses the necessary changes in business intelligent systems regarding their impact to management decision processes.

Keywords: Big Data, business analytics, business intelligence system, predictive analytics

I. INTRODUCTION

Although „Big Data“ term includes capturing, storing, processing, analysing and visualising huge quantities of information, Erl et al. [6] emphasize that Big Data does not only denotes data management problem. According to [6], more accurate explanation is that it is the process of solving business problems whose solutions are enabled by technology that can support the analysis of Big Data datasets. Big Data is not just a data warehouse in the cloud, because in addition to processing structured data, it can also deal with semistructured and unstructured data and documents [22]. Business analytics in the context of Big Data is emphasized as crucial for managing business performance today. The explosion of the quantity of data generated every day by companies, social networks, web sites, as well as diversity of data available in multimedia formats, documents, social media, etc. is enormous. The experts estimate that 35 zettabytes of data will be stored by 2020 [22]. The biggest challenge of today is how to exploit the enormous amounts of data and turn them into an efficient support for decision making. Traditional analytics packages such as statistical, mathematical tools, as well as tools that are included in companies' databases systems and in ERP systems cannot handle the quantity and variety of data within Big Data concept. Landrock et al. [9] emphasize that the era of making decisions based on simple reports on company's performance indexes (PKI's) is over, and that the new era of finding connections among PKI's and all available semi-structured and unstructured data on the web is beginning. The main purpose of Big Data analytics is well described by [22] in the sentence „When combined with advanced analytics, the right platform can turn mountains of data into intelligence that can be translated into action, turning our systems into intelligent processes“ [22].

However, inspite of emphasizing the importance of Big Data analytics, the IT experts still do not provide clear guideliness to business managers on how to select the appropriate tools and incorporate it into their information systems. This paper addresses that gap by comparing the characteristics of most popular Big Data platforms and evaluating their scope in addressing business problems. In order to provide guidelies for adopting predictive analytics in business, the paper furtherly suggests a strategy that could be used for an efficient development of Big Data predictive analytics in accordance to a company's goals.

II. PREVIOUS RESEARCH ON BIG DATA AND PREDICTIVE ANALYTICS

Due to known characteristics of Big Data which started with 3 V's (volume, variety, and velocity) and recently expanded with another 2 V's (verocity and value) [11], a careful choice of tools that can deal with such characteristics and incorporate them in decision making is needed. According to [6], data analytics includes the management of complete data life cycle, from collecting data, cleansing, organizing, storing, analyzing, to data governing. In the context of Big Data, data analytics is focused on managing large volumes of data using highly scalable distributed technologies. Data analytics can be divided into descriptive, diagnostic, predictive, and prescriptive analytics [6]. Predictive analytics aims to determine the outcome of an event that might occur in the future. It investigates strength and magnitude of the associations, trends, patterns and exceptions, in order to create models that could be used for future predictions. Finally, prescriptive analytics uses the results of descriptive and predictive analytics, suggests courses of action, and shows the likely outcomes of possible decisions. Descriptive statistics is the most common in analytical tools, such as Google Analytics, while the diagnostic, predictive, and prescriptive analytics require more advanced methods and models (such as machine learning, business rules, metaheuristics, etc.), and many companies still do not use it although it could bring more value to decision making.

How can decision makers select a predictive analytics tool that best suits their needs? The criteria of selecting business analytics tools are not well defined. The researchers emphasize that a company should consider nature of the data it works with, organizational budget, infrastructure (platform which enables data interconnectivity), the skillset of the company's staff, the level of coding expertize required, the visualization capabilities, flexibility, scalability, etc. [9] Loshin [10] suggests the following 5 criteria for comparing business analytics tools: (1) analyst expertize and skills, (2) analytical diversity, (3) scope of the data to be analyzed, (4) support for scalability and high performance, (5) collaboration, (6) vendor size and product integration, and (7) budget for licensing and maintenance. Among those that require lower analyst expertize he selects IBM SPSS Modeler, Oracle Advanced Analytics, and the Automated Analytics version of SAP Predictive Analytics. As more demanding for advanced skills he mentions IBM SPSS Statistics, KNIME Analytics Platform, Expert Analytics module of SAP Predictive Analytics, Microsoft Revolution Analytics, Teradata Aster Discovery Platform, and Oracle R Advanced Analytics. Regarding analytical diversity, Loshin [10] names Oracle Data Miner, IBM SPSS, SAS Enterprise Miner, Alteryx, and KNIME as tools that include the largest choice of advanced methods and models, such as neural networks, clustering, association analysis, link analysis, and visualization methods. Also, an important diversity feature of a system should be its integrativity with programming languages such as R. Regarding the scope of the data, analytics tools differ in a number of different data sources they support: conventional databases, cloud sources, Big Data platforms, and unstructured data. The criteria of scalability is important due to different sizes of companies and almost all the available tools satisfy this feature (from possible run on desktop computers to extension into Hadoop paralellism).

Regarding the budget there are free or low cost analytical tools such as Rapid Miner or R, but most of the tools have different editions that are suited to fulfill user budget and needs.

A systematic comparison of Big Data platforms on four main levels: strategy, processes, software, and hardware, is given by [9]. In this paper, we take a different approach by comparing the tools from a problem domain perspective, and methodological variety.

III. METHODOLOGY

In this section the methodology and technology aspects of using predictive analytics within Big Data will be described. Predictive analytics within Big Data has some specific charateristics comparing to predictive analytics which uses traditional database. The main differences are the following: (a) all datasets instead of samples are used for modelling purposes, (b) structured as well as unstructured data (graphics, documents, etc.) are used, and (c) social media data is also used (likes, shares, ratings, comments, etc.). Figure 1 illustrates those different characteristics by emphasizing that predictive analytics, when performed on transactional databases, can be used to analyze past events on structured data, while in Big Data it can be used for real-time modeling with structured, semi-structured and unstructured data. Methodology for predictive analytics on transactional databases includes various advanced statistical and machine learning methods, such as regression analysis, neural networks, decision trees, clustering, association rules, genetic algorithms, etc., commonly known as data mining methods for discovering hidden patterns in data [7]. In Big Data platform, the same methods could be used for predictive analytics, with additional methods that could process unstructured data, such as speech recognition, text mining, semantics, etc.

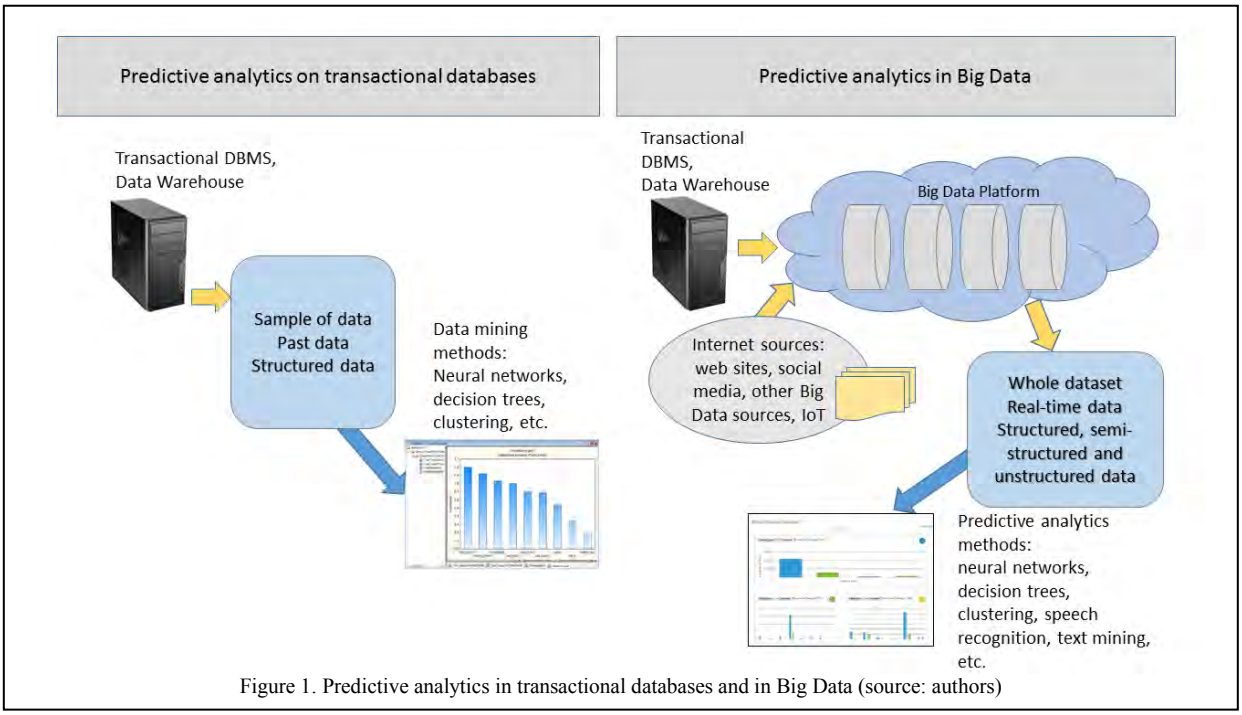


Figure 1. Predictive analytics in transactional databases and in Big Data (source: authors)

The sampling problem in predictive analytics on transactional databases is connected with the capacity and storage problems. Many companies erase their log data or other historical data because they have no capacity to store and process such large quantity of data. Big Data enables them to address those issues. Instead of taking a sample of data from a structured database, the storage advantages of Big Data platforms enable to estimate models on huge datasets. Another important benefit of such analysis is the possibility of using customers' oppinions from social media as input data almost in real time as the transactions are hapenning [22].

Some of the examples of possible efficient usage of predictive analytics in BigData are consumer analytics [5], fraud detection (in financial institutions, health care, insurance, etc.) [1], risk modelling and management, customer satisfaction and sentiment with product and service quality, customer loyalty prediction, and many more [10], [12], [19].

The technology that is able to support Big Data includes a new form of database platform (such as Hadoop) as a kernel system, and additional tools for managing analytics that operates in a cloud, therefore enabling companies scalability and on-demand usage. Such concept is also adjustable to small companies which do not have resources (financial, human experts) to conduct advanced analytics, but are usually high-growth companies that need analytics to be competitive at the market.

A. Big Data platforms Hadoop and Spark

According to [22], Hadoop is a computing environment (created by Apache Software Foundation) which uses a distributed clustered file system designed for very large-scale data operations. It means that it can quickly scan through large databases and produce the results within a scalable distributed batch processing system. It consists of three main parts: (1) Hadoop distributed file system (HDFS), (2) a programming paradigm MapReduce, and (3) Hadoop Common. Data in the file system are divided into smaller blocks, and distributed through clusters, therefore enabling scalability of the search. Clusters consist of a number of servers, where each server consists of internal disk drives, and only the disks where the data are actually stored can be searched with MapReduce using data locality concept, therefore enabling the speed of results. Also, data are stored in a redundant manner in order to prevent data loss in case of a failure in any hardware component, while a developer or a user does not need to take care of data locality – the Hadoop creates and manages the NameNode server which stores metadata on individual blocks. The third component, called Hadoop Common, is a set of libraries aimed to support Hadoop subprojects, such as Apache Avro for data serialization, databases Cassandra and Hbase, Hive (ad-hoc queries), Pig (Hadoop programming language), ZooKeeper for synchronization accross a cluster, etc. Big Data analytics providers, such as IBM, Oracle, SAS, SAP and others, embrace the capabilities of Hadoop and provide commercial applications built on Hadoop platform in order to make Big Data users an easier implementation of Big Data analytics.

The main advantage of Hadoop is in its ability to quickly process very large amounts of strucured and unstructured data. Disadvantages of Hadoop include its centralized metadata in the NameNode server, therefore a certain vulnerability. General limitation of all Big Data platforms is in the fact that there is a limited number of IT experts that are educated to install and manage the platform.

Spark is a Big Data open-source platform built originaly in the AMPLab at University of California Berkeley, later incorporated under Apache. It is not entirely independent from Hadoop because it is built upon Hadoop's distributed file system, but is an alternative to the Hadoop MapReduce component, and has been developed as a platform for Big Data analytics. Spark enables writing applications in Java, Scala or Python, supports SQL queries, machine learning, graph data processing, and streaming. It addresses some limitations of MapReduce component of Hadoop by introducing a concept of resilient distributed datasets (RDDS). In that way it supports a much wider class of applications than MapReduce.

Hardware configuration needed for Hadoop MapReduce or Spark platform is highly scalable and can be divided into a master (NameNode) server and worker nodes. According to [18] a small claster of servers with fewer than 20 worker nodes can consist of a dual quad-core 2.6 Ghz CPU, 24 GB of DDR3 RAM, dual 1 Gb Ethernet NICs, a SAS drive controller, and at least two SATA II drives. Larger clusters usually benefit from an additional 24 GB of RAM. If we want to compare the speed of Hadoop MapReduce and Spark, we can refer to several tests. One of them is the Databricks test [20] of sorting 100 TB of data (Daytona Greysort test) performed on 206 machines from the Amazon's EC2 cloud (the i2.8xlarge instances), with 6,592 cores, 49 TB of memory, and 1.2 PB of solid-state drive storage. The Spark running on Hadoop sorted that database in 23 minutes, which was three times faster than the Yahoo's test of MapReduce done previously (in 72 minutes) with 10 times more computing power. The extensive comparative tests of Shi et al. [17] show that Spark is faster than MapReduce for the operations of Word Count, k-means, and PageRank. When comparing their abilities for analytics, Spark is in advantage over MapReduce since it includes its own machine learning libraries. However, if a user only needs a platform to process large amount of structured data, without advanced modelling, the MapReduce is a better option.

It is also interesting to test the performance of databases that run of Hadoop platform. An extensive test conducted by End Point [4] compared the speed and throughput of Apache Cassandra, Couchbase, HBase, and MongoDB databases. The tests ran on Amazon Web Services EC2 instances, using the i2.xlarge class of instances (30.5 GB RAM, 4 CPU cores, and a single

volume of 800 GB of SSD local storage for the database nodes. Worker nodes consisted of c3.xlarge class of instances with 7.5 GB RAM, 4 CPU cores. The tests were perfomed on 1,2,4,8,16, and 32 database nodes by observing the operations of load, read, write, mixed read/write, read/modify/write, and some others. In overall, Cassandra showed the lowest and most consistent latency for each test. For instance, in read/modify/write workload, the latency of Cassandra with 32 database nodes was 40,357.81 microseconds, while the latency of MongoDb was 2,606,085.12 microseconds.

Both Hadoop MapReduce and Spark are free platforms, but require programming skills to manage data. By adding commercial analytics tools around Hadoop and Spark, managers are able to store, access, and analyze data by themselves. Davenport and Dyché [3] emphasize that a key skill needed to implement Big Data into companies is explaining big data outcomes to executives in visual and narrative form.

B. Big Data analytics tools

In comparing most known predictive analytics tools that run within Big Data platforms, we use the criteria of business problem domain and variety of methods included. Due to the lack of space, four major providers of Big Data analytics are compared: IBM Watson Analytics, SAS Enterprise Miner, SAP Predictive Analytics, and Oracle Advanced Analytics. The comparison is based on the information available at the web pages of those providers [2], [8], [15], [16]. Other, also popular providers of Big Data predictive analytics are Amazon (Amazon Machine Learning), Dell Big Data Analytics (Statistica), KNIME Analytics Platform, Microsoft Power BI, RedHat Access Insights (RHAI), Software AG (APAMA Streaming Analytics), Teradata (Aster Analytics) and others. By using the given comparison, we provide a potential user a possibility to estimate appropriateness of each tool to their business needs, and give an insight into its scope and methodological sofistication. Also, to provide even more clear guidance to managers who still lack their expertize in selecting and using Big Data predictive analytics, we suggest a strategy consisted of several important steps and decisions that are to be made in order to gain the benefits of Big Data predictive analytics in their businesses. In the suggested strategy, we focus on several predictive methods that have shown their success over standard statistical analysis, such as neural networks, decision trees, clustering, and association rules [13],[21].

IV. RESULTS – COMPARISON AND STRATEGIES

A. Comparison of predictive analytics tools in Big Data

According to previously described criteria, Table 1 shows a comparison of selected available predictive analytics tools that enable data mining and modeling.

TABLE 1. COMPARISON OF PREDICTIVE ANALYTICS TOOLS IN BIG DATA

Predictive analytics tools	Business problem domain	Advanced predictive analytics methods included
IBM Watson Analytics	customer buying and behavior trends, marketing campaign opportunities, discovering hidden insights in social media data, sales and revenue analysis, resource maintenance, detecting product issues, human resource performance	decision trees, linear and logistic regression, statistical comparison tests: model comparison test, influence test, analysis of variance (ANOVA), distribution test, Fisher r-to-t test, unusual high-low analysis, etc.
SAS Enterprise Miner	detecting and reduce fraud, measure credit risk, maximize cross-selling opportunities, retain customer, optimize marketing campaigns, response rates for marketing campaigns and curb customer, attrition reduce production failures, reduce asset downtime, optimize distribution, anticipate resource demands, consumer behavior analysis	clustering and self-organizing maps, market basket analysis, sequence and web path analysis, link analysis,dimension reduction techniques, linear and logistic regression, decision trees, gradient boosting, neural networks, partial least squares regression, two-stage modeling, memory-based reasoning, model ensembles, time series data mining, survival analysis, ratemaking for insurance, etc.
SAP (SAP Predictive Analytics	discovering market trends and customer needs, customer segmentation, revenue analysis	clasification/ regression model (decision trees), clustering, time series analysis, association rules, social analysis, frequent path analysis, text analysis, sequence analysis, etc.
Oracle Advanced Analytics	prediction of sales opportunities, discovering hidden customer segments, customer segmentation, finding most profitable selling opportunities, churn prediction, customer behavior analysis, market basket analysis, detection of social networks influencers, security and fraud detection, predictive workforce	logistic regression, decision trees, naive Bayes, support vector machine, multiple regression (GLM) support vector machine, anomaly detection, association rules, clustering, feature extraction (nonnegative matrix factorization), etc.

B. Strategy for using Big Data predictive analytics

The technology development enabled a shift of business analytics from local reporting to business-led self-service [14] in Big Data analytics. However, the managers are still introducing themselves with possibilities of business analytics in Big Data platform which is overwhelmed with powerful tools, yet has no clear guidance on which tools are more appropriate to their needs. Here we suggest a strategy which describes how the business goals will be achieved by the available

resources of business analytics tools. The strategy is built upon several important guidelines for an efficient usage of predictive analytics in business. We suggest the following strategic guidelines called OSTIM (by first letters of keywords describing each of the guideline):

(1) **O**pportunities - Addressing opportunities for improving business and business goals that could be realized by having knowledge on customer needs. This strategic guideline could be also formulated as answering the question: "What new knowledge do I need about my customers/products/services/suppliers in order to be more successful?"
(2) **S**ources - Identifying sources of structured and unstructured data (documents) that will be used for business analytics - .i.e. "Where could information needed for addressing goals or opportunities be found – databases, web content, social networks?"
This guideline addresses the issues of identifying all sources of information that could contribute, such as the selection of social networks where the customers could be targeted, social media where content could be reached, social data where information on customers' opinion could be collected, and sensor or other devices in IoT that will be used.
(3) **T**ools – Selecting and setting-up platform and tools that will be used for collecting, storing and analyzing data. The information from Table 1 could be useful in this step, extended with other popular tools for predictive analytics. The process of platform selection should include IT experts.
(4) **I**ntuitive usage of tools – After selecting and setting-up the platform and tools, managers will be able to perform near real-time predictive analytics - intuitively create models, get insight into patterns, make predictions, etc. The managers need no or little knowledge on the methodology behind. Methods such as neural networks, decision trees, association rules, genetic algorithms, clustering, will be performed according to their appropriateness to the data used for modeling. Details on modeling procedures in predictive analytics are given in [13]. The results can be displayed in numerical form, textual description and graphical visualization.
(5) **M**aking change in your business – after reading textual and graphical explanation of results, the managers use the knowledge to improve the business, make better sale, more satisfied customers and employees, innovations, and more impact to the society. In order to measure the change the decisions based on predictive analytics have made, the management needs to establish performance measures. For example, cross-selling index, conversion rate, sale increase, or any other measure is desirable.

C. Illustrative example – Predicting customer behavior in retail

In order to illustrate the suggested OSTIM strategy, an example is given of a Croatian retail company that needs to do customer profiling in order to increase their cross-selling index and the response rate of their marketing campaings. Five steps of the OSTIM strategy are performed. In step 1 (Opportunities), an opportunity to use predictive analytics was identified for the purpose of discovering the profile of a customer who is likely to make a purchase of an item. In step 2 (Sources) it was found that the necessary data will be collected from a transactional database that stores data on customers' purchases, and is loaded into a Big Data platform. In step 3, the Watson Analytics tool for predictive analytics built upon Hadoop Big Data platform is selected as an appropriate predictive analytics tool due to its easiness of usage (cognitive knowledge) and availability of data mining methods that could be used for the purposes identified in step 1. Such intuitive interface enables business users to create models almost without any expertize in data science and advanced data analysis. The fourth step is to run the predictive model in Watson Analytics by selecting data file that will be used for modelling.

The data file which captures purchases from a Croatian retail store within one month is used. The output of the model created for the purpose of identifying customer profiles consists of a binary variable (valued as 1 for the existence of purchase, and 0 for the absence of purchase of a specific item or a set of items in a store, while the customer demographics and previous purchase history was used as input variables, i.e. predictors. The available input variables included age, gender, education level, home ownership, home value, number of cars, number of kids, and other available descriptive variables, as well as some behavioral variables (such as number of previous purchases, time between purchases, etc.). The initial database consisted of 14012 transactions obtained by initial loading, although this file could be updated in near real-time and increase to very high-capacity.

The results of the Watson predictive analytics were obtained in several forms: (a) as a graphical insight to the predictive power of customer profile, (b) in textual narrative form, and (c) as numerical data. As an example, Figure 2 shows a graph which responds to the question: „What influences tuna purchase?". It was shown that the variables: *No_od_kids* (number of kids), and *Home Value* (home value) have the highest predictive power in describing which customer is going to purchase the tuna item. Such interpretation can be obtain for each item in the database. The characteristics of customers which show high predictive power in Figure 2 could be used to create a profile of customers who are most likely to buy tuna fish (or other item under observation). The model furtherly provides a deeper analysis of the relationships among customers' characteristics by producing a graphical decision tree of item purchase. A number of additional exploration techiques are enabled with very intuitive questions regarding the relationships among data presented to the user.

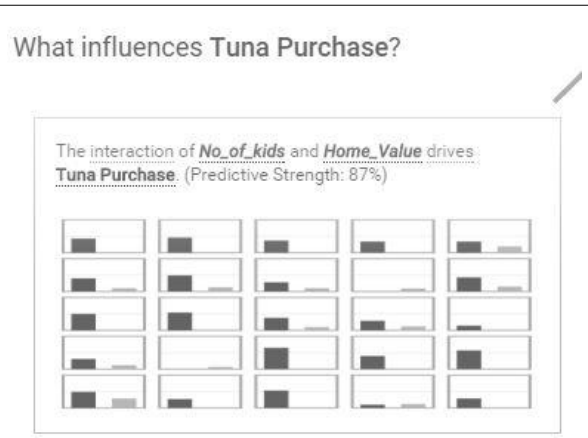


Figure 2. Results of an income prediction model obtained by machine learning methodology in Watson Analytics (source: authors)

Also, the analyses of social media in Watson Analytics brings the possibility to analyze multimedia content from social networks as well as data on customers ratings, evaluations, comments, etc.

V. DISCUSSION AND CONCLUSION

The purpose of this paper, besides explaining basic terms of predictive analytics and Big Data, was to compare the most popular tools and methods from a business user perspective and to suggest a strategy for an efficient usage of predictive analytics for adressing some new opportunities for improving business. The tools were compared in terms of problem domains and variety of methods included. In addition, a 5-step strategy OSTIM of using predictive analytics within a Big Data platform is suggested which could be used regardless the industry or problem specificity. The strategy has a purpose to encourage business managers to embrace predictive analytics and not be frighthened by advanced statistical and machine learning methodology behind. User-friendly intuitive tools with an initial help of a data science expert open a large spectar of possibilities for providing an efficient intelligent decision support system to managers. The strategy is illustrated on an example of the usage of one of the popular predictive analytics tools on simulated retail data on customer behavior, which provides intuitive interface for data load and result reporting and visualization. The paper is a preliminary research and planned to be extended by using retail data of higher volume, as well as unstructured data and social media. The results of such analysis, when performed on real huge datasets and in near real-time loading and processing could serve as an efficient intelligent support to decision makers, as well as to create an innovative environment within a company.

References:

[1] A. Bănărescu, Detecting and Preventing Fraud with Data Analytics, Procedia Economics and Finance, Vol. 32, 2015, pp. 1827- 1836.

[2] C. Berger, Big Data Analytics with Oracle Advanced Analytics in-Database Option, Oracle, 2012, http://www.oracle.com/technetwork/database/options/advanced-analytics/bigdataoracleadvanalytics11gr2-1930894.pdf, [accessed 18.02.2016]

[3] T.H. Davenport, J. Dyché, Big Data in Big Companies, International Institute for Analytics, T.H. Davenport and SAS Institute Inc., http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf, 2013 [accessed 22.02.2016]

[4] End Point, Benchmarking Top NoSQL Databases, http://www.datastax.com/wp-content/themes/datastax-2014-08/files/NoSQL_Benchmarks_EndPoint.pdf, [accessed 25.04.2016]

[5] S. Erevelles, N. Fukawa, L. Swayne, Big Data consumer analytics and the transformation of marketing, Journal of Business Research, Vol. 69, No. 2, 2016, pp. 897-904.

[6] T. Erl, W. Khattak, P. Buhler, Big Data Fundamentals: Concepts, Drivers & Techniques, Boston: Prentice Hall, 2016.

[7] S. Finlay, Predictive Analytics, Data Mining and Big Data, Hampshire: Palgrave McMillan, 2013.

[8] IBM Analytics, Predictive analytics, What is predictive analytics, http://www.ibm.com/analytics/us/en/technology/predictive-analytics/, [accessed 17.02.2016]

[9] H. Landrock, O. Schonschek, A. Gadatsch, Big Data Vendor Benchmark 2015, A Comparison of Big Data Solution Providers, Experton Group, Germany, https://www.t-systems.com/solutions/big-data-vendor-download/1298900_1/blobBinary/Big+Data+Vendor_Download-ps.pdf [accessed 04.02.2016]

[10] D. Loshin, Big Data Analytics, From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph, Amsterdam: Elsevier Inc., 2013.

[11] B. Marr, Big Data: Using Smart Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance, Chichester: John Wiley & Sons, 2015.

[12] M. Minelli, M. Chambers, A. Dhiraj, Big data, big analytics: emerging business intelligence and analytic trends for today's businesses. John Wiley & Sons, 2013.

[13] D. Nettleton, Commercial Data Mining, Processing, Analyzing and Modeling for Predictive Analytics Projects, Amsterdam: Elsevier, Morgan Kaufmann, 2014.

[14] J. Parenteau, R.L. Sallam, C. Howson, J. Tapadinhas, K. Schlegel, T.W. Oestreich, Magic Quadrant for Business Intelligence and Analytics Platforms, Gartner Inc., https://www.gartner.com/doc/reprints?id=1-2XYY9ZR&ct=160204&st=sb, 2016 [accessed 17.02.2016]

[15] SAP Community Network, SAP Predictive analytics, http://scn.sap.com/community/predictive-analytics, [accessed 17.02.2016]

[16] SAS, Advanced analytics, Where power meets purpose, http://www.sas.com/en_us/software/analytics.html, [accessed 17.02.2016]

[17] J. Shi, Y. Qiu, U.F. Minhas, L. Jiao, C. Wang, B. Reinwald, F. Ozcan, Clash of the titans: MapReduce vs. Spark for lLarge Scale Data Analytics, Proceedings of the VLDB Endowment, Vol. 8, No. 13, pp. 2110-2121.

[18] E. Summer, Hadoop Operations, O-Reilly, https://www.safaribooksonline.com/library/view/hadoop-operations/9781449327279/ch04.html, [accessed 20.04.2015.]

[19] R.S. Xin, J. Rosen, M. Zaharia, M.J. Franklin, S. Shenker, I. Stoica, Shark: SQL and Rich Analytics at Scale, SIGMOD 2013. June 2013.

[20] A. Woodie, Spark Smashes MapReduce in BigData Benchmark, Datanami, http://www.datanami.com/2014/10/10/spark-smashes-mapreduce-big-data-benchmark/, October 10, 2014 [accessed 20.04.2016]

[21] M. Zekić-Sušac, A. Has, Data mining as support to knowledge management in marketing, Business Systems Research, Vol. 6, No. 2, 2015, pp. 18-30.

[22] P.C. Zikopoulus, C. Eaton, T. Deutsch, D. deRoos, G. Lapis, Understanding Big Data, Analytics for Enterprise Class Hadoop and Streaming Data, New York: McGraw Hill, 2012.

The Analysis of CSFs in stages of ERP Implementation – Case Study in small and medium – sized (SME) companies in Croatia

Milorad Nikitović*, Vjeran Strahonja **
* College for Information Technologies, VSITE, Zagreb, Croatia
** University of Zagreb, Faculty of Organization and Informatics, FOI, Varaždin, Croatia
milorad.nikitovic@vsite.hr; vjeran.strahonja@foi.hr

Abstract - This paper elaborates the stages of ERP implementation in terms of critical success factors in small and medium-sized enterprises in Croatia. These stages and their associated success factors were evaluated by company executives, implementation managers and advanced users. The implementation is an extremely critical process especially in small companies which have limited resources, particularly human resources, where some people often have to perform more than one business function. The paper shows that the number of critical success factors of implementation significantly differs in six examined stages and the quantity of critical factors in the stage with the highest number of them is 100% higher than in the stage with the lowest number of them. If we sum up the factors obtained through the stages and compare the result to the global list of critical success factors of implementation – the results differ to a considerable extent.

I. INTRODUCTION

After years of procurement, implementation and application of solutions based on ERP philosophy, these became prevalent solutions for monitoring the business functions of companies. Given the problems which accompany the totality of the implementation process of such solutions, the authors divide the totality of the implementation process differently. Thus, [1] some divide them into five stages (Define, Design, Build, Transition and Go Live & Support), this paper [2] suggests seven steps (Strategic Plan, Readiness Assessment, Prepare for Vendor Selection, Vendor Selection, Post Implementation Assessment, Implement the ERP Solution and the Plan Implementation), Adam et al. [3] only the solution selection process is divided into 4 stages (Intelligence, Design, Choice and Review), Bajwa et al. [4] divide this process into 5 stages (Awareness, Selection, Preparation, Implementation and Operation). In his dissertation, Otieno cites the importance of each stage of implementation, and divides them into 5 stages (Project Preparation, Business Blueprint, Realisation, Final preparation and Go Live & Support) [5]. On the other hand, Yingjie provides a somewhat different approach [6], which basically divides the implementation process into four basic stages, the Chartering phase (Preparation, Analysis, Design), the Project phase and the Shakedown phase (which basically constitute Implementation) and Phase Onward & Upward (which is essentially Maintenance) [6]. Peslak and others [7] have presented the implementation of SAP in two divisions of a manufacturing company. In their study the implementation was divided into four phases Preparation & Training, Transition, Performance & Usefulness and Maintenance. They came to the conclusion that there are two key phases which most directly affect the success of the implementation - the first and the third phase (Preparation & Training and Performance & Usefulness). Accordingly, a strong focus on these phases is recommended. Kronbichler et al. [8] have divided the totality of the implementation process into three phases (Planinng-Preimplementation, Implementation and Stabilisation / Improvement and Post-Implementation). In the aforementioned paper there is a review of literature in which the critical success factors of implementation are linked to each of these phases. Hustad and Bechina [9] have made a comparison of the critical success factors of implementation in 4 companies, based on the previous research of the same methodology through the three phases of implementation (Pre-implementation, Implementation and Post-implementation). Bhatti divided the research on the critical success factors of implementation in 53 Australian companies into 4 stages of implementation (Implementation Planning, Installation, Final Preparation and Go Live) [10].

There are many articles describing the particular stages of implementation in terms of importance, i.e. which factors within them are especially critical. In one of them Hustad and Olsen [11] examine the ERP Pre-Implementation Process in a Small and Medium-sized Enterprise (SME), and identify critical issues in this process. The ERP Pre-implementation phase consists of selecting ERP software, vendors and consultants. Ara and Al-Mudimigh [12] present a review of the impact of project management in ERP project life cycle by studying various project management methodologies. Nicolau and Bhattacharya [13] empirically examine the extent to which discrete changes to ERP systems over a post-implementation time-frame affect the company's ability to deliver long-term organizational performance. Chen and Tsai [14] present a very interesting approach; they investigated organizational readiness through the three variables of

implementation (ERP adoption, process adoption and organizational resistance). In his research, Hedman [15] attempts to reduce the critical factors of success to organizational, project and technical factors, while at the same time, he tries to classify them according to six stages of the implementation (Initiation, Adoption, Adaptation, Acceptance, Routinization and Infusion). A different approach has been taken in the work of Ahmad and others [16]. They tried to correlate the 4 stages of implementation (Project Initiation, Project Preparation, Realization, Operation and Maintenance) and their associated activities with those responsible for their realization. Olson and Zhao [17] investigated the positions of those in companies responsible for IT through 4 phases of implementation (Assessment phase, Planning phase, Implementation phase and Renewal phase). Tsai and others [18] examined the relationship between the critical success factors of implementation and the success of the ERP system of implementation through 4 phases of implementation (charter phase, project phase, shakedown phase and upward phase). Hasibuan and Dantes [19] have established priority CSFs in the ERP implementation phase by using weighting factors. The success of implementation is measured by five indicators: system quality, information quality, service quality, tactical impact and strategic impact.

From the abovementioned papers, as well as many others, it is evident that the implementation process generates certain critical factors of success. It is especially important to identify which CSFs are crucial in particular phases of the implementation in terms of targeted management of implementation, i.e. paying particular attention to the key factors of success. Some studies suggest that in 2015 the value of the Croatian ERP solutions market should reach $ 59,000,000 [20], wherein the significance of the implementation process is undisputed.

The paper is divided into four chapters. In the introduction there is a partial literary overview of CFSs according to various phases of implementation, and a glance at some of them, in relation to their corresponding phases. The second chapter presents a methodology of implementation with an emphasis on the key research question. The third chapter presents the results related to the "ranking" of CSFs related to the 6 stages of implementation and the comparison of sum values of CSFs with research in which the same CSFs were treated in total at the level of implementation as a whole [21]. The fourth chapter is the conclusion.

II. RESEARCH METHODOLOGY

A questionnaire with 32 proposed critical success factors of ERP implementation as a whole, in alphabetical order was sent to 120 companies. These factors had to be ranked according to degree of importance ranging from critical (1) very high (2), high (3), weak (4) and low (5). The questionnaire was addressed to three parties within each of the companies: the company executives, project managers and advanced users. Replies were received from 60 company executives, 80 project managers and 72 advanced users. Therefore, we can claim that 212 individuals [21] have responded. Of these people, 35 of them were selected and were sent the same questionnaire but related to the next phase of ERP implementation:

- Selection of ERP solution (Phase 1),
- Project preparation (Phase 2),
- Analysis of the compliance of the organization's business functions with the options of ERP solutions (Phase 3),
- Physical realization of the compliance analysis of business functions in the organization and ERP solution options (Phase 4),
- Transition to the new solution (Phase 5) and
- Everyday operations running on ERP solutions with the implementers' support (Phase 6).

Distribution these 32 critical success factors of implementation had to be ranked according to above mentioned phases of implementation. Critical success factors of ERP implementation were listed in alphabetical order:

1. Analysis and motivation regarding the necessity for ERP,
2. BPR & minimum customization,
3. Business culture,
4. Change management,
5. Clear goals and objectives,
6. Complex architecture and high number of implementation modules,
7. Data conversion,
8. Data management,
9. ERP treated as a program not a project,
10. ERP software package selection,
11. ERP system quality,
12. Implementation approach methodology,
13. Interdepartmental cooperation,
14. Management expectations,
15. Monitoring and evaluation of performance,
16. Organizational fit,
17. Organizational communication,
18. Partnership with vendor,
19. Performance evaluation and management,
20. Project management,
21. Relationship of business and IT strategy,
22. Software development, testing and troubleshooting,
23. Steering committee of project ERP implementation,
24. Team competence composition,
25. Top management support,
26. Understanding key problems of ERP implementation,
27. Use of external consultant,
28. Use of vendor's tools,
29. User acceptance,
30. User involvement & participation & competence,
31. User training and education (timely defined) and
32. Vendor support.

The questionnaires were sent by e-mail and post to 35 companies which earlier provided replies in this research. This time separate responses from company executives, project managers and more advanced users were not required, and the emphasis was on those that were considered "the best", which implies a high level of awareness of the totality of the implementation of ERP solutions. These questionnaires were sent out in mid-April 2012. A time frame of one month was set within which the responses were expected. The basic intention of the "in-depth analysis" was to answer the following research question:

- To what extent do the critical success factors of ERP implementation deviate when they are summed up through phases of ERP implementation compared to the same factors when responding to the global questionnaire related to the totality of ERP implementation, with the same suggested CSFs?

III. RESULTS

Replies were received from 20 respondents, a little over 57%. These responses were collected within more than two months and more than 80% responded after several requests, or after paying an actual visit to the persons in question. After processing received responses, the reviews of the implementation phases are as follows. Factors which received more than 12 votes have been particularly stressed.

A. Selection of ERP solution

At this stage the basic framework should be set which will allow the selection of the appropriate ERP solution. This phase presumes a consensus on the need for a new solution, if there was a previous one, or the awareness that without ERP solutions market competition is not possible. One of the strongest insights of the authors, based on a large number of completed implementations, is that at some point of discontent the company demonstratively decides on the transition to a new solution. Therefore, the decision cannot be a momentary whim of anger or ignorance. This phase consists of the first 7 critical factors (Table 1):

- Analysis and motivation regarding the necessity of ERP,
- Clear vision and business objectives,
- ERP software package selection,
- Management expectations,
- Organizational fit and
- Relationship of business and IT strategies.

These first seven critical success factors of ERP implementation are undoubtedly logical choices, considering the suggested CSFs.

B. Project preparation

Project preparation should focus on all activities related to the adequate assembly of a project team as well as the project preparation phases with related activities and with an emphasis on resources. Table 1 shows the first 7 critical factors:

- Clear goals and objectives,
- Team competence composition,
- Top management support,
- Thorough understanding of the key problems of ERP implementation,
- BPR & minimum customization,
- ERP system quality and
- Analysis and motivation regarding the necessity of ERP.

C. Analysis of the compliance of the organization's business functions with the options of ERP solutions

This is a very important phase of the whole of the implementation process. The specificity of this phase is reflected in the potential need for overhauling business processes or customizing solutions according to methods of executing processes within particular business functions. Table I states separately all critical success factors of implementation, related to this stage:

- Clear vision and business objectives,
- Top management support and
- Thorough understanding of the key issues related to ERP implementation.

D. Physical realization of the compliance analysis of business functions in the organizations and ERP solution options

This is "hard work" with users in the actual business environment or, as it is often referred to in practice, implementation "in the strict sense". After this phase, both external and internal implementation teams and employees should be able to know how the business processes will be running when the implementation is completed, i.e. in real life. Table 1shows the most critical success factors of ERP implementation which are related to the fourth stage:

- User Involvement and participation,
- Interdepartmental cooperation and
- Team competence composition.

E. Transition to the new solution

This period includes the first few weeks of the operation, during which the thinking process is still adapted to the previous solution (and the habits associated with it), in case of replacing one solution with the other. Depending on the amount of work that has been invested in the previous phases, in terms of training, at this stage there is a degree of disorientation, fear and rejection, which is only natural. Based on this phase the most important critical success factors are:

- Data conversion,
- User involvement and participation and
- User training and education (timely defined).

It should be noted that when it comes to the degree of criticality, the votes in the questionnaire were the least critical.

F. Everyday operations running on ERP solutions with the implementers' support

This phase implies that the users have generally accepted the new solution and have become almost addicted to it. What is particular about this phase is that the users, apart from focusing on their immediate area of operations, are beginning to realize the overall context of the solution, all the activities that precede their activities as well as those which depend on their own correct performance. Based on this phase - Everyday operations running on ERP solutions with the implementers' support - provided the least number of outstanding critical factors:

- User acceptance and
- User involvement and participation.

G. The comparison of critical success factors according to the phases of ERP implementation

Based on degree of importance (critical-1) of all six implementation phases their total value can be summed up.

TABLE I: SUMMARIZING SIX PHASES ERP IMPLEMENTATION (P1-PHASE 1, P2-PHASE 2, P3-PHASE3, P4-PHASE4, P5-PHASE5) FOR DEGREE OF IMPORTANCE CRITICAL-1

Critical Success factors of ERP Implementation	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Total**
Top management support	9	15	14	12	8	5	63
User involvement and participation	4	8	4	16	13	15	61
Clear goals and objectives	18	18	18	2		2	58
Thorough understanding of the key issues related to ERP implementation	10	13	14	9			46
Relationship of business and IT strategy	14	10	10	6		4	44
Analysis and motivation regarding the necessity of ERP	18	15	4	4	2	2	45
BPR & minimum customization	7	13	12	6		4	42
User acceptance	2		5	6	8	16	37
ERP system quality	14	13	5	4			36
Steering committee for project ERP implementation	3	12	13	4	4		36
Team competence composition	2	17		14	2		35
ERP treated as program rather than project		10	11	10			31
Interdepartmental cooperation		4		14	4	9	31
Management expectations	16	4	4	3	2		29
Organizational fit	16	4	5	2			27
Monitoring and evaluation of performance		2	6	6	6	5	25
Project management		4	5	12	4		25
Organizational communication	3	6	2	7	4	2	24
Change management		4	6	4	3	6	23
User training and education (timely defined)				12	9	2	23
Partnership with vendor	4		6	4	5	2	21
Data management			2	6	9	4	21
Implementation approach and methodology	8	2	6	4			20
Data conversion			2	4	13		19
ERP software package selection	18						18
Performance evaluation and management			2	5	8	2	17
Software development, testing and troubleshooting	1	2		3	8	2	16
Complex architecture and high number of implementation modules		2	2	2	4	3	13
Business culture		2	2		2	4	10
Vendor support			3	4			7
Use of vendor's tools				2			2
Use of external consultant							0

Based on the Table I it could be concluded:

- The tree most CSFs in terms of importance (critical-1) are: Top management support, User involvement and participation and Clear goals and participation,
- CSFs which are rated degree of importance in all phases of the life cycle ERP implementation are: Top management support, User involvement and participation, Analysis and motivation regarding the necessity of ERP and Organizational communication,
- CSF Clear goals and objectives is given 18 votes in first three phases of the life cycle ERP implementation,
- Most votes, from the point of importance (critical-1) are given: Clear goals and objectives, in first three phases and in phase 1 Analysis and motivation regarding the necessity of ERP and ERP software package selection and
- Interesting is the perception of respondents towards suppliers and consultants. Vendor support is given only 7 votes in 3th and ^{4th} phases of the life cycle and Use of vendor's tools only 2 votes in 4th phases of the life cycle ERP implementation. Use of external consultant did not get any vote.

Of all the implementation phases the 'most dangerous' for the respondents was the actual physical realization of the implementation, i.e. practical adaptation to business operations through ERP solution. The project preparation stage and the selection of the appropriate ERP solutions are almost as important to them. It should be noted that the last two phases are not so criticality.

In this way, defined critical success factors of ERP implementation are somewhat different (Table II) from the ranking of critical success factors of implementation in the paper of Nikitović and Strahonja [21]. In that paper the same questionnaire for three categories of respondents, company executives, project managers (or IT managers) and advanced users, was offered for the totality of the implementation process.

TABLE II: COMPARISON OF RANKING SUMMARY ACCORDING TO IMPLEMENTATION PHASES WITH GLOBAL RANKING

Critical success factors of implementation	**According to implementation phases (total)**	**According to global evaluation of the average grade**
Top management support	1	2
User involvement, participation and competence	2	5
Clear goals and business objectives	3	4
Understanding the key problems of ERP implementation	4	10
Raltionship of business and IT strategy	5	19
Analysis and motivation regarding the necessity of ERP	6	1
BPR & minimum customization	7	21
ERP system quality	8	11
Steering committee for project ERP implementation	9	28
User acceptance	10	7
Team competence composition	11	27
ERP treated as a program not a project	12	6
Management expectations	13	18
Interdepartmental cooperation	14	23
Organizational fit	15	12
Project management	16	9
Organizational communication	17	25
Change management	18	8
Monitoring and evaluation of performance	19	20
User training and education (timely organized)	20	15
Partnership with vendor	21	17
Implementation approach methodology	22	13
Data management	23	14
ERP software package selection	24	3
Software development, testing and troubleshooting	25	22
Data conversion	26	30
Performance evaluation and management	27	24
Complex architecture and high number of implementation modules	28	31
Business culture	29	29
Vendor support	30	16
Use of vendor's tools	31	26
Use of external consultant	32	32

IV. CONCLUSION

The implementation process of such an expensive and long-term solution must be carefully planned. In this paper the authors tried to compare the global approach to implementation with the implementation in phases, from the standpoint of critical success factors of implementation. Given that the study involved three groups of respondents, company executives, project managers and advanced users of the implemented solutions, the obtained responses gain significant importance. The research issue of this study, when it comes to responses, indicates that it is much better to conduct the implementation in ERP implementation phases, in terms of criticality, than when it is done at the level of global responses to the totality of the implementation. That way, each success factor associated with implementation can be examined more accurately. Undoubtedly, such a method of implementation is more expensive, but in the end can lead to the more successful completion. According to the paper [21], the attitudes of groups of participants in the implementation process are clearly different. Therefore, external participants in the implementation process (vendors, consultants, implementers) should receive special training for each of these groups, and within each phase of the implementation process, no matter how many there are.

The difference which can be seen in Table II, in terms of the importance of critical factors for the totality of the implementation process in relation to its phases, has two potential causes. The first was the decision to reduce the number of respondents from 212 which completed the global questionnaire to 35 carefully selected respondents (with more experience and knowledge). The second is related to the possibility to define more accurately the importance of each phase in relation to the totality of the implementation process.

Table I in our opinion deserves special attention. It shows the totality of the critical success factors of implementing ERP. According to this table, users found that the physical realization of the analysis of the compliance of the organization's business functions with the options of ERP solutions was a 100% more "dangerous" phase (188 critical success factors) than the one with the least of

them – the phase Everyday operations running on ERP solutions with the implementers' support (90 critical success factors). It should be noted that just the aforementioned phase "everyday operations" is a challenge for many researchers.

REFERENCES

[1] Nazir, M. M. ERP Implementation in Oil Refineries. In *Daily Business Recorder*, pages 103-114, Karachi, 2005.

[2] Cornelius, E. T. Seven Steps in the ERP Process. Collegiate, 2007.

[3] Adam, F.; Sammon, D.; Carton, F. Project Management Issues in Implementing ERP. Business Information Systems, University College Cork, Cork, Ireland, 2008.

[4] Bajwa, D. S.; Garcia, J. E.; Mooney, T. An Integrative Framework for the Assimilation of Enterprise Resource Planning Systems: Phases, Antecedents and Outcomes, *Journal of Computer Information Systems.* 81-90, 2004.

[5] Otieno, J.O. *Enterprise Resource Planning System Implementation and Upgrade (A Kenyan Study).* PhD thesis, Middlesex University, London, 2010.

[6] Yingjie, J. *Critical Success Factors in ERP Implementation in Finland.* M.Sc., The Swedish School of Economics and Business Administration, Vassa, Finland, 2005.

[7] Peslak, A.R.; Subramanian, G. H.; Clayton, G. E. The Phases of ERP Software Implementation and Maintenance: a model for predicting preferred ERP use, *Journal of Computerr Informattion Systems.* 25-33, 2008.

[8] Kronbichler, S. A.; Ostermann, H.; Staudinger, R. A Review of Critical Success Factors for ERP-Projects, *The Open Information Systems Journal,* 14-24, 2009.

[9] Hustad, E.; Bechina, A. A study of the ERP Project Life Cycles in Small-and-Medium-Sized Enterprises: Critical Issues and Lessons Learned, *World Academy of Science, Engineering and Technology 60,* 110-116, 2011.

[10] Bhatti, T. Implementation Process of Enterprise Resource Planning (ERP): Empirical Validation, European and Mediterranean Conference on Informaton Systems (EMCIS) 2006., Costa Blanaca, Alicante, Spain, 2006.

[11] Hustad, E.; Olsen, D. H. Exploring the ERP Pre-Implementation Process in a small-andmedium-sized Enterprise: a case study of a Norwegian retail company, University of Agder, Kristiansand, Norway, 2012.

[12] Ara, A.; Al-Mudimigh, A. S. The Role and Impact of Project Management in ERP Project Implementation Life Cycle, *Global Journal of Computer Science & Technology,* Volume 11, Issue 5, 2011.

[13] Nicolaou, A. I.; Bhattacharya, S. Organizational performance effects of ERP systems usage: The impact of post-implementation changes, *International Journal of Accounting Information Systems,* Volume 7, 28-35, 2006.

[14] Chen, H.-H.; Tsai, A Study of Successful ERP – From the Organization Fit Perspective, Chungshan North Road, No. 40, Sec. 3, Taipei, Taiwan, 2003.

[15] Hedman, J. ERP Systems: Critical Factors in Theory and Practice, *CAICT COMMUNICATIONS 2010/2*, Copenhagen Business School, Frederiksberg, Danemark, 2010.

[16] Ahmad, R. M. T. R. L.; Othman, Z.; Mukhtar, M. Campus ERP implementation framework for private institution of higher learning environment in Malaysia, *WSEAS Transaction on Advances in Engineering Education,* Issue 1, Volume 8, January 2011.

[17] Olson, D. L.; Zhao, F. CIO's perspectives of critical factors in ERP upgrades projects, *Enterprise Information Systems,* Vol. 1, No. 1, February 2007., 129-138.

[18] Tsai, W.-H.; Chen, S.-p.; Chen, H.-C.; Hsu, J.-L.; Yang, C.-C. The Relationship between Critical Success Factors (CSFs) and Performace of ERP Systems Implementation – View of Implementation Phases, 2008.

[19] Hasibuan, Z. A.; Dantes, G. R. Priority of Key Success Factors (KSFS) on Enterprise Resource Planning (ERP) System Implementation Life Cycle, *Journal of Enterprise Resource Planning Studies,* Vol.2012., Article ID 122627, 15 pages, 2012.

[20] http://www.tportal.hr/scitech/tehno/138299/Domace-ERP-trziste-u-2010-palo-za-7-3-posto-html (downloaded: oktober, 12th 2012.

[21] Nikitović, M.; Strahonja, V. Critical Success Factors Aspects of the Enterprise Resource Planning Implementation, JOIS, Vol. 36 No. 2, (2012), Varaždin, Croatia, 2012.

Model optimizacije procesa s primjenom na punjenju bankomata

Ivan Osman, Krešimir Bokulić
Multicom D.o.o. , Zagreb, Hrvatska
ivan.osman@multicom.hr, kresimir.bokulic@multicom.hr

Sažetak - U pristupu smanjenju troškova u procesima kao što je sustav punjenja bankomata se dolazi do velikog broja mogućih problema. To su jednostavni problemi poput nošenja sa "expert inputom" tj. stručnom intervencijom pa sve do problema složenosti samoga procesa optimizacije u smislu broja varijabli i kompleksnosti algoritma.

Cilj našeg rada je stvaranje jednostavnog riješenja za optimizaciju procesa kojim bi se maksimalno snizila cijena održavanja kontinuirane usluge bankomata. U svrhu toga je razvijen algoritam za nalaženje optimalnih rješenja. Stvoren je i sustav za automatiziranu primjenu dobivenog algoritma u svrhu minimalizacije cijena punjenja bankomata kroz neko razdoblje. Efikasnost sustava optimizacije je naposljetku i testirana na bankomatima simulirane potrošnje novčanica.

I. UVOD

Kao ciljna varijabla optimizacije je odabrana ukupna cijena niza pružanja usluge punjenja bankomata. Sama pojedinačna cijena punjenja bankomata je podijeljena na 8 faktora (cijena novca, cijena punjenja, cijena pražnjenja, cijena praznog bankomata, itd.) te jednu vremensku komponentu. Vremenska komponenta je definirana kao iznos koliko će ukupno dana bankomat vršiti uslugu sa sumom novaca koja je u danom punjenju stavlja u bankomat. Tu informaciju te posljedično i informaciju kada će biti potrebno odraditi sljedeće punjenje se dobiva iz prediktivnog modela transakcija na bankomatima. Ono što preostaje je sistematizacija nalaženja mogućih kombinacija datuma na koje bi se moglo provesti punjenje bankomata.

Sami sustav se s obzirom na primjenu u punjenju bankomata može podijeliti na 3 glavne komponente –

A. Algoritam optimizacije punjenja bankomata
B. Prediktivni model dnevnih transakcija na bankomatima
C. Implementacija sustava optimizacije u SQL-u i PL/SQL-u

II. ALGORITAM OPTIMIZACIJE PUNJENJA BANKOMATA

Algoritam se može jednostavno podijeliti na tri najbitnije komponente:

A. Nalaženje mogućih kombinacija datuma na koje bi se moglo provesti punjenje bankomata

B. Izračun troškova svih kombinacija punjenja bankomata uzetih u razmatranje

C. Odabir optimalnog rješenja za punjenje bankomata

A Nalaženje kombinacija datuma na koje bi se moglo definirati punjenje bankomata

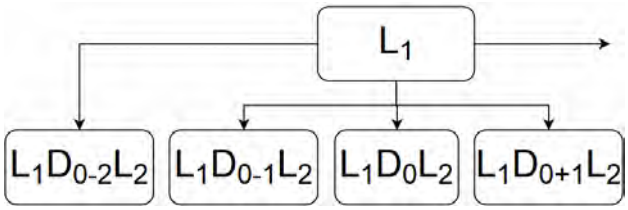
Za nalaženje datuma mogućih punjenja bankomata su potrebne informacije o trenutnom stanju količine prisutnog novca te određeni broj predikcija budućih stanja gotovine na bankomatu (dobiveno pomoću prediktivnog modela). Postavljena je pretpostavka da se informacija o trenutnom stanju gotovine na bankomatu dobiva barem jednom dnevno. Uz to je postavljena pretpostavka o dostupnosti predikcija budućih stanja od barem n+m dana od trenutka razmatranja; gdje je n broj dana do trenutka kada će bankomat ostati bez novaca a m označava koliko široki krug mogućih dana punjenja želimo razmatrati (optimalno između 2 i 6).

Računajući na dostupnost predikcija i određenosti broja n; algoritam traži dan kada bi bankomat trebao prijeći u negativni saldo - D_x. To je dan u čijoj neposrednoj okolini tražimo dane na koje je moguće puniti bankomat. D_x se određuje kumulativnom sumom predviđenog prometa. Traži se broj dana taka da od zadnjeg poznatog stanja na bankomatu suma predviđenih budućih transakcija bude jednaka ili veća od vrijednosti zadnjeg poznatog stanja. Ako p_i = promet na i-ti dan i SPP = suma zadnjeg poznatog stanja tražimo:

$$Min_i(\sum_{i=0}^{n} p_i) \geq SPP$$

S obzirom na D_x se u presjeku s kalendarom mogućih punjenja (dana kada je teoretski moguće izvršiti punjenje bankomata) uzima u razmatranje prvih n datuma u prošlost te prvih n datuma u budućnost(gdje je n postavljeni parametar od strane korisnika). Kao bitan za daljnju primjenu se nameće dan D_0 (najčešće najbolji odabir). D_0 je definiran kao prvi dan na koji je moguće punjenje bankomata koji je ranije ili jednak danu D_x. Do razlike između D_0 i D_x dolazi pošto ne mora nužno svaki dan u mjesecu biti mogući dan punjenja; nego se gleda kalendar mogućih dana punjenja. D_x može biti bilo koji dan u mjesecu ali određivani niz datuma koji uzimamo u obzir za moguće datume punjenja (kao i D_0) oko D_x mora biti iz kalendara mogućih punjenja. Za kandidate datume punjenja (n dana u okolici D_x) se izračunava dugotrajnost punjenja za svaki od strane korisnika postavljeni mogući iznos punjenja. Time se završava prva iteracija algoritma u određivanju datuma punjenja.

U slučaju odabira kombinacije tj. niza od više punjenja se za dani bankomat svako sljedeće potencijalno punjenje (iteracija) oslanja na prethodnu. Za specifični potencijalni datum punjenja unutar n dana se iz okolice datuma D_x prethodne iteracije (dana kada će po prethodnoj iteraciji bankomat ostati bez novaca) odabire niz mogućih datuma punjenja te s obzirom na niz potencijalnih iznosa punjenja ponovo izračunava trenutak kada će bankomat ostati bez novaca (novi D_x i D_0). Nakon toga se ako je i predodređeno od strane korisnika kreće u sljedeću iteraciju algoritma. Primjer danog principa rada algoritma se može vidjeti iz slike 1. Na slici oznaka L_1 predstavlja matricu L_1 stupnja algoritma (primjer jedne takve matrice je slika 2). Iz te matrice se za svaki unos stvara nova matrica (L_2) koja je pozicionirana oko datuma kada je predviđeno da će za asocirani datum (oznaka D) i sumu novca iz L_1 stupnja algoritma ponestati novca u bankomatu. Time datum i suma punjenja bankomata iz L_1 iteracije i njemu pripadna matrica s datumima i sumama punjenja iz L_2 iteracije algoritma predstavljaju niz od nxm mogućih kombinacija punjenja bankomata (gdje su n i m brojevi redaka i stupaca pripadne matrice u L_2).



Slika 1, Primjer grananja iteracija algoritma L_1-L_2

B. Izračun troškova svih kombinacija punjenja bankomat uzetih u razmatranje

Pri izračunu troškova za svako moguće punjenje potrebno je uzeti u obzir:

CN – cijena novca

CPU- cijena punjenja

CPP – cijena pražnjenja

COP – cijena ophodnje

CPB – cijena praznog aparata

SD – predviđena suma novca za dan za koji se razmatra moguće punjenje

SD – suma punjenja kojom bi se napunio bankomat

S – kumulativna suma svih stanja bankomata od trenutka punjenja do trenutka praznog uređaja s obzirom na prediktivni model budućih transakcija

N – broj dana od potencijalnog punjenja bankomata do trenutka kada po predikcijama s danim iznosom punjenja u uređaju ponestaje novaca

Time se može izračun cijene punjena bankomata definirati pomoću svih uključenih faktora kao:

$$C_{ij} = \frac{S*CN + SD*CPP + SP*CPU + COP + CPB}{N}$$

S time uzetim u obzir možemo krenuti u sistematizaciju izračuna spektra rješenja. U tu svrhu se stvaraju matrice cijena punjenja. U matrici su stupcima određeni datumi na koje je moguće napuniti bankomat dok su redci definirani sumama valuta punjivih u sami bankomat. Primjer tipične matrice punjenja se može vidjeti na slici 2 gdje je $t_i = D_0$ (prvi datum prije ili jednako D_x na koji je moguće punjenje). S obzirom na broj iteracija algoritma se broj ovakvih matrica povećava za vrijednosti po svakom slučaju dostupnom iz zadnje iteracije algoritma tj L-a (L_1 = prvi iteracija itd.). Same matrice cijena punjenja definiraju lokalne vrijednosti tj. cijene samo za zadani L_i stupanj strategije punjenja.

	t_{i-4}	t_{i-3}	t_{i-2}	t_{i-1}	t_i	t_{i+1}	t_{i+2}	t_{i+3}	t_{i+4}
600k									
300k					C_{ij}				
200k									

Slika 2, Primjer matrice lokalnih cijena punjenja za nedefinirani L_i stupanj optimizacije

C. Odabir optimalnog rješenja za punjenje bankomata

Za izračun završne cijene se gledaju kombinacije punjenja od L_1 do L_n gdje je n broj zadnje iteracije algoritma. Cijene komponenti (lokalne cijene unutar mogućih kombinacija punjenja) se zbrajaju te dijele s

brojem dana koliko će sama kombinacija punjenja trajati. Time se dobiva globalna cijena po dugotrajnosti same kombinacije punjenja. Ovim pristupom možemo raditi usporedbu svih kombinacija punjenja. Određivanjem minimuma cijena danih kombinacija punjenja dobivamo traženu strategiju punjenja.

III. PREDIKTIVNI MODEL DNEVNIH TRANSAKCIJA NA BANKOMATIMA

Preciznost modela predikcije budućih potrošnja bankomata je odlučujući faktor u smislenosti same optimizacije. Uz dovoljno neprecizni model predikcije se donošenje strategije niza budućih punjenja gubi u stalnoj potrebi za korekcijama. Sama vrsta modela za predikciju ne igra bitnu ulogu za optimizaciju ali s obzirom na generalni primjer punjenja bankomata se u prvi plan stavljaju mogući kandidati kao

A. Regresijska analiza vremenskih serija

B. Regresijska generalizirani linearni model

Regresijska analiza vremenskih serija se temelji na prethodnim tendencijama. Na primjer za predviđanje cijene dionice se odluka donosi na temelju prijašnjih kretanja cijene s obzirom na dan, mjesec ili neki drugi vremenski faktor. Točnije odluka se temelji na statističkoj obradi vremenske serije bazirano na vremenskoj domeni. Princip su osmislili George Box i Gwilym Jenkins.

GLM iliti generalizirani linearni model je klasa linearnih modela koja obuhvaća modele sa specijalnim strukturama grešaka, kategorijskim ili uređenim varijablama odaziva te multinomijalnim (zavisnim) varijablama. GLM je koliko robusna tako i široko iskoristiva klasa modela te je tako i vrlo pogodan u procesima optimizacije.

U svrhu testa modela optimizacije korištena su oba tipa regresijske analize. Korištene su i informacije o stvarnim transakcijama u svrhu otkrivanja nelogičnosti na samome principu rada algoritma. Generalno se za svaki bankomat stvara posebni model predikcije, baziran na faktorima tipa dan u mjesecu, dan u godini, dan u tjednu itd. Dobiveni niz predviđenih dnevnih transakcija pridodaje se modelu optimizacije za donošenje niza strategija punjenja te krajnje i završne odluke.

S obzirom na generalnu nesavršenost prediktivnog modela, sam sustav optimizacije mora uzimati u obzir i mogućnost povećanja troškova punjenja u svrhu smanjenja utjecaja prediktivnog modela na završni rezultat. Dodatni troškovi se mogu pojaviti u slučaju preniske predikcije što rezultira promjenom tipa punjenja bankomata (na skuplju verziju), hitne intervencije ili u najgorem slučaju nemogućnosti izvršavanja osnovnih usluga samoga bankomata (prazni bankomat). Radi smanjivanja mogućnosti stvaranja takvih troškova se definira niz preventivnih mjera.

Korisne intervencije za minimaliziranje utjecaja nesavršenosti prediktivnog modela uključuju:

A. *Preferiranje veće sume punjenja do predefiniranih n% ukupne cijene punjenja. Time se smanjuje broj punjenja uz neznatno povećanje cijene. Posljedično se smanjuje i broj situacija u kojima bankomat može ostati bez raspoložive svote ili bi morao žurno biti napunjen.*

B. *Preferiranje kasnijeg datuma punjenja u do n% ukupne cijene punjenja (rezultati su isti kao u br. 1).*

C. *Forsiranje regularne vrste punjenja ako „upper bound" predikcija potrošnje predviđa nedostatak novaca u neko dogledno vrijeme bez obzira na to što generalna predikcija predviđa dulji vremenski rok mogućnosti opsluživanja.*

D. *Korištenje granice do koje ako se bankomat isprazni dolazi do automatskog naručivanja punjenja (smanjuje mogućnost nemogućnosti pružanja usluge bankomata)*

E. *Kažnjavanje najave punjenja za ponedjeljak i/ili utorak (smanjuje se utjecaj multiplikacije greške prediktivnog modela tijekom vikenda, kada ne postoji mogućnost punjenja bankomata)*

F. *Ručni unos predikcija tj. takozvani „Expert input" (ostavlja se mogućnost intervencije na predikcije od strane stručnjaka kada postoji svijest o događajima i okolnostima podobnim za veću ili manju potrošnju u zadanom intervalu).*

IV. IMPLEMENTACIJA SUSTAVA OPTIMIZACIJE U SQL-U I PL/SQL-U

Objašnjenja bitnih termina:

D_x – dan prije nego bankomat ostane bez novaca

D_0 – dan prije ili jednak D_0 na koji je moguće obaviti punjenje bankomat

D_{02} - D_0 u sljedećoj iteraciji

L_1 – prvi stupanj algoritma tj. nalaženje prvog najboljeg dana punjenja

L_i – i-ti stupanj algoritma tj. nalaženje i budućih punjenja

U implementaciji algoritma za punjenje bankomata se koristi Oracle baza podataka te sql i pl/sql tehnologija. Implementacija samog L_n algoritma je pisana isključivo u ORACLE SQL tehnologiji te je rastavljena na 7 dijelova:

A. *Dobivanje prognoziranog stanja na dan: krećemo od zadnje dostupne informacije o stanju količine novčanicana bankomatu i od njega oduzimamo kumulativne sume prognoziranih dnevnih transakcija.*

B. *Dobivanje kalendara punjenja za vremensku okolinu na kojoj algoritam djeluje. Kalendar pruža informaciju o datumima na koja je moguće i koje vrste punjenja je moguće odraditi (izvanredno, regularno).*

C. *Nalaženje bitnih datuma: - D_0 , D_x , D_{02}.*

D. *Dobivanje svih kombinacija datumskih scenarija punjenja unutar zadane okoline datuma oko D_x vrijednost. Raspon okoline je definiran u posebnoj tablici OPT_PARAM.*

E. *Za sve kombinacije datumskih scenarija i scenarija punjenja izračunavaju se troškovi.*

F. *Agregiraju se sve moguće kombinacije punjenja te na nivou kombinacija izračunava cijena.*

G. *Zapis svih rezultata.*

V. ANALIZA RUBNIH SLUČAJEVA TE SPECIFIČNIH OSOBINA ALGORITMA

Za analizu primjenjivosti sustava optimizacije punjenja bankomata je potrebno uzeti u obzir spektar mogućih situacija koje bi potencijalno mogle stvarati troškove ili tjerati sustav optimizacije na donošenje štetnih odluka. U obzir su uzeti rubni slučajevi te mogući scenariji kao:

A. *bankomat se puni jako često (min)*

B. *bankomat se puni jako rijetko (max)*

C. *vrlo male sume punjenja*

D. *vrlo velike sume punjenja*

E. *promjene u sezonalnosti (ljeto, zima)*

F. *stagnacija stanja u bankomatu*

G. *greške u podacima potrebnim za modeliranje, validaciju ili optimizaciju samih punjenja*

H. *bankomat je prazan na nedefiniranu količinu vremena*

I. *nemogućnost definiranja D_x , D_0 ili D_{02}*

J. *skokovi u potrošnji koji su veći od teoretski moguće količine novca pohranjivog u bankomatu*

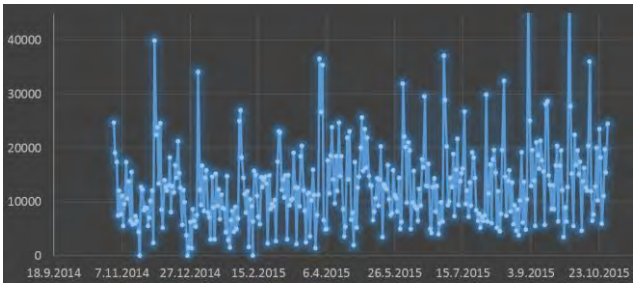
K. *dva ili više skoka u potrošnji na bankomatu koji su vremenski vrlo blizu jedan drugome*

L. *multiplikacije greške u predikciji s obzirom na neradne dane tj. nizove dana na koje je nemoguće puniti banakomat*

M. *Nemogućnost punjenja bankomata duži niz vremena prije većeg skoka potrošnje na bankomatu*

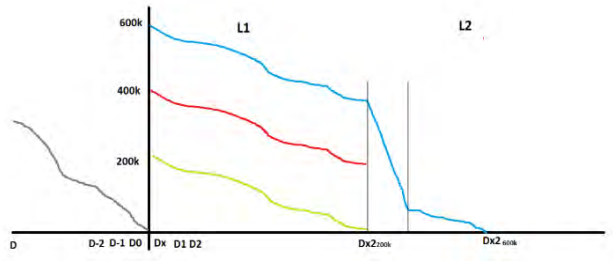
S obzirom na navedene slučajeve za analizu se nameću dani zaključci:

Nameće se primjena minimalno dvije iteracije algoritma (L_1+L_2) radi nužnosti optimizacije situacija kada se potrošnja na samom uređaju naglo poveća ili smanji. Primjeri toga su slučajevi nagle sezonalnosti ili slučajevi specifično velike potrošnje na određene dane u mjesecu. Za primjer možemo uzeti simulirani problem s velikim potrošnjama na određene dane u mjesecu – slika 3. Na prikazanom grafu Y os predstavlja visinu potrošnje a X os protok vremena (datume). Predstavljeni su simulirani podaci za od 1.11.2014 do 1.10.2015.



Slika 3, transakcije po danu na simuliranom bankomatu

U slučaju primjenjivanja samo prvog stupnja algoritma bi dolazilo do situacija kada je sustav suočen s odlukom o definiranju dana punjenja netom prije predviđeno velike potrošnje na bankomatu. Algoritam bi sugerirao punjenje s maksimalnom mogućom svotom novaca u pokušaju da zadovolji potrebe potrošnje za sljedećih n dana. U trenutku netom prije velike potrošnje izvršio bi još jedno punjenje maksimalne vrijednosti (inače ne bi mogao zadovoljiti predviđenu potrošnju na bankomatu). U slučaju korištenja L_1 + L_2 stupnja optimizacije bi algoritam uzimajući u obzir 2 buduća punjenja te cijenu novca i cijenu opsluživanja bankomata predložio minimalno veliku količinu novca za prvo punjenje. Tako ne ponestaje novca točno do predviđene veće potrošnje. Predviđenu veću potrošnju bi tada zadovoljilo u međuvremenu zakazano drugo punjenje. To drugo punjenje bi se odradilo na najbliži mogući datum točno prije povećanja potrošnje. S time je maksimalno minimizirana cijena ove kombinacije punjenja s obzirom na dano razdoblje. Primjer takvog problema je prikazano na slici 4. Pod pretpostavkom da je maksimalno moguće punjenje 400 000 Hrk se vidi da bi nakon dana D_{x2} bankomat ostao bez novca bez obzira kojom svotom bio punjen na dan D_0. Prema tome nameće se opcija punjenja s 200 000 Hrk na dan D_0 praćena sa punjenja od 400 000 Hrk na dan D_{x2}.



Slika 4, Primjer problema pri korištenju samo L_1 iteracije algoritma

VI. TESTIRANJE

U testu je korišteno 30 simuliranih bankomata različitih cijena opsluživanja, kalendara punjenja te količina i vrijednosti dnevnih transakcija. Transakcije pojedinačnih bankomata su podešene kako bi simulirali rad sustava bankomata. Nakon simulacije svih transakcija odsimuliran je i niz ručnih punjenja bankomata koji reprezentira standardni pristup korišten na europskom tržištu. Nasuprot tome je korišten sustav za optimizaciju punjenja bankomata u kombinaciji s GLM predikcijskim modelom. Korištene su verzije optimizacije samo s L_1, L_1+L_2 te L_1+L_2+L_3 stupnjem optimizacije. Korištenje L_4 ili većeg stupnja optimizacije se u početnim testiranjima pokazalo kao veliko povećanje u vremenu izračuna uz minimalnu ili nepostojeću korist u samim rezultatima. U testiranu su korišteni i različiti rasponi količine novaca koji se mogu pohraniti u punjenju bankomata.

Rezultati:

Raspon od između 100 000 Hrk do maksimalno 600 000 Hrk u bankomatu. L_1+L_2 stupanj optimizacije.

Simulacija ručnog punjenja:

Broj punjenja	Cijena punjenja	Cijena pražnjenja	Cijena novca	Cijena ophodnje	Ukupno
248	23.715,00 kn	3.701,00 kn	28.809,00 kn	50.432,00 kn	106.657,00 kn

Automatski optimizirana punjenja:

Broj punjenja	Cijena punjenja	Cijena pražnjenja	Cijena novca	Cijena ophodnje	Ukupno
148	18.524,00 kn	1.955,14 kn	41.239,00 kn	30.954,69 kn	92.672,83 kn

Raspon od između 100 000 Hrk do maksimalno 600 000 Hrk u bankomatu. L_1+L_2 stupanj optimizacije.

Simulacija ručnog punjenja:

Broj punjenja	Cijena punjenja	Cijena pražnjenja	Cijena novca	Cijena ophodnje	Ukupno
272	28.541,00 kn	5.869,52 kn	32.451,00 kn	51.967,00 kn	118.828,52 kn

Automatski optimizirana punjenja:

Broj punjenja	Cijena punjenja	Cijena pražnjenja	Cijena novca	Cijena ophodnje	Ukupno
183	22.325,00 kn	3.495,19 kn	45.122,00 kn	37.651,69 kn	108.593,88 kn

Uključujući ova dva testa je sveukupno testiranje pokazalo da optimizirani pristup punjenja bankomata u pravilu donosi snižavanje cijene opsluživanja samih bankomata za između 5% i 15%. Automatizirano i optimizirano punjenje bankomata isto tako nije dovelo do situacija kad bi bankomat ostao bez raspoloživog novca. U pravilu se pokazalo i da što veći raspon mogućih količina novaca punjivih u bankomat, to je prednost više na strani algoritamski optimiziranog pristupa.

VII. ZAKLJUČAK

Kroz testiranje L_1 te L_1+L_2 pristupa se sustav optimizacije pokazao puno djelotvorniji, brži i pokazao puno bolje rezultate od ručnog odabira vremena i sume novaca u punjenju bankomata. U slučaju korištenja samo L_1 stupnja optimizacije se dolazi do niza predloženih punjenja za period od mjesec dana u manje od 10 sekundi komputacije. Pri korištenju L_1+ L_2 stupnja optimizacije se dolazi do traženih rezultata za manje od 20 sekundi. Dodavanje daljnjih stupnjeva optimizacije se pokazalo nedovoljno korisno. L_1 +L_2 pristupom je potvrđeno omogućen izračun korisnih rješenja. Takvu razinu optimizacije bi pristupom ručnog izračuna u kompleksnim sustavima bilo teško te možda i nemoguće dobiti. Efikasnost sustava optimizacije je i demonstrirana na simuliranoj primjeni u bankarstvu pri optimizaciji punjenja bankomata. U toj primjeni sustav optimizacije postiže značajno bolje rezultate od ručnog odabira punjenja od strane stručnjaka. Algoritam za optimizaciju ostavlja i mogućnost daljnje primjene na procesima s većim brojem i vrstama faktora bitnih za izračun ciljne varijable. Sustav je generalno i lako primjenjiv uz SQL i PL/SQL tehnologiju. Sustav isto tako u radu ne zahtijeva dodatne informacije ili nadgledanje od strane stručnjaka za optimalnu iskoristivost.

LITERATURA

[1] Broda, P. and Levajković, T. and Kresoja, M. and Marčeta, M. and Mena, H. and Nikolić, M. and Stojančević, "Optimization of ATM filling-in with cash" European Study Group with Industry > ESGI 99 (Novi Sad, Serbia, Feb 3-7, 2014).

[2] prof.dr. I. F.A. Vis, prof.dr. E. Hagdorn van der Meijden, „Optimizing loggistics processes in cash supply chains", Vrije University doctorate disertation 2014.

[3] E. Greičius, S. Minkevičius, L. Sakalauskas "Maintenance of an ATM network: modeling of cash flows, analysis of cash demand and customer habits".neobjavljen

DE/GLGPS

International Conference on

DIGITAL ECONOMY AND
GOVERNMENT, LOCAL GOVERNMENT, PUBLIC SERVICES

Steering Committee

Chairs: Mladen Mauher, Polytechnic of Zagreb, Croatia
Edvard Tijan, University of Rijeka, Croatia

PAPERS

The modern approach to the analysis of logistics information systems

Ana Iskra [1], Edvard Tijan [2], Saša Aksentijević [3]
[1] E-mail: ana_iskra@yahoo.com
[2] University of Rijeka, Faculty of Maritime Studies
Studentska 2, 51000 Rijeka, Croatia
Tel: +385 51 33 84 11 Fax: +385 51 33 67 55 E-mail: etijan@pfri.hr
[3]Aksentijević Forensics and Consulting, Ltd.
Gornji Sroki 125a, Viškovo, Croatia
Tel: +385 51 65 17 00 Fax: +385 51 65 17 81 E-mail: axy@vip.hr

Abstract - The terms of logistic management in organizations and the use of information and telecommunication technologies in support of the management are inextricably intertwined. The fundamental benefit which business systems obtain by using logistics business information systems is the optimization of business processes. By using the logistics information systems the focus is shifted to those business processes that result in the highest added value. The highest level of development of logistics business information systems results in the construction of an integrated logistics information system that allows data mining processes and supports managerial decision making. This paper analyzes the development of logistics information systems, their types and functionalities in the context of modern business environments.

I. INTRODUCTION

To understand the logistics information systems it is necessary to clarify the concepts of logistics management and computerization. Logistics is the science which developed from the war planning, and its modern form dates from the 17th century [1]. The use of computers in business began to grow significantly in 1950s after the launch of the first commercial electronic computer UNIVAC [2] [3]. The computer, as well as logistics, was first used to obtain military superiority (for example in deciphering enemy messages) but the link of the two concepts is not the original military purpose, but the manipulation of data and information [4].

The logistics management can be interpreted through the general division of the business system, consisting of input and output information, and executive and management processes. Therefore, each logistics function takes and processes the data necessary for the operating of core activities, but the logistics management level integrates the relevant data, adds importance and value, and uses the information to debug and optimize the logistics processes [5].

A growing number of companies at the end of the 20th century launched a redesign of their business processes [6]. The classic form of business had to change completely because information has become one of the basic business resources. Reengineering or redesign of the business processes changed the traditional approach to business based on the work, natural resources and capital. The construction of information systems in enterprises created a new market division by applying the modern work concepts and electronic business [7] [8].

Although the modern information systems have evolved to a great extent, the human component remains indispensable because the information system as useful only to an extent to which a person is able to analyze and make use of the information provided. Likewise, the business information system is useful only if well implemented to pursue the objectives of enterprise. The company is not an isolated system, and in decision-making and management the wider environment must be considered [9]. The information system must support the management in a way that in addition to information related to the company it includes the external information available. This provides a comprehensive solution on the basis of which the management makes their decisions [7].

Smaller companies develop their information systems to the level they need for decision making. If the management of a small company has a clear insight into the business and the entire logistics process, they will opt for an operational and/or executive information system layer. For such companies, the most important goal is to automate the input and output of data and to simplify the processes in order to save time. If there is no heterogeneity of business functions, data discrepancy or poor communication, then there is no need for data warehousing, complex reports and analysis [9] [10].

The combination of various forms of information systems with the integration of the strategic system creates an integrated information system. The information is input within one organizational unit and is simultaneously used in all the other units when needed. Such integrated information system links the activities in the logistics chain horizontally and vertically, and the processing time is reduced to a minimum. The largest advantage of such systems is the creation of a homogeneous organization on the basis of sharing information, therefore the integrated information systems form the basis of virtual organizations [7].

II. THE ROLE OF LOGISTICS SYSTEMS IN IMPROVING BUSINESS PROCESSES

The purpose of the logistics system is the optimization of business processes and maximum utilization of resources. Compared to the traditional profit-oriented companies, logistics companies leave the narrow-minded orientation directed toward financial framework of revenue and expenditure. With the development of modern logistics at the beginning of the 20th century, successful business could no longer be based only on the retrospective financial indicators. These were the beginnings of modern business, more complex and yet more simple. Its complexity lies in the extensive surveillance and monitoring of the business functions. In the short term any such company makes great effort to get updated information from all business functions, in order to serve the business purpose. Data from one of the business functions, such as procurement, must be useful, unchanged, visible and usable in all other functions, from production to distribution. On the other hand, the establishment of logistics information network within the company enabled the management (at all levels) the visibility of the entire material, information and financial flows. This in the long term led to more manageable business processes [10].

Companies which utilize the logistics management model focus on flows and indicators that give added value. The logistics concept is not about producing to cover all costs and achieve a profit; it is about fulfilling customer requirements in the shortest possible time and at the right time. If customer demands are met, the company receives positive feedback, and thus the financial gains.

The modern business discards the traditional physical and intermediary model and reduces the number of active participants in the logistics chain. That does not mean the end of continuous structures with actors from production, procurement, transportation, storage, wholesale and/or retail trade and end customers, but the integration of certain functions in one segment. The inclusion of heterogeneous functions was enabled by the development of computer networks. Considering the various functions as one means faster response to unpredictable situations [11].

Through ICT solutions, the data from the logistics chain arrive into one place, a database, and can be extracted at any time and in different ways, depending on the purpose. A quick way of entering data in procurement, storage, production sales, distribution and the possibility of mutual exchange of data provided the basis of the logistics management model. Complete information support between business functions is needed to optimize the space- time flow of goods, energy, information and knowledge. Furthermore, by linking various stages in the logistics chain, the company achieves lower operational costs mainly related to supplies or stock. Optimization of production within the logistics chain requires that the product comes to an end point or to the end customer in the shortest possible time. Information linking and dissemination of information between participants in production, distribution and sale reduces the waiting time resulting in faster cash flow and positive customer reaction. Information and communication leads to the integrated logistics information system, and integrated logistics information system leads to fully optimized logistics system [1] [12].

III. INFORMATION TECHNOLOGY AND LOGISTICS INFORMATION SISTEMS

Although lowering the costs is the main reason for implementing ICT solutions, another factor appeared later in the evolution of ICT in business. It is the duration of the business process. After initial calming of the information revolution and the profiling of each of the companies that have successfully carried out the adjustment by applying an appropriate information system in every logistics chain, the new types of competition arose. In that regard, there was a focus on the customer and meeting his needs. Companies developed strategic plans in which the objectives were the development of new quality products or services in a more efficient and effective way, as soon as possible. Information technology became a tool to save time in the business. Modern companies with their ICT systems received not only the easier view of the entire logistics chain, but the system on which they can rely on, which can give them a clear view of the previous operations, whose storage capacity and processing time surpasses humans and saves a lot of time and money in planning, development, procurement, distribution, sales, marketing, etc. The other reasons for ICT implementation may include: modernizing administration, better cash flow, energy conservation, and accelerated distribution of vital resources [13].

IV. COMPONENTS AND FUNCTIONS OF THE LOGISTICS INFORMATION SYSTEM

In logistics, information systems maintain constant communication between vertical and horizontal positions in the logistics chain. Information support enables more control over the current business as well as greater insight into the previous business, on the basis of which a business decision can be made to change and improve. The goal of information technology is not the automation of all business activities as this would lead to the elimination of the human factor which is, and always will be, a key link in the process of computerization. In logistics processes, the information technology plays a key role in the integration of information sharing, joint planning, coordination of business process and acceptance of new business models and technologies [13].

The components of logistic information systems are hardware (material and technical component), software (intangible component) lifeware (the human component), Netware (network) and orgware (organizational component [7]. The most influential component in the realization of the logistics information system is the human component - the users of the system who are directly involved from the development of software to using the final results. Therefore, all the people who

appear in the logistics chain as outsourcers, suppliers, transporters, distributors, etc. and use the logistics information system constitute lifeware. The orgware includes functional, coordinated and synchronized alignment of the remaining four components of logistic and information systems. Network connection is the prerequisite for information and communication links [7] [15]. Many types of computer networks exist, depending on the geographical location and structure, for example PAN (Personal Area Network), LAN (Local area network), HAN (Home Area Network), SAN (Storage Area Network), CAN (Campus area network), MAN (Metropolitan area network), WAN (Wide area network), EPN (Enterprise private network), VPN (Virtual private network) and GAN (Global Area Network) [16]. A large increase in network and Internet traffic is measured in exabytes today. According to Cisco forecasts, traffic will by 2018 exceed 715 Exabyte's per month, or 8.6 zettabytes per year. [17].

Some authors mention dataware, which represents the data in structured databases and on any other media. All information passing through the information system must be archived in a way to allow them to be easily accessed at a crucial moment. It is easy to conclude that if one component fails to create a connection with the whole, the system will not perform its designated tasks [18]. The data in databases and data warehouses are usually collected from various sources. Efficient database operation is based on the structured relationships between data from different sources, so a qualified user can extract, transform and read useful information and draw conclusions [19].

V. INTRODUCTION AND APPLICATION OF LOGISTICS INFORMATION SYSTEMS

Before the implementation of information technology or information systems, the management must decide which type of information system matches the business and logistics processes. First, it is important to plan which business functions the information system needs to support, improve, speed up or replace.

1960s saw a growth of both the customer's awareness and product portfolio to meet the growing demand. This was a major factor in the distribution channels, as the company had to work with multiple products and at the same time maintain the cost of transport and stocks. In the 1970s, the use of computers in business increased. The change was significant because the business processes were firstly redesigned and then computerized. One approach of business redesign in the 1970s was TQM or Total Quality Management [20]. In the 1980s, due to recession and rising interest rates, many companies re-focused on cost reduction, which gradually lead to the development of integrated logistics [21].

The information systems and database facilitated the process of monitoring the business and enabled business development based on continuous corrective procedures. Information systems almost completely eliminated the human error when combining data from different sources, but did not eliminate the need for specific human knowledge. Today, humans and computers work together, discovering new methods and new knowledge. These two spheres form a virtual organization and are the foundation of connecting with new partners, suppliers, distributors, customers and so on [3] [22]. What the information technology ultimately enables are: structured transactions, automation, improving information capabilities, shortening the cycle time, knowledge management, inventory control and reducing organizational complexity [23]. Information systems present an important and irreplaceable role in connecting with suppliers. Starting with the Just-In-Time concept in Japan and later in Europe, the relationship with suppliers was completely changed. To ensure that production flowed continuously, with the delivery at the right time and the right place, the suppliers gradually merged with the production and have become strategic partners [24].

The information link between companies and suppliers was enabled by the commercialization of Electronic Data Interchange (EDI) in the 1980s and 1990s. The company could achieve a cost advantage and establish a better market position by connecting closely with its suppliers, Through EDI system, a company can send timely information regarding inventory, debit lists, information regarding demand, or possible customer complaints, and thereby reduce costs of intervention, reduce unnecessary inventory, speed up the response time, etc. Of course, the information system can always be created as desired, thereby providing a control over output information, because the increasing links with external partners may lead to the risk of leaks of classified information [14] [25].

VI. THE APPLICATION OF INFORMATION SYSTEMS IN LOGISTICS ACTIVITIES

The company is a complex organization consisting of a variety of business functions in a business cycle. Looking from the perspective of a common information system, it can be concluded that it is impossible to apply only one business information system which will be equally beneficial to all business functions. For example, in the human resources department the employee database will be the most important (with their income, their working hours, annual leave and similar), while this information will be completely irrelevant in the production process.

It often happens that a variety of business functions use completely different information systems that meet the basic purpose, but slow down the process chain. For example, if the same material specifications must be input in procurement, warehousing, production or sales, it extends the logistics process and thereby increases the risk of errors at each entry. In order to solve such a problem an integrated information system has been developed, that connects all levels vertically and horizontally and provides optimum data dissemination within the company. It achieves the functional and process connection within the company, because

operational, tactical and strategic levels of management use the same information system. The integrated system has eliminated the problem of multiple information entry and has linked the process flow within the company. For this reason, the integrated information system has had greatest impact on the logistics processes. Also, it eliminated the potential conflict between the vertical levels because it enabled the use of a single system and the same data in all positions within the company [7] [9].

VII. INTEGRATED LOGISTICS INFORMATION SYSTEM

Logistics management model is applicable to every business aspect, in which the company wants the most efficient way to get to the finish line. Logistics means the coordination of system components and strategic linking to obtain a harmonious whole that leads to the pre- set target. Many companies are facing the problem of disunity of operating information and data within the organization. The same company can use a multitude of different databases: virtual repositories, systems for processing documentation data, electronic mail systems, structured and unstructured knowledge databases and more. Due to the use of information from different sources, the common problems are inconsistency and contradiction which leads to inefficiency of information systems to support management [26].

Often the concept of integrated logistics is mixed with the concept of supply chain management because they share the idea of a continuous and uninterrupted flow of information. Integrated logistics management implies the management of physical distribution, materials management, business logistics and business engineering. Integrated logistics is the process of anticipating customers' needs and desires [21]. From the definition, it is clear that the integrated logistics deals with all logistics activities: transportation, warehouse management, inventory, materials management and communication and information flow management. The last activity is crucial for the whole system because it holds the system together. For this reason, the study and development of integrated information systems should be a priority and main consideration in the implementation of integrated logistics.

The process of information management begins with collecting basic data related to transportation, raw materials, employees and so on. Such information is processed in the integrated logistics information system to finally reach integrated logistics managers. Different subsystems exist: for data processing, for research and development, for decision support, for reporting etc. All of them ultimately support and link the functional areas of integrated logistics [21].

The data processing system is comparable to customer service and is of great importance due to the direct contact with the customer. Computer solutions that are used at the lowest level of management are Transaction Processing Systems or Office Automation Systems. The system for research and development includes activities such as planning the integration of logistics and information technology, the coordination of information and logistics network, staff training, decision-making on outsourcing of the IT solutions and more. Such a system is backed up with strategic management level solutions. Decision support systems are also a part of the strategic and tactical levels of the organization and are based on analytical modeling. A collective term is sometimes Management Support Systems (MMS). MMS are based on simulation models, projections and database analysis for the purpose of finding accurate and relevant information for decision making.

The principle of knowledge discovery in databases is popularly referred to as data mining. Data mining is a method of extraction of knowledge from information and is often used at the highest levels of management [9]. Strategic managers apply decision support systems to extract data entered at the tactical or operational level and put it in context in order to reach solutions. Four types of management support systems exist [27]: Decision Support Systems (DSS), Group decision support systems (GDSS), Expert Systems (ES) and Executive Support Systems (ESS).

In order to support management in the decision-making process, various integrated information systems are being used, such as ERP systems (Enterprise Resource Planning), CRM (Customer Relationship Management), SRM (Supplier Relationship Management), PLM (Product Lifecycle Management) , BI (Business Intelligence) and others [28] [29]. They are essential for today's management and timely decision-making, but are not limited to the management role. If implemented, each employee (depending on his professional position) will have access to the system. In short, the employees are the ones who have direct contact with the system as they input information from their field of action, whereas the higher levels decide on the basis of reports that the system is creating, not on the basis of the information entered. The information system follows the information flow within the company. Regardless of the type of integrated information system, all of them are implemented in order to accelerate business processes, increase efficiency and reduce costs [28] [30].

VIII. CONCLUSION

Information, as the most important resource in the business, opened the business organization to the environment and encouraged internal reorganization and redesign of processes, structures and business functions. Companies have turned to modern way of doing business – the network structures that allow for greater autonomy of the individual business functions, while on the other hand the concentration of decisive power centralizes in the form of uniting all business information in a virtual unit called data warehouse.

As the information connection is critical in large companies, increased development of tools, processes and applications is required in order to better communicate between the disparate functions, locations and employees.

Because of more specific activities, production dislocation, integration with other companies, cooperation with international partners, and so on, the use of information technology in business is inevitable. Large corporations that work in different locations are using IT solutions for direct communication, but, more important, for the processing, storing and archiving information and data that is further processed and extracted in the form of reports required for the future operation of the company.

The existence of multinational companies, and international cooperation in general, would be unthinkable without the information support today. Most modern companies have invested heavily in the development of information activities and their ICT departments, which are now considered one of the central business functions. Logistics information systems provide greater autonomy and efficiency of business functions while imposing technical constraints to control the authenticity of the entered data. The entire business process takes place within the same system. Cooperation through the information system thereby reduces the operating costs of sending physical documents, saves time and ensures the timely dissemination of accurate information.

Some companies accept the information technology only to reduce costs of the overall business process, while others aim for the integration of business functions on the basis of information connectivity. Therefore, many companies see their vision in a successful integration process in order to speed up the logistics flow and communication, avoid loss of information, increase their competitiveness and in the end reduce operating costs.

REFERENCES

[1] Zekić, Z., 2001, „Logistički model dinamičke optimizacije poslovanja poduzeća“, Ekonomski pregled 52 (3-4)

[2] Brock, G.W., 2009, The Second Information Revolution, Harvard University Press, Library of Congress Cataloging-in-Publication Data, USA

[3] Srića, V., Spremić, M., 2000, Informacijskom tehnologijom do poslovnog uspjeha, Sinergija, Zagreb

[4] Ciglenečki, N., „Razvoj računala kroz povijest“, 20.04.2015: http://www.zbrdazdola.com/infobible/infobible/razvoj_racunala_kroz_povijest.htm

[5] Andrijanić, I., Grgurević, D., 2011, Poslovna logistika, Visoka škola za ekonomiju, poduzetništvo i upravljanje „Nikola Šubić Zrinski“, Zagreb

[6] Davenport, T.H., Short J.E., 1990, „The new industrial engineering: Information technology and business process redsign“, Management Review, MIT Sloan School, Summer 1990, vol.31/4, 09.03.2015: http://sloanreview.mit.edu/article/the-new-industrial-engineering-information-technology-and-business-process-redesign/

[7] Panian, Ž., Ćurko, K., 2010, Poslovni informacijski sustavi, Sveučilište u Zagrebu, Element, Zagreb

[8] Zelenika, R., 2005, Logistički sustavi, Ekonomski fakultet u Rijeci, Rijeka

[9] Čerić, V., Varga M., 2004, Informacijska tehnologija u poslovanju, Ekonomski fakultet u Zagrebu, Element, Zagreb

[10] Luić, Lj., 2009, Informacijski sustavi, Veleučilište u Karlovcu, Karlovac

[11] Pupavac, D., 2006, „Optimalizacija proizvodnje unutar logističkog lanca za 21. stoljeće“, Zbornik radova Ekonomskog fakulteta u Rijeci, vol.2

[12] Ivić, K., 2008, “Utjecaj informacijskog sustava na management logistike”, VIII. Znanstveni skup Poslovna logistika u suvremenom menadžmentu, Ekonomski fakultet u Osijeku

[13] Müller, J., 2001, “Upravljanje informacijskom tehnologijom u suvremenim tvrtkama te hrvatska poslovna praksa korištenja informacijskih tehnologija”, Ekonomski fakultet Sveučilišta u Zagrebu, Ekonomski pregled, 52 (5-6)

[14] Jujnović, I., 2011, „Utjecaj informacijske tehnologije na integraciju logističkih procesa“, 11. međunarodni znanstveni skup, Sveučilište J.J. Strossmayera, Ekonomski fakultet u Osijeku

[15] Mrkonjić, M., 2007, „Utjecaj informacijskih sustava za podršku odlučivanju na uspješnost poslovanja poduzeća“, magistarski rad, Sveučilište u Zagrebu, 15.04.2015: http://www.srce.unizg.hr/fileadmin/Srce/dokumenti/katalog_radova/8_2007_Mrkonjic.pdf

[16] Mitchell, B., no date, “Introduction to LANs, WANs, and Other Kinds of Area Networks“, ABOUT.COM,: 09.06.2015: http://compnetworking.about.com/od/basicnetworkingconcepts/a/network_types.htm

[17] CISCO, 2014, “Cisco Visual Networking Index: Forecast and Methodology, 2013–2018”, June 10, p. 1-41 09.05.2015: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf

[18] Segetlija, Z., 2007, “ Razvoj poslovne logistike”, Sveučilište Josipa Jurja Strossmayera u Osijeku, Ekonomski fakultet u Osijeku, Zbornik radova

[19] Palace, B., 1996, “Data Mining: What is data mining?”, Anderson Gradute School of Management at UCLA, Management 274A, June 1996, 10.03.2015 http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[20] Svijet kvalitete, 2012, „Potpuno upravljanje kvalitetom“: 12.04.2015 http://www.svijet-kvalitete.com/index.php/upravljanje-kvalitetom/368-potpuno-upravljanje-kvalitetom-tqm

[21] Bloomberg, D.J., LeMay S., Hanna J.B., 2006, Logistika, Zagrebačka škola ekonomije i managementa, Mate d.o.o., Zagreb

[22] Ivaković, Č., Stanković, R., Šafran, M., 2010, Špedicija i logistički procesi, Fakultet prometnih znanosti Sveučilišta u Zagrebu, Zagreb

[23] Meter, M., 2006, “Potreba intenzivnijeg korištenja i mogućnosti unapređenja računovodstveno informacijskih sustava u hrvatskim poduzećima”, Ekonomski fakultet Sveučilišta u Zagrebu, Ekonomski pregled, 57 (7-8)

[24] Department of trade and industry, no date, „The Evolution of Quality“: 12.04.2015. http://www.businessballs.com/dtiresources/quality_management_history.pdf

[25] Clarke, R., 1998, „Electronic Dana Interchange (EDI): An Introduction“, Business Credit, October 2011, vol.103, no.9, str. 23-25“, 28.04.2015 http://www.rogerclarke.com/EC/EDIIntro.html

[26] Asproth, V., 2007, „Integrated Information Systems – A Challenge for Long-Term Digital Preservation“, Interdisciplinary Journal of Information, Knowledge and Management, vol.2, str. 89-98, 15.04.2015 http://www.ijikm.org/Volume2/IJIKMv2p089-098Asproth233.pdf

[27] Mrkonjić, M., 2007, „Utjecaj informacijskih sustava za podršku odlučivanju na uspješnost poslovanja poduzeća“, magistarski rad, Sveučilište u Zagrebu, 15.04.2015. http://www.srce.unizg.hr/fileadmin/Srce/dokumenti/katalog_radova/8_2007_Mrkonjic.pdf

[28] Pavlić, H., 2005, „Racionalizacija logističkih mreža“, Znanstveni magistarski rad, Ekonomski fakultet u Rijeci, Rijeka

[29] Holt, E., 2008, „SAP Business All-in-one Overview“, YouTube video uploaded 25.09.2008, viewed: 23.04.2015. https://www.youtube.com/watch?v=9_YY6z-_rsw#t=235

[30] CRMSwitch, 2013, “CRM and ERP: What is the difference?”, uploaded: 08.08.2013, viewed: 09.05.2015 http://www.crmswitch.com/crm-value/understanding-crm-erp/

Development of the data warehouse model for public authorities accounts in Croatia

Miodrag Sretenović* Božidar Kovačić** Vladan Jovanović***
* MI&DA d.o.o., Karlovac, Hrvatska, m.sretenovic@mida.hr
** Odjel za informatiku, Sveučilište u Rijeci, Hrvatska, bkovacic@inf.uniri.hr
*** Georgia Southern University, Statesboro, USA, vladan@georgiasouthern.edu

Abstract – An ongoing process of modernization, rationalization and implementation of public administration reform, in order to increase efficiency and economy, would be infeasible without improvement and continuous investments in new information technology. The main task of public administration is to be a service to citizens and legal entities from their domain. With the purpose of improving its functions, public authorities must have complete information in order to increase the quality and speed of information exchange with all stakeholders in the process. This paper describes a model of data storage. It describes the process of building and organization of data warehouse and ETL process of extracting, cleansing and transformation of data in a multidimensional data warehouse model, which is based on the metadata model. Furthermore it describes a concept and basic methods used in the development of business intelligence systems for the construction and use of a data warehouse. In addition it states options for the analysis of the data in the warehouse, which are shown based on dimensional storage concept, and it gives guidelines for future development.

Keywords: Data warehouse, public administration, ETL process, data mart.

I. INTRODUCTION

The introduction of automated business processes to the domain of public administration reduces costs and increases its efficiency. Business processes in the domain of public administration generate a large amount of information that needs to be stored. Laws regarding the keeping of documents prescribe the obligation to keep such documents for a long period of time.

The reasons that contributed to the architecture and development of data warehouses was the inability of databases to support complex requests for processing in a reliable and safe manner, and an increasing need for analysing a large amount of data and information.

A. Business Intelligence

Business intelligence is the process of collecting data and information from internal and external business environments and their transformation into business knowledge used for making business decisions[1]. By applying business intelligence to the analysis and synthesis of information in the field of public administration and their transformations, makes it easier to make business decisions in order to increase the efficiency of public administration as a public service of society.

B. Data Warehouse

Data warehouse is a segment of business intelligence, provided that the data within the warehouse are filtered. There are several ways of using filtered data. For example, by making inquiries that cannot be made in transactional databases because data are unstructured or semi-structured, or because inquiries could slow down the usual operation of databases. Data warehouse represents a central source of previously cleaned and transformed data, which are used for providing support in making business decisions[2]. Data warehouse is a set of procedures, technologies and tools that enable end users to perform data analyses, which help in the decision-making process and improve information resources[3].

By implementing data warehouses provide long-term data backup, and provide an analysis of financial flows of public administration.

C. Data Mart

Data mart is a part of data warehouse, containing summary data, which has certain functions and options within a single segment or area of business operations[4]. One of the leading experts in the field of application and implementation of data marts is Ralph Kimball, who founded Red Brick Systems in 1986[5]. By applying the relational structure of databases, access to data warehouse has been optimized, which results in a significantly improved performance with respect to the former technological level of databases.

D. Data Warehouse Architecture

Data warehouse architecture includes the process of defining data warehouse elements, methods for creating and connecting warehouse elements in accordance with set requirements. Data warehouse architecture represents a formal description of a data warehouse system[6].

This paper describes the development of data warehouses in chapter two, and presents the concept of data warehouses in chapter three. Chapter four contains description of a data warehouse model, the model implementation is presented in chapter five, and conclusion in chapter six.

II. DEVELOPMENT OF DATA WAREHOUSE

When organising a data warehouse, the structure of local government needs to be taken into account. Units of local government consist of multiple administrative departments, with clearly defined scope of work.

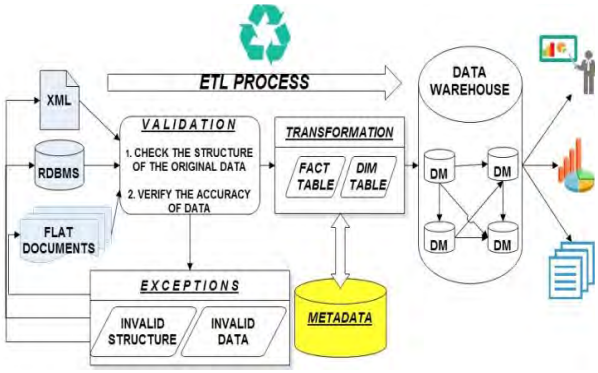


Figure 1. Shows the process of building a data warehouse.

In the organisation of data warehouse, this resulted in data marts for different administrative departments. The exact number of data marts varies according to the structure of local government.

Each change in legal regulations in the domain of public administration results in a change in the structure of databases at the source. Such changes increase requirements towards the data warehouse with respect to storing data and all previous changes, including the structure of data and metadata, covering a longer period of time. For this reason, agile methodology is crucial when it comes to the development of data warehouse[7].

When developing a flexible data warehouse, the ability to respond to legal and various user requirements becomes imperative. The solution is the application of agile methodology for designing a data warehouse.

Agile methodology represents a reaction to software development and enables fast presentation of concrete results and concrete answers to the changes in the environment[7]. Therefore, modelling requires:

- Conceptual modelling of data,
- Business process modelling according to ETL (Extraction, Transformation and Loading) process elements:
  - data extraction process,
  - data cleansing process,
  - data transformation process,
- Modelling of data marts within data warehouses,
- Enabling the presentation of data and flexible access to all relevant data, in order to allow system users further analyses.

A. *Conceptual Modeling of Data*

Conceptual modelling of data is a highly abstracted process of data modelling. The structure of data within the information system for tracking must reflect the structure of prescribed abstracted reports of local government. The conceptual data model is a complete, consistent and concise description of data within an information system[8].

B. *ETL Process Modeling*

The process of building a data warehouse is shown in figure 1 and starts by defining ETS processes (Extract, Transform, Load) [9].

- The first step includes extracting the data from data sources. The most common data sources are relational databases, with the exception of data stored in various file formats generated from incompatible applications. In the process of data extraction, it is crucial to trace the source of data, date and time of their entry and data on entry operators, in order to enable complete insight to auditors. Data Vault concept emphasises the need to leave a trace of the source and time of storing data in the database[10]. At the moment ETL tools pull data from relational databases, they pull metadata about their structure. Data structure in relational databases must reflect the final, abstract and prescribed form of report. Therefore, we may say that the tax data warehouse is based on the model of metadata and the model of prescribed legal rules.
- Transform means the transformation of data from the source to the warehouse. The preparations for loading data include an abstracted process of preparation and transformation of format and structure, validating the accuracy and quality of source data. Therefore, it requires an automatic process of cleaning and removing useless, incomplete or inaccurate data and returning them to the source for reprocessing.
- Load is a process of loading transformed data into the warehouse. These data will be suitable for further processing and will enable the user to access the creation of user reports.

C. *Modeling of Data Marts*

Each public administration data warehouse must contain multiple data marts, which are coordinated and synchronised, and can be classified as supermarts according to their structure. Supermarts consist of multiple mutually coordinated and synchronised tables. Table synchronisation is performed using foreign keys coming from conformed dimension. Dimension is a set of data that have been physically connected into multiple database tables using the same structure of attributes and keys[11]. The organisational structure of public administration must be reflected in the organisational structure of data warehouse, by mutually connecting the data marts of administrative departments using shared dimensions and dimension attributes.

D. *Modeling of Reports*

Part of public administration reports is abstracted and defined by precise legal content, while the rest enables the designers to create user reports. The ultimate aim of data warehouse is to enable end users a fast access to high-quality information.

III. CONCEPT OF WAREHOUSE FOR PUBLIC ADMINISTRATION INVOICES

The organisational structure of local administration consists of administrative departments. Administrative departments consist of several sections, depending on their scope of authority and operation.

The task of the administrative department for accounting and finance is to perform payment operations, record income, expenditure, payables and receivables

towards all natural and legal persons in the process, and to calculate salary and other benefits for employees[12]. Sources of local government revenue include[13]:

- city and municipal taxes,
- city and municipal duties and fees,
- city and municipal fines,
- other revenue of city and municipal budgets,
- revenue from foreign aid,
- revenue from ongoing assistance within general
- government,
- revenue from capital assistance within general government.

Expenditure of public local government is associated with planning, execution and control of expenditure realization. Each expenditure is documented by certain debtor-creditor documents, which consist of:

- incoming invoices,
- administrative documents and local government
- decisions.

It should be noted that within the segment of decentralization and transfer of power to local government units, a part of city and municipal revenue is being tracked through bookkeeping documents of local government. For the rest, the realization of expenditure is tracked exclusively at the level of local government. For this reason, a data warehouse should be planned and built with quality. Figures 2 and 3 show Entity-Relationship Diagram (ERD) of data marts, according to Martin Entity-Relationship Diagram[14]. Entity is anything we wish to collect and store information about, entity-relationship method is a diagram of mutually connected groups of data within a system in question [15].

Local government must comply with legal regulations from the domain of accounting and financial operations of non-profit organisations, and must keep track of the list of assets and liabilities, the principle of presenting assets, liabilities and own resources and the recognition of income, expenditure, receipts and expenses, financial reporting, presentation of receivables within business records and bookkeeping documents.

Fundamental documents for recording payables and receivables, prescribed by law, are as follows:

1. incoming invoices register,
2. outgoing invoices register.

A. Local Government Outgoing Invoices Register

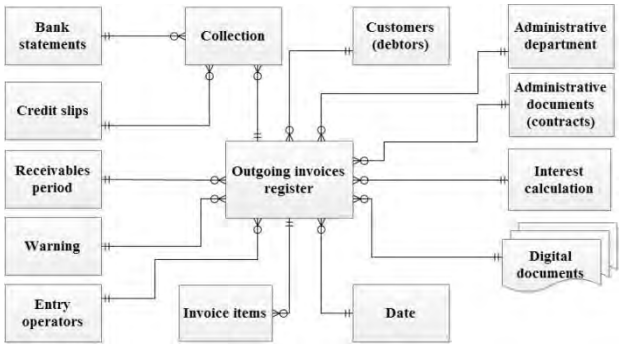


Figure 2. Local Government Outgoing Invoices Register

Outgoing invoices register in figure 2 represents receivables from customers (natural or legal persons) based on the delivery of goods, services or a decision of local government with the character of a receivable.

All receivables are grounded in administrative documents-cases, and must be stored as an integral part of debtor-creditor documents.

From the entity-relationship diagram, entity relationships of outgoing invoices register with the following entities are visible:

- *Collection* – represents the entity of collection of receivables, consisting in its structure of collection entities through:
- *Bank statements* – collection performed through transaction account,
- *Credit slips* – collection performed in cash at the local government point of sale,
- *Customers* (debtors) – legal or natural persons against whom financial resources are claimed,
- *Administrative department* – part of organisational structure of local government,
- *Administrative documents (contracts)* – legal documents for debtor-creditor relationship,
- *Interest calculation* – data on interest accrued due to untimely collection,
- *Invoice items* – data on products-services based on which receivables have been made,
- *Entry operators* – data on persons creating the document,
- *Warning* – document used in case of untimely collection,
- *Receivables period* – data on receivables period and the occurrence of debtor-creditor relationship,
- *Digital documents* – digital administrative documents based on debtor-creditor relationship,
- *Date* – date of change that has occurred.

B. Incoming Invoices Register

Incoming invoices register in figure 3 represents payables towards suppliers-creditors (natural or legal persons) based on received goods, services or a contract regulating contractual relations.

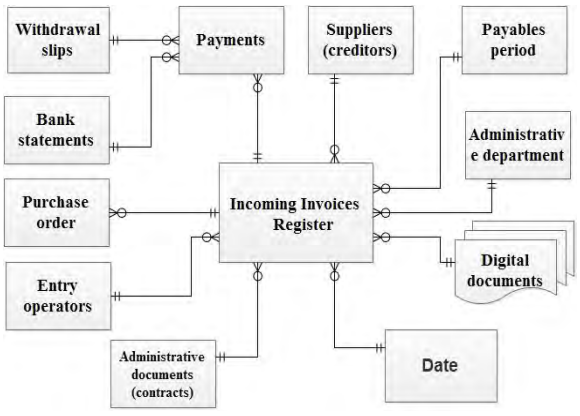


Figure 3. Local Government Incoming Invoices Register

All payables are based on documents in form of invoices, payment certificates, contracts with the character of an invoice and debtor-creditor relationship, and such documents need to be stored as an integral part of debtor-creditor documentation.

From the entity-relationship diagram, entity relationships of incoming invoices register with the following entities are visible:

- *Payments* – represents the entity of debt payment, which in its structure consists of payment entity through:
- *Bank statements* – payment performed through transaction account,
- *Withdrawal slips* – payment performed in cash at the supplier's point of sale,
- *Suppliers* (creditors) – legal or natural persons to whom certain amounts are due based on received goods or services,
- *Administrative department* – part of organisational structure of local government,
- *Entry operators* – data on persons recording the document,
- *Purchase order* – data on the number of documents based on which goods or services are requested from the supplier,
- *Payables period* – data on the payables period and the occurrence of debtor-creditor relationship,
- *Digital documents* – digital documents based on debtor-creditor relationship,
- *Administrative documents (contracts)* – legal documents for debtor-creditor relationship,
- *Date* – date of change that has occurred.

Control processes are indispensable in the process of invoice recording. Only perfectly reliable persons can be allowed to access record control systems as well as invoice liquidation systems. Due to an extremely sensitive area of indirect work with financial transactions and bank accounts, it is crucial to introduce the Data Vault concept, which emphasises the need to track the source and moment of storing the data into database[10].

IV. WAREHOUSE MODEL FOR PUBLIC ADMINISTRATION INVOICES

Kimball's model [16] was used as a model for designing the warehouse model for public administration invoices. In the dimensional approach of warehouse architecture, adopted by Kimball's model[16], data are divided into fact tables, which are connected to one or more dimension tables, which contain attributes, and with the application of adequate tools, enable multidimensional view to the user[17], which we call a "hypercube", based on its structure. Within the warehouse, dimensions are independent by their structure, so they can be shown as an n-dimensional coordinate system. Public administration data warehouse consists of multiple data marts, whose exact number depends on the structure of public administration. Incoming and outgoing invoices registers are located within the data warehouse each in its own data mart. Incoming invoices and outgoing invoices data marts are connected through the use of joint dimension tables.

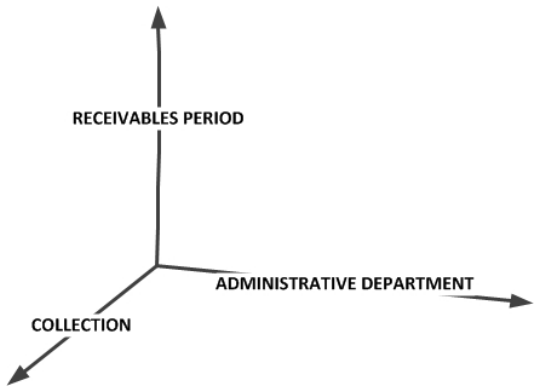


Figure 4. Three-dimensional display of books outgoing invoices local government

Figures 2 and 3 show the organisational scheme of data marts, with fact tables and dimension tables visible. By modelling a "hypercube" for storing mandatory registers, we enable tracking of payables and receivables by using dimension tables. This enables a complete financial insight into payables and receivables per various categories of dimension tables at any time.

Advantages of modelling:

- separating the transactional part from the warehousing part of information system.
- providing a complex overview per various dimensions.
- security.

An example of three-dimensional representation of outgoing invoices register in figure 4 provides an insight into the status of collection within a certain period and per administrative departments. By the same procedure, n-dimensional representations can be created by combining dimension tables. An example of dimensional representation with six dimensions includes the following dimension tables: administrative department, collection, receivables period, warnings, interest, customers-payers.

An example of five-dimensional representation of incoming invoices register in figure 5 provides an insight into the status of payments of suppliers-creditors within a certain period and per administrative departments and purchase orders.

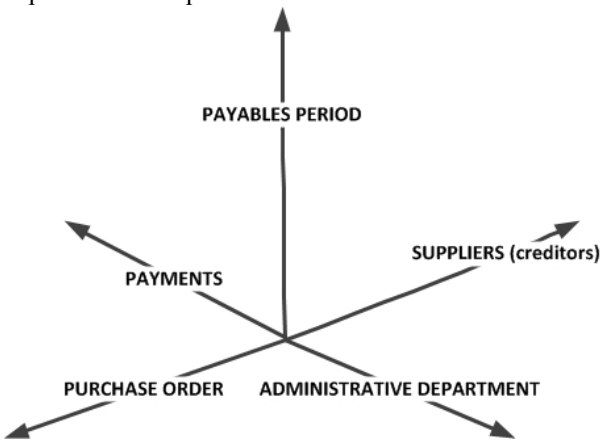


Figure 5. Five dimensional view of book incoming invoice of local government

Naziv — Naziv - naslov Adresa, Mjesto, — **Knjiga PRIMLJENIH(ULAZNIH RAČUNA)**

OIB — oib — 27.09.2015 — 27.09.2015 — Stranica 1

	RAČUN		DOBAVLJAČ (ISPORUČITELJ DOBARA I USLUGA)			
Rdbr	Broj računa dobavljača	Datum	Dobavljač - vjerovnik	OIB	IZNOS	Interni broj
1	2	3	4	5	6	7
343	0000225960102	27.09.2015	VIPnet d.o.o. Vrtni put 1 10 000 Zagreb	29524210204	9.675,75	33-09-2015
344	2340-	27.09.2015	PRIVREDNA BANKA ZAGREB d.d. Radnička cesta 42/I 10000 ZAGREB	02535697732	437,49	35-09-2015
				UKUPNO	10.113,24	

Figure 6. The book incoming invoice of local government

Naziv — Naziv - naslov Adresa, Mjesto, — **KNJIGA IZDANIH (IZLAZNIH RAČUNA)**

OIB — oib — 07.10.2015 — 07.10.2015 — Stranica 1

	RAČUN		KUPAC (PRIMATELJ DOBARA ILI USLUGA)		
Rdbr	Broj računa	Datum	Kupac	OIB	IZNOS
1	2	3	4	5	6
401	3992015	07.10.2015	HP-HRVATSKA POŠTA d.d. Branimirova 4 10000 ZAGREB	87311810356	5.000,00
402	4002015	07.10.2015	INA d.d. Avenija V Holjevca 10 10000 ZAGREB	27759560625	7.500,00
				UKUPNO	12.500,00

Figure 7. Book of outgoing invoices local government

V. IMPLEMENTATION

Structured reports of incoming and outgoing invoices represent indicators of total payables and receivables. Invoices as a bookkeeping document commit the debtor, and the creditor at the same time claims the amount stated on the invoice. Therefore, during the implementation all aspects of the protection of creditors and debtors through accurate bookkeeping must be taken into account. Figure 6 shows an example of public administration incoming invoices register. In the structure of reports shown in the incoming invoices fact table, the content of the incoming invoices fact table is visible. Mandatory possession of incoming invoices register is prescribed by legislation.

The incoming invoices register in figure 6 is a result of data search within the accounting period of the fact table INCOMING INVOICES REGISTER, as well as corresponding dimension tables, which is shown in figure 3. The incoming invoices register registers all the invoices received from the suppliers for received goods or services, and consists of: date of bookkeeping change that has occurred,

- data on suppliers of goods or services,
- data on invoice number of the supplier,
- data on the amount due to the supplier,
- data on the internal record number,

Figure 7 shows an example of public administration outgoing invoices register. In the structure of reports shown in the outgoing invoices fact table, the content of the outgoing invoices fact .table is visible. Mandatory possession of outgoing invoices register is prescribed by legislation. Outgoing invoices register is a document recording all invoices issued to customers, from which the following can be read:

- date of bookkeeping change that has occurred,
- data on customers receiving goods or services,
- data on the amount of receivables from customers,

The report is created within an accounting period and is a result of data search from the fact table OUTGOING INVOICES REGISTER, as well as corresponding dimension tables, which is shown in figure 2. A warehouse structured in this way is able to show a large number of reports shown in figures 4 and 5, with the exception of fundamental bookkeeping documents in figure 8, which shows the report "supplier invoices card". The said documents chronologically show obligations and payments. The purpose of this document is to provide information about the status of obligation; this is the document used for coordinating the status of obligation with the supplier. This representation includes the following dimension tables, which is visible from figure 3 of the entity, namely: Payments *(Bank statements,Withdrawal slips),* Suppliers (*creditors*) and Date. Figure 9 shows the content of the document "Statement of customer's open items on the reporting date". The said document shows the structure of invoices not paid by customers, and serves for coordinating the status of receivables from customers.

This overview includes the following dimension tables, which is visible from figure 2 of the entity, namely: Collections (*Bank statements,Credit slips),* Suppliers (creditors) and Date. The future research needs to be directed towards architecture and building of a centralised public administration data warehouse, where data of all units of local government would be stored. This would enable a unique overview of all relevant sustainability indicators of individual units of local government. This would also enable auditors to have a complete overview of all elements. Such architecture and building entails multiple benefits, from more cost-effective maintenance, centralised system of tracking costs, to the overview of all revenues of local government units. The introduction of data warehouses would mean the following improvements for local government:

- separating the transactional part of information system from the data warehouse,
- reducing pressure on operational databases in the sense of operation,
- analysing data and creating reports faster,
- multidimensional analysis of input-output parameters,
- current status of payables and receivables,
- reducing the maintenance costs,
- increasing the efficiency of business process,
- increasing the satisfaction of debtors-creditors with the speed and accuracy of data,
- data security.

When building the architecture and the centralised data warehouse for public administration, it is necessary to pay special attention to trends and changes in legal regulations.

Naziv - naslov
Adresa, Mjesto

Stranica 1

Promet kartice dobavljača 76

HRVATSKE AUTOCESTE d.o.o.
Širolina 4
10000 ZAGREB
57500462912

Od datuma: **01.01.2015**
Do datuma: **31.12.2015**

Rdbr	Broj UF	Dat.knj	Dat.dok	Rok	Dospijeće	Opis	Duguje	Potražuje	Saldo
1	8-02-2015	10.02.2015	10.02.2015	0	10.02.2015	37727-607-600		**1.173,50**	**1.173,50**
2	8-02-2015	10.02.2015	10.02.2015			IZV 26	**1.173,50**		**0,00**
							1.173,50	**1.173,50**	**0,00**

Figure 8. Supplier card

Naziv - naslov
Adresa, Mjesto,

Stranica 1

Izvod otvorenih stavki kupca na dan izještaja 31.12.2015 15:50:24

INA d.d.
Avenija V Holjevca 10
10000 ZAGREB

Vaš račun prema našim poslovnim knjigama se sastoji od slijedećih otvorenih stavaka

IF	Datum	Rok	Dospijeće	Opis	Duguje	Potražuje	Saldo
400-2015	07.10.2015	0	07.10.2015		7.500,00	0,00	7.500,00
					7.500,00	0,00	7.500,00

DUŽNIK:

NAPOMENA: Osporavam iskazano stanje u cjelosti-djelomično za iznos od: Potvrđujem usklađenost otvorenih stavki

iz ovih razloga:

datum i mjesto potpis

Figure 9. Statement of customer's open items on the reporting date

VI. CONCLUSION

The work is based on Croatian legislation, which is compatible with international law. The paper describes the model of data warehousing with respect to public administration invoices, with the application of ETL process. By implementing tools and technologies that can provide support in managing invoices, the management and control of the status of payables and receivables is also achieved, which increases financial discipline. In the process of planning and analysis, we achieve the possibility of realistic planning and analysis, of planning the exact resources available, and as a result of realistic expenditures.

The aim of this paper is to help the authorized persons decide on the creation of a warehouse, and to propose guidelines and a model for building a warehouse to designers, with the application of proposed model.

In circumstances of compliance with legislation which provides for high fines in case of lack of or possession of inaccurate tax records or their possession on non-compliant forms (reports), and in circumstances in which the security of information systems requires increasing amounts of human resources, as well as financial resources, the need for creating data warehouses has become imperative.

Next steps in the development and implementation of data warehouse public administration is the choice of hardware components, and then to the described model selection destiny data, functional verification and active use of the warehouse.

REFERENCES

[1] Javorović, B., Bilandžić, M.: Poslovne informacije i business intelligence, Golden marketing – Tehnička knjiga, Zagreb 2007.

[2] O'Brien J. A., Marakas G. M., Management Information Systems, McGraw-Hill, New York, 2006

[3] Data Warehouse Design: Modern Principles and Methodologies, Matteo Golfarelli and Stefano Rizzi, McGraw-Hill, 2009.

[4] Golfarelli M., Rizzi S., Saltarelli E.: WAND: A CASE Tool for Workload-Based Design of a Data Mart. SEBD 2002

[5] http://www.kimballgroup.com/html/about.html

[6] Huang X., Data Warehouse Architecture: Practices and Trends, IGI Global, 2010.

[7] R. Huges, Agile Data Warehousing for the Enterprise- A Guide for Solution Architects and Project Leaders, MK 2015.

[8] Mladen Varga, Baze podataka, Društvo za razvoj informacijske pismenosti (DRIP) 1994,

[9] R.Kimball, J.Caserta, "The Data Warehouse ETL Toolkit. Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data", 2004

[10] D.Linstedt, Olschimke, Building a Scalable Data Warehouse with Data Vault 2.0, MK 2015

[11] Casters M., Bouman R., van Dongen J., Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration, Wiley Publishing, 2010

[12] Koprić I., Marčetić G., Musa A., Đulabić V., Lalić Novak G.; Upravna znanost - Javna uprava u suvremenom europskom kontekstu; Zagreb: Pravni fakultet Sveučilišta u Zagrebu (2014)

[13] The law on the financing of local and regional governments (Narodne novine 117/93, 69/97, 33/00, 73/00, 127/00, 59/01, 107/01, 117/01, 150/02, 147/03, 132/06, 26/07, 73/08, 25/12, 147/14, 100/15)

[14] Martin, J. Principles of Object Oriented Analysis and Design Prentice Hall, Englewood Cliffs, 1993

[15] Mile Pavlić, Oblikovanje baza podataka, Rijeka: Odjel za informatiku Sveučilišta u Rijeci 2011,

[16] R. Kimball, M. Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. Wiley Computer Publishing, 2002

[17] Budišćak, Ivan; Mornar, Vedran; Mlinarić, Danijel. Analiza podataka u svrhu predlaganja studijskih programa u Nacionalnom informacijskom sustavu prijava na visoka učilišta // *MIPRO 2014 37th International Convention Proceedings* MIPRO, 2014. 1672-1676.

The Future of Digital Economy in some SEE Countries (Case study: Croatia, Macedonia, Montenegro, Serbia, Bosnia and Herzegovina)

Marijana Vidas-Bubanja, PhD *, Iva Bubanja, Mr.sc.
* Belgrade Business School, Belgrade, Serbia
** Belgrade Business School, Belgrade, Serbia
E-mail: marijana.bubanja@bbs.edu.rs; ivabubanja@gmail.com,

Abstract - This paper examines some South East European countries` readiness to develop digital economy by analyzing and comparing several key measuring indicators. Although these countries have created the basement for digital economy development, more intensive activities are needed in the future in order to create proper environment for further development of digital economy (education, innovation, organizational transformation, etc.). For these countries digital economy will be an opportunity to realize long-term economic development and overcome recession.

I. INTRODUCTION

Emergence of information communication technologies (ICT) as the key general purpose technology in the last 40 years affected almost every aspect of economic and social activities. The world has been experiencing positive effects from ICT on its economy, business growth, and living standards. Both developed and developing countries are trying to put in place right enabling conditions in order to tap all the innovation and growth opportunities offered by proper ICT implementation in different sectors of national economy and society.

Selected South East European (SEE) countries for our analysis (Croatia, Macedonia, Montenegro, Serbia, Bosna and Hercegovina) are faced with economic and financial instability and budgets constrains. For them, it is very challenging to overcome all uncertainties and barriers for the role ICT can play in economic and business performance. Although in the past two decades ICT have generally improved their performance in analyzed SEE countries and growth of ICT sector has been reasonable strong, they still need to reform their economies in the way that allows ICT to become "the great enabler" it can be.

II. WHAT IS DIGITAL ECONOMY

The 19th and 20th centuries saw three industrial revolutions – mechanization, electrification, and automation. The 21st is seeing the fourth – digitization. Digital opens up unprecedented possibilities. Digital technologies and innovations are powerful, pervasive and have multiple, indirect impacts. These innovations are changing economies and markets, and reinventing relationships between organizations, suppliers and customers thus becoming critical elements for growth, innovation and job creation.

Paper realized by research under the project 179038 sponsored by Ministry of education, science and technol. development of Rep. Serbia

The digital economy isn't just about speeding up communication across borders or changing the skills workers need; it is about changing the very nature of consumption, competition and how markets operate [1]. Such reformation toward digital is very comprehensive and profound. The success in this digital world can only happen by embracing change, that is, by having pledged massive investment in future-oriented technologies, solutions, and business models. That is, only widespread and systematic use of ICTs by all stakeholders—individuals, businesses, and government—can trigger transformation to digital economy.

Though affecting every industry and business function, the impact and pace at which digitalization takes place differs across industries and businesses. The future of doing business in the digital age will mainly depend on the pace of digital development and adoption of new technologies, as well as on the outlook for the global and national business climate.

III. ADVANTAGES OFFERD BY DIGITAL ECONOMY

From an economic point of view, ICTs boost productivity and reduce transaction and information costs, but their impact extends well beyond productivity gains. For example, for Europe is estimated that one third (0.7 percent) of the average 2.2 percent GDP growth rate from 1995-2007 can be traced to ICT, of which more than half came from investment in ICT, one third from productivity gains in ICT production, and the remainders from productivity through ICT use [2].

In the same time, ICTs allow new models of collaboration that increase workers' efficiency and flexibility. Over a short period of time, digital technologies have moved from simply allowing businesses to do what they've always done, but more efficiently and effectively, to enabling businesses to completely change not only what they are doing, but also the business models they use to create value [3]. In this way ICTs foster entrepreneurship and create new business models.

The modern company supports the idea of digital disruption in order to realize all potentials across a spectrum of digital opportunities by building on the way it currently uses digital technologies and organizes business processes. 'Digital disruption' refers to changes, both positive and threatening, and will affect companies in three ways [1]:

- Customer insights combined with the ability to reach out to customers more effectively.
- Operating models - the way daily operations and processes are organized.
- Business models - the way value is created, delivered and captured.

Digital innovation is also driving a significant shift in the balance of power between organizations and individuals. The explosion in connectivity and the availability of information is putting today's consumers, employees, citizens, patients and other individuals in a controlling position [1].

Enterprises must go further than just responding to and meeting customer needs. A higher-level of user-centricity is required that involves users in every business process, from requirement collection, product ideation and design, R&D, testing, production, marketing, to after-sales services. Only by leveraging users' collective wisdom can enterprises prosper, together with customers.

With the reach of broadband expanding, and everything else smartening up, new industry ecosystems are being created, thus reshaping the business landscape to suit the needs of the information era.

ICTs are vectors of not only economic but also social transformation. ICTs offer significant social benefits, notably by enabling access to basic services, including financial services and education. Widespread ICT use by businesses, government, and the population at large is a precondition for all these benefits and opportunities to materialize [2].

IV. WHY ANALYZED SEE COUNTRIES NEED DIGITAL ECONOMY

The digital society and economy represent an irreversible trend. An inevitable choice for either developed or developing countries is to embrace digital reformation by applying the Internet mindset and ICT technologies in order to establish competitive strengths for today and the future. This is true for analyzed SEE countries too, as they have to accelerate the transformation of its business landscape through the development and smart use of digital technologies.

Faced with stagnant or decreasing growth rates, rising unemployment and worsening living standard of their citizens in the last decade analyzed SEE countries need new digital sources of growth that will enable further reform, modernization and innovation in order to boost growth rates and create employment (Table I).

TABLE I. SELECTED ECONOMIC INDICATORS

	HR	MK	ME	RS	BA
GDP growth annual %					
2010	-1,7	3,4	2,5	0,6	0,8
2014.	-0,4	3,8	1,8	-1,8	0,8
GDP per capita constant 2005 US$					
2010.	10.524	3.690	4.525	4.107	3.325
2014	10.547	3.979	4.770	4.246	3.441
Consumer price index %					
2010.	1,0	1,5	0,7	6,1	2,2
2014	-0,2	-0,3	-0,7	2,1	-0,9
Current account balance (% of GDP)					
2010	0,7	-2,1	-23,0	-6,0	-6,1
2014	-0,4	-1,4	-15,2	-11,5	-7,7
General government gross debt (% of GDP)					
2010	52,8	24,1	40,9	43,7	39,1
2014	90,4	44,9	75,5	68,6	38,3
Unemployment rate % of total labour force					
2010	11,5	32,1	19,7	20,0	27,2
2014	17,1	28,0	19,1	19,7	27,5

Legend: HR-Croatia, MK-Macedonia, ME-Montenegro, SR-Serbia, BA-Bosnia and Herzegovina.
Source: http://knoema.com/atlas/topics/Economy

There are at least two reasons why analyzed SEE countries need creation of digital economy and society. The first one is connected with the potentials offered by digital transformation of existing industries and enterprises in selected SEE countries that will open an opportunity to increase productivity, competitiveness and therefore job-creating ability [3].

Innovating and transforming isn't easy and businesses in analyzed SEE countries are not transforming quickly enough. Embracing digital technologies means adapting processes, organizational structures and workforces to the digital world, and moving away from business as usual means taking risks and implementing new tools quickly and at scale.

These are challenges that not all companies in these SEE countries, especially the smallest ones, can successfully overcome. Many feel more comfortable with the status quo and will need a confidence boost to set out on the path to digital transformation.

Secondly, a deeper and more disruptive wave of digital technologies is already beginning to impact analyzed SEE countries and they need to be ready to take full advantage of it. This second wave includes advanced mobile communication, social media, cloud, big data analytics, smart devices, connected objects and sensors. These technologies, and the breakthrough innovations they enable, have massive transformative power, and will be crucial tools to enable a smart, sustainable and inclusive economy in the future (Table II).

Companies able to make use of this second wave of advanced accelerating technologies, and by doing so, are performing 10 times better than their peers. According to some authors this is the biggest transformation in business the world has seen in over a century [4]. The untapped potential here really is enormous and it represents big challenge for companies and industries in analyzed SEE countries.

TABLE II. THE SECOND WAVE OF NEW DIGITAL TECHNOLOGIES

Mobility and mobile apps: Technologies that enable voice and data connections between people, and increasingly between objects, while on the move. Applications that take advantage of this and in some cases make use of location data.
Social media: Enterprise social media describes companies' use of social media tools for business purpose. These tools may include social networks (e.g. Facebook, LinkedIn, etc.), microblogging (e.g. Twitter), blogs, internal wikis and/or other enterprise collaborative software.
Cloud: cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, software, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.
The Internet of Things (IoT): describes the network of physical objects that feature an IP address for Internet connectivity, and the communication that occurs between these objects and other Internet-enabled devices and systems.

Source: [3], pg. 10.

VI. PROGRESS TOWARDS DIGITAL ECONOMY IN ANALYZED SEE COUNTRIES

Progress to digital economy in selected SEE countries will be analyzed by using and comparing some key infrastructure development indicators such as fix and mobile telephone subscriptions, fix and mobile broadband subscriptions, individuals, households and enterprises using the internet but also complex measuring ICT index such as network readiness index. Except Republic of Croatia who has become full EU member state since 2013, all analyzed countries (Republic of Serbia, FJR of Macedonia, Montenegro and Bosna and Hercegovina) have been taking part in an ongoing process of the enlargement of the European Union. The collecting and comparing these data was possible as selected countries base their statistics in information society area on Eurostat methodology. The aim was to follow the progress in the development of digital economy and society in this part of SEE region in the framework of European digital agenda tasks.

Fix voice telephony market and mobile market - The decline in the total number of fixed telephone lines in the selected SEE countries have been accelerated since 2012, thanks to continued fixed-to-mobile substitution. Although there is reduced fixed line penetration rated per 100 population in all analyzed countries, there remains significant variation between countries, with penetration ranging from over 39% in Republic of Serbia, to just 22,2% in Bosna and Hercegovina (Table III).

TABLE III. FIX AND MOBILE SUBSCRIPTIONS

	HR	MK	ME	RS	BA	EU28
Fixed lines per 100 population						
2010	39,3	20,2	27,7	41,4	24,7	40*
2014	35,0	18,93	26,6	39,6	22,2	38,3
Mobile subscriptions per 100 population						
2010	143,5	106,1	173,9	132	87,5	124,4*
2014	104,0	108,69	163,8	130,04	90,0	120,5

*EU 27 Source: Eurostat and for 2010 [5] p.16 and 19. For Macedonia www.aek.mk

The total number of mobile subscriptions rises. Growth rates are not as high as before owing to penetration going over 100% in all countries except Bosna and Hercegovina.

Fix and mobile broadband subscriptions -The fixed broadband market in the selected SEE countries measured by the number of active connections grew but still is below the European average of 29,6 % in 2015 (Table IV) [6].

Mobile broadband is the fastest growing and most dynamic segment of the electronic communications market in this region as well. Despite the fact that many fixed broadband operators are offering users access to higher-speed packages, mobile broadband remains appealing, thanks to its flexibility. Indeed, in some areas where there is no fixed broadband coverage, mobile networks provide the only form of broadband access.

Mobile-broadband penetration levels are highest in Europe and the Americas, at around 78 active subscriptions per 100 inhabitants, while in SEE penetration rates are not that high [6]. In Serbia and Croatia the mobile broadband penetration rate (measured for all devices including handhelds and dedicated data cards/modems) exceeded 50% in 2014 and was close to the EU-27 average at end-2012 of 54%.

TABLE IV. FIX AND MOBILE BROADBAND SUBSCRIPTIONS

	HR	MK	ME	RS	BA	Europe
Fixed broadband subscriptions per 100 inhabitants						
2012	18,3	12,5	8,3	11,2	8,2	25,7
2014	20,7	13,7	8,4	12,9	10,6	28,6
2015	21,5	15,1	12,8	14,2	-	29,6
Mobile broadband subscriptions per 100 inhabitants						
2012	15,5	6,5	22,0	4,1	9,2*	54,2
2014	53,9	25,1	27,5	52,1	12,2	69,3
2015	65,3	38,0	23,1	53,7	-	78,2

*2013 Source: www.knoema.com/atlas for Europe [6].

Use of Internet in selected SEE countries - With regard to Internet usage, the Digital Agenda sets the target to increase regular Internet usage from 60% to 75% by 2015, and from 41% to 60% among disadvantaged people [9]. In all analyzed countries Internet usage by individuals is close to 70%, but still below EU average penetration of 79 in 2015.

In EU28 83% of households had Internet access in 2015 [10]. The number of households with Internet access is rising in analyzed part of SEE region as well. The highest proportion of households with Internet access in 2015 was recorded in the Croatia (77%) and Macedonia (69%). For all other analyzed countries more than 60% of households had Internet access at home (Table V).

The use of information and communication technologies influences the way that enterprises are run, internal communications organized, information shared with business partners, and customers communicated with. Nearly all enterprises in Serbia and Montenegro with at least 10 persons employed had Internet access in 2015. Percentages are little lower in Croatia (90) and

Macedonia (93). Many enterprises may consider it important to be visible on the Internet and close to three quarters (75%) of enterprises in SEE region had a website in 2015. The percentage of enterprises having web site is something lower in Macedonia (53).

TABLE V. USE OF INTERNET

	HR	MK	ME	RS	BA	EU28
Internet use by individuals %						
2010	54	52	37,5	40,9	42,7	68
2014	69	68	64,5	62,5	60,8	78
2015	70	70	67,2	65,8	-	79
Households with Internet access at home						
2010	56	46	51,4	39,0	29,8	70
2014	68	68	63,6	62,8	47,5	81
2015	77	69	67,5	63,8	-	83
Enterprises with Internet access						
2010	95	84	95	96,8	-	94
2014	96	93	98	100	-	97
2015	90	93	99	99,1	-	97
Enterprises with WEB sites						
2010	61	43	-	67,5	-	67
2014	66	53	73.3	74	63,2	74
2015	71	52	74,3	75,2	-	75

Source: http://appsso.eurostat.ec.europa.eu/nui/show.do
For Serbia: [7] and [8].

Use of e-commerce and e-government services - 50% of EU citizens should use e Government by 2015, with more than half of them returning filled-in forms [9]. Selected SEE countries also show rapid increase of e Government usage when suitable services become available, and Croatia and Montenegro are regional leaders in this respect in 2015 (Table VI).

TABLE VI. E-COMMERCE AND USE OF OTHER E-SERVICES

	EU-28	HR	SR	MK	MN
Individuals					
% of individuals using the Internet for interaction with public authorities					
2010	41	19	-	14	-
2014	47	32	37,4	27	22,0
2015	46	35	27,8	23	31,3
% of individuals using e-banking services					
2010	36	20	-	4	-
2014	44	28	13,5	9	-
2015	48	33	-	17,6	-
% of individuals using the Internet for ordering goods or services					
2010.	40	14	6,1	4	-
2014.	50	28	21,6	11	7
2015.	53	31	22,7	11	10,1
Enterprises					
	EU-28	**HR**	**SR**	**MK**	**MN**
% of Enterprises using the Internet for interaction with public authorities					
2013	87	88	87,6	74	-
2014	88	93	92,0	77	-
% Enterprises receiving orders online (at least 1%)					
2010	13	22	20,7**	4	8,1,*
2014	15	-	22,9	5	14,7
2015	17	20	-	7,7,	24,3
% Enterprises purchasing online (at least 1%)					
2010	27	23	40,1**	4	11,7*
2014	22	-	40,3	4	24,3
2015	23	11	-	3	24,1
% Enterprises using ERP					
2010	21	15	-	9	-
2014	31	18	16,2	22	43,8
2015	36	29	-	18	43,8
% Enterprises using CRM					
2010	17	11	-	11	-
2014	20	12	14,9	16	38,2
2015	21	15	-	14	37,7
% Enterprises using cloud services					
2014	32	16	3,2	-	12,5

***2012; ** 2011 Source:** Eurostat, for Montenegro www.monstat.org,

Another Digital Agenda target is that 50% of the population should buy online by 2015, with 20% buying cross-border [9]. This indicator is particularly low in all monitored SEE countries, same as the indicator of e-banking services use. SEE countries are significantly legging in this respect to EU28 average.

Despite having WEB sites, companies in selected SEE countries still are modest in use of Internet for purchasing or selling online. Around a quarter of companies in Croatia, Serbia and Montenegro receive orders online, and Serbia is the leader concerning the percentage of companies having purchased online in 2014 (40, 3%).

Generally, large enterprises made greater use of information technology than small enterprises, and sell more online what is similar to situation in EU. According to Digital Agenda Scoreboard for 2015 only 14.5% of European SMEs sell online [10]. Even the best performing EU countries are far from the EU-wide target of 33% by 2015. Large companies, on the other hand, are much more active, with 35% of them selling online. Having this in mid reached percentage of companies in monitored SEE countries is in line with their gradual progress on the road to digital economy.

The proportion of EU-28 enterprises that used enterprise resource planning (ERP) software applications reached 36% in 2015 while a smaller proportion of enterprises (21%) used customer relationship management (CRM) applications [10]. Enterprises in analyzed SEE countries are very slow in accepting these practices in their work and are lagging behind EU average in both aspects (according to national statistics the only except are the enterprises from Montenegro).

Network readiness index - Since 2001, The Global Information Technology Report series published by the World Economic Forum in partnership with Cornell University and INSEAD has measured the drivers of the ICT revolution using the Networked Readiness Index (NRI). The NRI 2015 covers 143 economies, which together account for 98.4 percent of world GDP [11].

Networked Readiness Index is a composite indicator made up of four main categories (*subindexes): 1)* Environment subindex (Political and regulatory environment, Business and innovation environment), 2) Readiness subindex (Infrastructure indicators, Affordability, Skills), 3) Usage subindex (Individual usage, Business usage, and Government usage), 4)

Impact subindex (Economic impacts, Social impacts). This represented one of the first attempts to make conceptual sense of the complex ICT reality, identifying the common factors that enable countries to use technology effectively [11].

Not unexpectedly, advanced economies are better than developing ones at leveraging ICTs. High-income economies dominate, taking the first 31 places in the overall NRI rankings in 2015. The performance of countries largely mirrors their position on the development ladder: a higher level of income is typically associated with a higher NRI score.

Europe is home to some of the best connected and most innovation-driven economies in the world. In particular, the Nordics—Finland (2nd), Sweden (3rd), Norway (5th), Denmark (15th), and Iceland (19th)—continue to perform well. Indeed, these five countries have featured in the top 20 of every NRI edition since 2012 [11].

TABLE VII. NETWORK READINESS INDEX

NRI	SR	MN	HR	BIH	MK
2009/2010 (138)	94	42	51	110	73
2010/2011(138)	93	44	54	110	72
2012 (142)	85	46	45	84	66
2013 (144)	87	48	45	78	67
2014 (148)	80	52	46	68	57
2015 (143)	77	56	54	-	47

Source: [11]. In the brackets are numbers of covered countries for that year.

According to performed analysis, selected SEE countries have created a good basement for digital economy development and ICTs are becoming increasingly affordable in these economies. NRI ranking, on the other side, represent complex SEE ICT reality, and proves that these countries are more or less lagging in efforts and commitment to fully develop and leverage ICTs to boost their economic development and competitiveness. Looking at the trends since 2009 reveals that all analyzed countries have improved their overall performance to some extent but they are still positioned in the middle of the 2015 list taking positions between 47th (Macedonia) to 77th Serbia [11].

Despite this progress, the analyzed SEE countries continue to suffer from low rates of e-business, a weak political and regulatory framework and a poor business and innovation environment. These factors affect their capacity to further leverage ICTs to boost their economies and benefit from higher rates of products and service innovation [12].

Our analysis highlights that monitored SEE countries need to adopt harmonious and comprehensive strategies that do not focus only on improving access to ICTs. While important, access is only one ingredient in the recipe for success. Improving the ecosystem for spurring entrepreneurship and strengthening the conditions that enhance innovation are also crucial to boost competitiveness and well-being, to enhance economic growth, and to create jobs in selected SEE economies.

VII. WHAT NEED TO BE CHANGED

In an economic environment of analyzed SEE countries that is threatened by the short-term concerns of financial instability and budget constraints and that suffers from a long-term structural growth deficit it is not easy to create conditions for digital transformation and proper use of ICT potentials. The future of the digital economy in selected SEE countries depends on the realization of five important tasks on the transformation road to digitally driven economy [1]:

1. Increase industry digital transformation. Increase the take-up and use of digital technologies by industries and companies (big and SMSs) in order to transform existing business and operating models thus fostering productivity and competitiveness.

2. **Create a digital entrepreneurial culture.** Improve the image of digital entrepreneurs and promote their role in society.

3. **Attract, develop and retain high-end digital skills and talent.** Increase the quantity and quality of digital entrepreneurial skills and talent; boost the development of a unique blend of creative, technology and entrepreneurial skills.

4. **Ease the access to finance and enhance investments.** Improve access to finance for each stage of enterprise development and support the growth of digital entrepreneurial ventures.

5. **Boost the digitally powered market.** Improve the ease of doing digital business by equally stimulating the demand and supply of digital technologies and creating economies of scale.

According to the results of performed analysis in monitored SEE countries there is rising number of citizens on the network and businesses that use ICTs have generally improved their performance. But despite of these trends, it is evident that infrastructure looks unfit to cope with future demands from the next wave in ICT, especially the use of big data and cloud computing.

For this reason the key pre-conditions for reaping the ICT growth benefits need to be secured by a high-quality and affordable infrastructure in all sectors, capable of supporting the growing cloud, big data, and including high-speed fixed and mobile broadband. Selected SEE countries still need to further develop hard infrastructure (broadband, broad access, interoperability).

In order to use infrastructure in a proper way monitored SEE countries also need to provide the skills able to exploit this infrastructure and all the opportunities ICT can create. This includes hard and soft skills: technical skills, business and management skills, and the technological savvy to identify business opportunities in new technologies and applications [12]. Entrepreneurial skills, which must be enabled by framework conditions,

including a dynamic business environment, are also lacking in analyzed countries, as is a cultural readiness to embrace technology and change. Government and business can work together to develop and foster the skills and willingness to use ICT.

Realizing a greater impact from technology and innovation depends on regulation (in ICT-related areas, but also in product and labor market regulation and other areas), the cost and ease of doing business, and access to finance. The complexity of regulatory frameworks ill-adapted to new technologies and innovations is also a huge barrier to reaping the benefits of ICT. That is why analyzed SEE countries urgently require some measures related to reducing regulatory barriers, improving market integration, simplifying administrative rules and procedures, and improving the allocation of funding budgets already available [14].

VIII. CONCLUSION

Digital technologies are one of the most important sources of growth for national economies. They enable economies to create more jobs, improve people's lives and build better and greener societies. However, performed analysis in selected South East European countries proved that despite progress in last decade in creation of basement preconditions for digital economy, the huge potentials of ICT still remains untapped. There is a need to stimulate a more innovative and entrepreneurial mind-set and accelerate smarter use of 'digital' technology in various sectors of economies in monitored countries.

Under current economic and political conditions policymakers in these SEE countries face pressures to cut spending, and ICT-related spending increasingly competes with alternative spending objectives. But, despite these circumstances, creators of ICT policy framework should put a stronger emphasis on the larger role of technology and innovation and on the transition of current business models to new digital forms. Infrastructure availability, educated work force, proper legal framework and simulative business environment are important conditions to be fulfilled and simulative topics for further examination and future work.

Only in this way analyzed SEE countries will be able to produce more and better goods and services at lower prices for European and world market. Digital economy is the great chance for selected SEE countries to use the opportunities afford by ICT in order to take their economies out of recession in the short run and to resume a healthier structural growth path for the medium and long run in the future.

REFERENCES

[1] Deloitte, Doing business in the digital age: the impact of new ICT developments in the global business landscape, Europe's vision and action plan to foster digital entrepreneurship, April, 2013.

[2] EC, Unlocking the ICT growth potential in Europe: Enabling people and businesses using scenarios to build a new narrative for the role of ICT in growth in Europe, Brussels, 2013.

[3] EC, Digital transformation of European industry and enterprises, A Report of the strategic policy forum on digital entrepreneurship, Brussels, 2014.

[4] Salim Ismail, Exponential Organizations, 2014.

[5] Cullen International, Monitoring regulatory and market development for electronic communications and information society services in enlargements countries, Report 4, 2014, pg.22 and 29.

[6] ITU, ICT facts and figures, The world in 2015, Geneva, 2015.

[7] RZS, Upotreba informaciono-komunikacionih tehnologija u Republici Srbiji 2014, Beograd, 2014.

[8] RZS, Upotreba informaciono-komunikacionih tehnologija u Republici Srbiji 2015, Beograd, 2015.

[9] EC, A digital agenda for Europe, com (2010) 245 final, Brussels, 2010.

[10] EC, Digital agenda targets progress report, Digital agenda scoreboard 2015, Brussels, 2015.

[11] WEF and ISEAD (2015), The global information technology report 2015, ICTs for Inclusive Growth, Geneva, 2015.

[12] Bilbao-Osorio, B et al. (2013), The network readiness index 2013: Benchmarking ICT uptake and support for growth and jobs in a hyper-connected world, The global information technology report 2013, WEF, Geneva.

[13] Welsum, D. van, and B. Lanvin, e-Leadership skills – Vision report, Prepared for the European Commission, DG Enterprise and Industry, October, 2012.

[14] Vidas-Bubanja, M., "ICT contribution to the development of some SEE countries in transition", 36th International Convention MIPRO 2013, Opatija, Croatia, May 20-24, 2013, pg. 1513-1518.

[15] http://knoema.com/atlas/topics/Economy, accessed on 17. 3. 2016.

[16] http://ec.europa.eu/eurostat, accessed on 17. 3. 2016.

[17] www.aek.mk, accessed on 23. 3. 2016.

[18] www.knoema.com/atlas, accessed on 17. 3. 2016.

[19] www.monstat.org, accessed on 9. 3. 2016.

Effects and Evaluation of Open Government Data Initiative in Croatia

Tomislav Vračić*, Mladen Varga** and Katarina Ćurko**
* Ministry of Public Administration, Zagreb, Croatia
** Faculty of Economics & Business, University of Zagreb, Zagreb, Croatia
tvracic@uprava.hr, mvarga@efzg.hr, kcurko@efzg.hr

Abstract — The paper is a follow-up to a previous paper where the authors described how users in a small transitional country perceive open government data. It gives an overview of the characteristics of the open data and the current state of a recently started implementation of an open data initiative in Croatia. Since the process involves a series of actions, the following questions need to be answered: what are the effects of an open data initiative and how could these effects be valued? As a starting point in answering these questions we provide an overview of current research initiatives on how to value effects of open government data. Approaches to open data research are usually classified into three broad groups: open data readiness assessments; open data implementation studies; and impact studies. We also bring an overview of the evaluation frameworks already in place in countries where open data initiatives are on the top of the political and social agenda. As a result, in this paper we propose actions in the areas of law and policy, as well as in the areas of organization and technology, which could benefit Croatia's economic growth and democratic empowerment, as suggested by open data advocates.

Keywords — open data, government data, open government data, public sector information, implementation, evaluation framework

I. INTRODUCTION

While accomplishing their institutional tasks government bodies collect, produce, reproduce and disseminate public data in many areas of activity. Government data (GD), also referred to as public sector information (PSI), includes data, such as geographical, statistical, weather, transportation, public health data, etc. that are of public interest, belong to the whole community and every citizen is entitled to know and use them. GD could be opened for re-use for citizens, business and industry, science, media, civil society and others. The benefits of opening GD are numerous, even if it is impossible to predict what value it will create in the future [1].

In [2] we presented a small country user's perspective on opening government data: the position and the early experience with an open data initiative in a small transitional country (the Republic of Croatia with about 4.3 million inhabitants), the objectives of opening government data to the public, end-users' requirements of open data, the impediments to end-user's adoption of open data, and the possible government measures to boost open data usage.

The paper is organized as follows. The characteristics of open data are discussed in Section II, the state of implementation of open data initiative in Croatia is described in Section III, ideas on how to value an open data initiative are considered in Section IV, possible approaches to open data research are presented in Section V, evaluation frameworks proposed and already implemented are described in Section VI and recommendations on further activities in Croatia are given in Section VII.

II. CHARACTERISTICS OF OPEN DATA

Open Data Institute [3] defines open data as information that is available for anyone to use, for any purpose, at no cost. OpenDefinition.org [4] defines open data as data that can be freely used, reused and redistributed by anyone. Private data, i.e., sensitive personal data, are not considered open data.

Open government data are data produced by government or its bodies that can be freely used, reused and redistributed by anyone, resulting in a fairer and better government.

As discussed in the literature [5] government data shall be considered open if it is made public in a way that complies with the principles below:

1. Complete - All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.

2. Primary - Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.

3. Timely - Data is made available as quickly as necessary to preserve the value of the data.

4. Accessible - Data is available to the widest range of users for the widest range of purposes.

5. Machine processable - Data is reasonably structured to allow automated processing.

6. Non-discriminatory - Data is available to anyone, with no requirement of registration.

7. Non-proprietary - Data is available in a format over which no entity has exclusive control.

8. License-free - Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

Based on common forms of practice for using open data that have emerged, a "standard model" of open data has been constituted [6], in which open data initiatives are understood to release pre-existing public sector information resources, and technical skilled intermediaries engage in finding and extracting the social and economic value from that data.

III. State of Implementation of Open Data Initiative in Croatia

With the publishing of open data portal[1] in March 2015 Croatian government joined the majority of EU countries that have a national open data portal [7]. Furthermore, according to the overall score, Croatia was ranked 14th in the Open Data and PSI re-use throughout the EU [8]. Although usage of Internet for e-Government in Croatia is increasing [9], compared to other EU member states the overall position of Croatia is not satisfactory, as Croatia was ranked 27th in the area of Digital Public Services and Open data [10] in 2015. Croatian citizens are aware that greater government transparency and accountability are necessary and because of that it is crucial for Government to introduce additional improvements.

In 2015 Croatia amended the Right to Information Act [11] by adopting Directive 2013/37/EU that amends Directive 2003/98/EC on the re-use of public sector information [12]. By adopting these changes Croatia has set a favorable regulatory framework for the implementation of an open data initiative even if the bylaws on regulating licensing and exclusivity rights still have to be discussed.

In order to describe the other aspects of the current state of implementation of the open data initiative in Croatia we will make an analysis based on our previous work. In the previous paper [2] we presented government measures for boosting the usage of Open data. The measures that were introduced afterwards we will classify as conceptual (1-11) and technical measures (12-15):

(1) Governments often act as secrecy keepers, not openness leaders and the recommendation is that the Government should define a strategy to change its culture.

- Placing government institutions as the leaders of data-driven government initiative has mostly triggered government-related accessibility. Croatian Government implemented the national open data portal, data.gov.hr, in line with similar practices for national portals in the EU and the rest of the World (e.g. data.gov.uk, data.gov.it, data.gv.at, data.gov, etc.). Croatian Government took a leader's role by making open government data one of the priorities of the e-Croatia strategy, drafted for the period 2016-2020.

(2) Government should provide a useful definition and explanation of both public and open data that is accepted by law, administration and citizens.

- Within the preparation of Right to Information Act [11] a kind of consensus among all interested stakeholders was reached.

(3) The difference between open and private data must be clearly defined and explained.

- Recommendations were published [14] in order to define and explain how to make a distinction between open and personal data.

(4) Citizens should be educated that they have the right to seek public data.

- Several workshops have been organized with the support from European Commission Projects [15], [16]. As a result, educational modules developed within these projects [15] can be re-used in public and private education initiatives.

(5) Since the marketing of open data has not been satisfactory, the civil society and government should do their best to market open data.

- Over the last two years, open government data often became a subject of public discussion or events. The initiative from the public sector usually comes from the Ministry of Public Administration[2] and Office of the Information Commissioner[3]. As regards the Civil Society, there are several groups organized within the civil society initiatives, e.g. "Imamo pravo znati"[4] Although there has been no strong marketing campaign, the actions taken by the abovementioned stakeholders have surely helped to promote open government data in the media and among the general public.

(6) Government should open a channel for communicating with users.

- After launching the Open Data Portal, a direct channel for communication with citizens has been established through the contact form on the portal. The citizens are now able to make requests[5] on missing datasets for re-use and in the first ten months thirty-three such requests were made.

(7) Users should be attracted by high value datasets.

- When the Open Data Portal was published in Croatia, it contained 100 datasets (e.g. air pollution data, sea pollution data, list of medicaments, broadband network availability, list of public institutions, registry of political parties and registry of non-government organizations). At the moment, the national Open Data Portal includes 196 datasets that are classified by EuroVoc, which also enables users to access the European open data portal[6]. The authorities also plan to publish datasets according to the G8 Open Data Charter [17].

(8) Non-governmental organizations, charity organizations and business associations have to be involved in open data initiatives.

[1] https://data.gov.hr

[2] https://uprava.gov.hr/
[3] http://www.pristupinfo.hr/
[4] http://imamopravoznati.org/
[5] http://data.gov.hr/data-request
[6] http://www.europeandataportal.eu/

- A platform for cooperation has been established through the implementation documents of the Open Government Partnership in Croatia [13].

(9) Governments should act like product developers and measure the outcome of their activities.

- The outcome of Open data initiative activities in Croatia has been measured through ePSI platform [8]. At the moment, there are no specific national measurements and this will be discussed in the following chapters.

(10) Government should define and explain to users what open data as opposed to private data are.

- The differences between open data and personal data have been explained in documents published by the Ministry of Public Administration [14].

(11) Data openness should be governed through an independent body.

- The office of the Information Commissioner is an independent body that answers to the Parliament. It is in charge of settling disputes if a public administration body refuses to open their data.

(12) Government should ensure the quality of open data.

- Datasets published at the Open Data Portal have been checked before publishing against data formats, licenses, metadata, broken links etc. These actions are helpful for achieving higher quality in process of publishing.

(13) Web portal should be equipped with as many data applications as realistically possible so that a proper user-friendly interface would be a catalyst of using open data.

- The implementation of the Open Data Portal in Croatia was based on the free software created by the UK Government [18]. It has two main sections: datasets for re-use and applications made on open data with 19 applications published.

(14) Open data portal should use the accepted standards.

- The Open Data Portal in Croatia includes a documents section with technical and user's documentation as well as handbooks for the data publishers [14], which is based on [25]. The recommended formats for reports are HTML, TXT or XML while SV, TSV, JSON, XML or RDF are the recommended data formats.

(15) Datasets should be certified.

- Dataset certification published in the Open Data Portal is aligned with the five stars classification scheme [19].

IV. How to value an open data initiative?

The majority of authors agree that open government data movement arose from the Open Knowledge movement [4] offering a point for consolidation for many distinct groups, all seeking some form of greater access to government datasets. These groups, with diverse interests, have come to constitute a global movement for open data [20], [21]. Although open data movement has been present globally for a number of years, no rigorous studies have attributed large-scale impacts to open data.

In [6] the authors gave examples of a number of common forms of practice for using open data. These range from hack days to "data journalism". The authors also offered the "standard model" of open data, in which open data initiatives release pre-existing public sector information resources while technically skilled intermediaries engage in finding and extracting the social and economic value from the said data.

If the "standard model" of open data is used to analyze the situation in Croatia, one notices that all elements are already in place. For example, there are civil society open data initiatives that organize hack days[7], regularly supported by government organizations. On the other hand, the Open Data Portal offers a unique place for technically skilled intermediaries to find datasets and create values from them. Some applications published at the Open Data Portal have already achieved commercial value[8].

Although commercial applications are exceptions and not a rule, when could we expect the visible results? The analysis [22], which is based on a number of open data implementations in the European Union, shows that the real starting point is reached when public service bodies (PSBs) dramatically lower the charges for the re-use of their public service information (PSI), when they make it free or almost free. This results in a whole array of subsequent effects, which can be shown in three phases: (a) the sowing, (b) the growing and (c) the harvesting phase.

In the sowing phase, in which Croatia is about to enter, following the price cut, the increase in demand in terms of volume and number of users is often by factor 2-100. Lowered prices attract new types of users, but the income of PSBs is very low because they lose revenues and the costs increase.

In the growing phase the market becomes more dynamic: new parties enter the market, existing users innovate and upgrade their services. The high-end market usually consists of open data re-users that provide services to professional clients. The business model in the low-end market is usually connected to integrating PSI into mass services oriented to non-professional consumers, possibly via apps on mobile devices.

In the harvesting phase the investments in PSI made by PSBs outweigh the cost, mostly through increased tax returns. The market dynamism leads to economic growth and produces various economic and non-economic benefits.

[7] https://codeforcroatia.org/
[8] http://data.gov.hr/apps/invest-croatia

The report [22] concludes "(a) that a strict marginal cost regime of PSI without exemptions has to be seen as an investment, (b) that this investment does not only pay off in terms of taxation gains but also in efficiency effects in the public administration, and (c) that it takes time to start the harvesting phase and in this timespan the relevant PSB will suffer losses in their revenue".

Having in mind the establishment of the "standard model" of open data in Croatia, as well as the recognized starting point in the sowing phase, there is a need to research the real value of open data initiative. In the next chapter we will present an overview of different approaches to open data research

V. Approaches to Open Data Research

According to [6], general approaches to open data research, could be classified into three broad groups: open data readiness assessments; open data implementation studies; and impact studies, as shown in Figure 1.

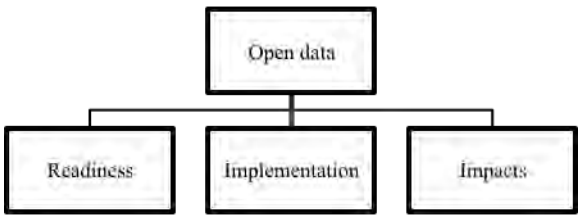


Figure 1 Approaches to open data research

The first group of studies, *readiness*, researches whether the conditions in a country, city or sector might be appropriate for an open data initiative to be effective, and may also highlight areas where investments or effort would be needed to get ready for an open data initiative. Readiness assessment is usually done through readiness studies that generally need to establish what an open data initiative would consist of, and to decide which factors are important for making an open data initiative a success.

Placing government institutions as the leaders of data-driven government initiative has for the most part proved *readiness* in Croatia. Firstly, in November 2012, Croatian Government asked its ministries [26] what datasets were prepared online for re-use. As a result, a list of seventy-eight datasets, applications and other forms of information resources in the last phase of preparation re-use was created. An online survey followed in February and March of 2013 asking members of the Croatian Software Exporters Union about their interest in public sector information for commercial reuse. Seventy-seven percent of respondents confirmed they were already using public sector information. Survey results revealed that there is interest in datasets in the following areas: economics, justice, demography, culture, agriculture, meteorology, telecommunication, media and finance. In 2014 additional research was made on the forms of public information already published on the official public administration institutions websites. The results showed 398 datasets and application for re-use. At that time, information was being proactively published in the area of the judiciary, environment protection, agriculture and fishery, geographical area, meteorological area, public procurement, statistics and culture.

The second group of studies, *implementation*, researches whether the conditions for open data, or open data itself, actually exist in a country, city or sector: that is, whether open data policies are being implemented. Implementation studies that are produced within this group have to determine what good open data looks like, which datasets are important to survey, and how to measure the level of implementation.

Implementation in Croatia has been studied within the PSI Scoreboard that represents a "crowdsourced" tool used to measure the status of Open Data and PSI re-use throughout the EU. It does not monitor government policies, but aims to assess the overall PSI re-use situation, which includes the open data community's activities.

The third group of studies, *impacts*, researches whether open data has led to change. Generally, they focus on whether open data has led to economic growth or democratic empowerment, i.e. the specific benefits suggested by open data advocates.

When it comes to researching economic growth, studies that analyzed the potential market value [23] and economic benefits of open data initiatives [24] estimated an additional increase in GDP level in Croatia by approximately 2 percent in the period 2013-2020. It is expected that analyses of this kind should create a great challenge for researchers in the future.

VI. Evaluation Frameworks Overview

According to [27], the evaluation of open data initiatives has become an increasingly pressing concern.

Many global organizations are helping countries develop open data initiatives by providing methodological frameworks or comparable indicators. For example, The World Bank's Open Government Data Working Group developed an Open Data Readiness Assessment (ODRA) methodological tool that can be used to conduct an action-oriented assessment of the readiness of a government or individual agency to evaluate, design and implement an Open Data initiative [31]. As part of the Open Government Data Toolkit, this tool is freely available for others to adapt and use. They have also provided a list of available assessments[9] done in mainly less developed regions or countries, such as: Ulyanovsk - Russian Federation, Antigua & Barbuda, Peru, Mexico, Dominican Republic, Burkina Faso and Serbia.

Furthermore, Web Foundation, as part of the Open Data for Development group, started the Open Data Research Network that is currently focused on the Open Data in Developing Countries project (ODDC). In the first phase of this project, over a hundred researchers from the Global

[9] http://opendatatoolkit.worldbank.org/en/odra.html#available-assessments

South developed 17 qualitative case studies[10] with findings that spanned 13 countries, from Indonesia to Brazil. Generally, six key areas that each case study should address were highlighted [6]:

- The context of open data
- The supply of open data
- Technical platforms and standards
- The context of the specific governance setting
- Intermediaries and data flow
- Actions and impacts

The Exploring the Emerging Impacts of Open Data in Developing Countries (ODDC) project[11] is working to develop a shared toolkit of research methods that can be used in understanding the nature, use and emerging impacts of open data in a range of different country and governance contexts around the world. Examples of governance and evaluation frameworks are:

- COBIT framework[12] – is a business framework created by ISACA for the governance and management of enterprise IT. The current version, COBIT 5 is based on five key principles:

  1. Meeting Stakeholders Needs
  2. Covering the Enterprise End-to-End
  3. Applying a Single Integrated Framework
  4. Enabling a Holistic Approach
  5. Separating Governance From Management

  In Ireland [28], an example was given by using COBIT framework to explain the separation of governance from management of the Open Data Ireland initiative.

- The Logical Framework Approach – is an analytical process and set of tools used to support objectives-oriented project planning and management. It provides a set of interlocking concepts that are used as part of an iterative process to aid structured and systematic analysis of a project or program idea. For the purpose of evaluation of the open data initiative, this approach could be used in the phases of project identification and formulation (the analysis stage, the planning stage).

- Most Significant Change – is a form of participatory monitoring and evaluation. It is a form of monitoring because it occurs throughout the program cycle and provides information to help people manage the program. For the evaluation of an open data initiative this approach could be used to compare best practices and to find changes based on the usage of open data that had the most significant impact.

On the basis of the examples shown above we can conclude that there is no unified expert opinion on how to evaluate and research open data initiative impacts. Therefore, the global initiatives that support countries in open data development are creating an excellent basis for future research activities.

VII. Recommendations on further activities in Croatia

The majority of the published datasets on the Open Data Portal in Croatia are governmental data. Still, there are a lot of data from other sources that could be opened. Example is business data collected by the chambers of commerce or trade.

The World Bank offered a "menu" [29] of services to promote and support "Open Data Literacy", the goal of which is to catalyze, engage, and inspire strategic multi-stakeholder groups to see the value and potential of open data, and what it means for local, national, and regional development in a practical, hands-on way. This could be a helpful tool for Croatia to stimulate the demand-side of open data and increase the open data *readiness* of all stakeholders.

Additional activities should be taken in the area of open data *implementation*. Besides, participating in analyses that compare results in different EU countries, Croatia should also participate in global initiatives, such as Open Data Barometer[13], that aims to uncover the true prevalence and impact of open data initiatives around the world or the Global Open Data Index[14], an annual effort to measure the state of open government data around the world.

When it comes to the open data *impact* area, additional initiatives that include dialog and exchange of experiences with the local communities should be taken in the future. Having in mind the financial limitations of the majority of local communities in Croatia financing future open data initiatives will be challenging. One possible solution is to obtain financing through different financial instruments (EU funds, EU programs), which is an underway process in Croatia. Another solution is to find resources through public-private-partnerships.

To clarify the objectives of open data in Croatia, the Government of the Republic of Croatia should further define short, medium and long-term objectives. This could be a helpful tool in setting a perspective for future research studies.

VIII. Conclusion

In this work we showed that the goal of sustainable data-driven government has lately been put on the agenda of the Croatian Government. And the long-term sustainability of the open data initiative has also been recognized by

[10] http://opendataresearch.org/project/2013/oddc
[11] www.opendataresearch.org/
[12] http://www.isaca.org/Cobit/pages/default.aspx
[13] http://www.opendatabarometer.org/
[14] http://index.okfn.org/

promoting participation in the exchange of experience and ideas around implementing open data policies in the public sector [30]. But, to lead the open data initiative in Croatia to a self-sustaining level it is crucial to find the right balance between the expectations of open data users and open data publishers. Because of that, one of the upcoming challenges in the open data in Croatia surely will be a work on appropriate legal framework for charging and protecting intellectual property rights.

Based on these conclusions we proposed actions in the law, policy, organizational and technical area that could bring benefits from opening datasets in Croatia – such as economic growth or democratic empowerment. We are aware that one of the future research challenges will be to understand effects of these actions on readiness, implementation and impact. Having this in mind and since finding adequate evaluation frameworks is on the top of the agenda of researchers in many areas, we also expect to continue our work in this area.

REFERENCES

[1] Open Knowledge Foundation, "Open Data Handbook Documentation, release 1.0.0", 2012

[2] Varga, M., Ćurko, K., Vračić, T. "Open Government Data: Small Country User's Perspective". Proc. of the Eighth International Conference on Digital Society ICDS 2014, 33-38, Barcelona, 2014

[3] Open Data Institute "Guides". Retrieved February 3, 2016, from http://theodi.org/guides/what-open-data

[4] Open Knowledge Foundation. "Open Knowledge Definition". Retrieved February 5, 2016 from http://opendefinition.org/

[5] "Open Government Data Principles". Retrieved February 3, 2016 from https://public.resource.org/8_principles.html

[6] Davies, T., Perini, F., Alonso, J.M., "Researching the emerging impact of open data", ODDC conceptual framework, World Wide Web Foundation, July 2013

[7] European Data Portal Project, "Insight Report n1", European Union, 2016

[8] ePSI Platform, "The PSI Scoreboard". Retrieved February 5, 2016 from http://www.epsiplatform.eu/content/european-psi-scoreboard

[9] Croatian Bureau of Statistics, "Usage of Information and Communication Technologies (ICT) in households and by individuals, 2015, first results". Retrieved February 5, 2016 from http://www.dzs.hr/Hrv_Eng/publication/2015/02-03-02_01_2015.htm

[10] European Commission, "The Digital Economy and Society Index". Retrieved February 5, 2016 from https://ec.europa.eu/digital-agenda/en/desi

[11] "Right to Information Act", Official Gazette, No. 85/2015. In Croatian. Retrieved February 5, 2016 from http://narodne-novine.nn.hr/clanci/sluzbeni/2015_08_85_1649.html

[12] "Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information," Official Journal of the European Union L 175/1 27.6.2013. Retrieved February 5, 2016 from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:EN:PDF

[13] Government of the Republic of Croatia, "Action plan for the implementation of the Open Government Partnership in Croatia for the period 2014-2016". In Croatian. Retrieved February 5, 2016 from https://vlada.gov.hr/UserDocsImages//Materijali%20za%20istaknuto/2014/Otvorena%20vlada//Akcijski%20plan-POV-8-7-2014-final.pdf

[14] Ministry of Public Administration, "Recommendation for publishing dataset for re-use". Retrieved February 5, 2016 from http://data.gov.hr/sites/default/files/library/Preporukezaobjavu.pdf

[15] European Commission, "Open Data Support". Retrieved February 5, 2016 from https://joinup.ec.europa.eu/community/ods/description

[16] European Commission, "ePSI Platform". Retrieved February 5, 2016 from http://www.epsiplatform.eu/content/about-us

[17] UK Cabinet Office, "Open Data Charter", 18 June 2013. Retrieved February 5, 2016 from http://www.gov.uk/government/publications/open-data-charter

[18] "Data.gov.uk To Go". Retrieved February 5, 2016 from http://data.gov.uk/blog/datagovuk-to-go

[19] Berners-Lee, T., "Linked Data". Retrieved February 5, 2016 from http://www.w3.org/DesignIssues/LinkedData.html

[20] Davies, T., & Bawa, Z. A. "The Promises and Perils of Open Government Data". The Journal of Community Informatics, 8(2). Retrieved February 5, 2016 from http://ci-journal.net/index.php/ciej/article/view/929/926

[21] Davies, T., & Bawa, Z. A. "The Promises and Perils of Open Government Data", An updated version, The Journal of Community Informatics, 8(2). Retrieved February 5, 2016 from http://ci-journal.net/ index.php/ciej/article/view/929/955

[22] De Vries, M., Hittmair, G. "Open data is fact now – When does the reuse start?". Retrieved February 5, 2016 from https://www.w3.org/2013/share-psi/wiki/images/a/a3/Lisbon_ deVries_hittmair.pdf

[23] G. Vickery, "Review of recent studies on PSI re-use and related market developments," Information Economics Paris. Retrieved February 5, 2016 from http://ec.europa.eu/information_society/newsroom/cf//document.cfm?doc_id=1093

[24] Buchholtz, S., Bukowski, M., Siegocki, A., "Big and open data in Europe, A growth engine or a missed opportunity?". demosEUROPA – Centre for European Strategy Foundation, Warsaw Institute for Economic Studies, 2014

[25] European Commission, "Guidelines on recommended standard licenses, datasets and charging for the re-use of documents". Official Journal of the European Union, C 240/1 24.7.2014. Retrieved February 5, 2016 from http://ec.europa.eu/newsroom/dae/document.cfm?action=display&doc_id=6421

[26] Government of the Republic of Croatia, "Guidelines for making internet site," In Croatian only. Retrieved February 5, 2016 from https://vlada.gov.hr/UserDocsImages//Materijali%20za%20istaknuto//Gov.hr%20-%20Smjernice%20za%20izradu%20internet%20sjediste.pdf

[27] Davies, T., "Notes on open government data evaluation and assessment frameworks". Retrieved February 5, 2016 from http://www.opendataimpacts.net/2013/02/506/

[28] Lee, D., "Open Data Ireland: Evaluation Framework", Insight Centre for Data Analytics, NUI Galway, May 2014.

[29] The World Bank, "Open Government Data Toolkit". Retrieved February 5, 2016 from http://data.worldbank.org/about/open-government-data-toolkit/demand-for-od-engagement-tools

[30] Share PSI 2.0. Retrieved February 5, 2016 from http://www.w3.org/2013/share-psi/

[31] The World Bank, Open Government Data Working Group, "Open Data Readiness Assessment (ODRA) Methodology". Retrieved February 5, 2016 from http://opendatatoolkit.worldbank.org/docs/odra/odra_v3.1_methodology-en.pdf

ICT technologies and structured dialogue: experience of "Go, go, NGO!" project

Nikola Kadoić*
* Faculty of organization and informatics, Pavlinska 2, Varaždin
nkadoic@foi.hr

Abstract - Structured dialogue can be defined as communication between the decision makers – government units (local, regional, national, European) – and citizens who act personally or through different organizations, such as institutions, enterprises, associations (non-governmental organizations, NGOs) or groups. The main focus of this paper is determine importance of ICT in that process. Paper contain presentation of a methodology which will help increase the level of structured dialogue between local government units (LGUs) and associations (NGOs), especially when it comes to NGOs dealing with the younger population. The suggested methodology has been applied in the project "Go, go, NGO!" and the resulting ideas should help increase the structured dialogue level in 5 local government units. Some of those ideas include the application of ICT technologies. Project included a city and 4 municipalities. All have a development index below the average of the Republic of Croatia and are a part of a county with development index below 75% of Croatian average. Project results show improvement of the structured dialogue, but also the significant role that ICT technologies played in that process. Naturally, we cannot expect the improved structured dialogue to have an immediate impact on the development index, but it can have long-term effects.

I. INTRODUCTION

In the scope of this paper, the term *structured dialogue* refers to the dialogue with the youth through government units. The structured dialogue is defined as a process through which public bodies (on local, regional, national and European level) ask the citizens and consult with them on various important topics [1]. This dialogue encompasses not only the youth and decision makers who discuss certain topics but also different institutions, organizations, associations, groups, experts, and individuals. According to a report of the European Youth Forum, the highest non-governmental youth body in Europe, the structured dialogue in Europe is defined through: (1) National Working Groups consisting of representatives from National Youth Council(s), the ministry in charge of youth affairs; (2) national agencies that lead Youth in Action program and conduct consultations with the young people and policy makers at national, and, whenever possible, local and regional levels; and (3) European Steering Committee, consisting of representatives from European Commission, the Trio presidency and the European Youth Forum that compiles the reports, including inputs from national working groups, international non-governmental youth organizations and other international partners, into one or more background documents intended for an EU Youth Conference [2].

Although these bodies define how structured dialogue should be carried out (strategic view), problems arise at a local (operational) level where the structured dialogue is implemented through youth councils, local governments units, non-governmental organizations, and institutions. Implementation of the structured dialogue in local government units is often hampered by the fact that there are no youth experts who could direct the decision makers in the right path of action. One research on a local level found that the decision makers do not even know what a structured dialogue is [3]. Eurodesk and Agency for Mobility and EU Programs conducted a survey on how the structured dialogue is put into practice in Croatia [4]. Because the survey included various groups (not only on a local level), the results were somewhat better.

The project "Go, go, NGO!" aimed to improve the structured dialogue mainly on a local level – in small cities and municipalities – and was carried out in 5 local government units. Their development index is below Croatian average in 2013 [5]. LGUs have very low amount of funds set aside for the structured dialogue. First present state analysis showed:

- All LGUs have low budgets; they do not have an employee dealing with the structured dialogue. In most cases, they do not employ young people nor people with expertise in managing a structured dialogue.
- LGUs have no youth councils established, no local programs defined, youth policies or any other contents recommended by the European Commission (mainly because they are not aware of them).
- The youth support in LGUs for the most reflects through (low) scholarships and school transport co-financing, but that system is not well organized and offers no support to the young people who leave LGUs after graduating. Most of the young people relocate to bigger cities.

The young people who took part in the project and live in the participating LGU also did not know the meaning of the structured dialogue. They can be categorized into a single group – youth with fewer opportunities who face geographical obstacles (sometimes even economic obstacles) [6].

LGUs do support various social initiatives put in motion by institutions, non-governmental organizations, schools, groups of (young) people and individuals, but the support system is not always transparent nor clear.

The main project goal was to increase the structured dialogue level, but there were also some subgoals: (1) building a transparent support system for NGO sector; (2) establishing youth councils; and (3) stimulating youth participation in NGOs. To achieve those goals, we created a methodology designed to increase the structured dialogue level by using the Four-Phase PDCA (Plan-Do-Check-Act) Model (Deming's Cycle) in combination with the BSC (Balanced Scorecard) [7].

II. STRUCTURED DIALOGUE AND USING ICT TO ESTABLISH AND MAINTAIN A STRUCTURED DIALOGUE

Literature review showed that there is no a lot of examples of organized conducting of structured dialogue and using ICT in that process, especially in small cities and municipalities in Croatia. Before describing the methodology, we will mention a brief analysis of papers dealing with the structured dialogue and youth.

- Authors of the paper [8] explored the possibilities of a structured dialogue between Turkish and Greek communities in Cyprus. Using the Structured Dialogic Design Process, they proposed 27 options (for establishing dialogue) and developed an influence map.
- A structured dialogue can also be achieved through volunteering on different projects in local governments. Survey [9] explored the role of local government agencies in attracting and managing volunteers.
- Youth councils, for the most part funded by adults, can significantly contribute to local communities. The paper [10] describes the successes and perceived challenges of youth councils.
- NGOs and LGUs can efficiently collaborate and successfully deal with employment problems. NGOs play an active role in the implementation of the ALMPs in Finland and Sweden by (1) employing the long-term unemployed; (2) providing social services, mobilizing local resources and undertaking other types of activities that create jobs for the unemployed; and (3) providing voluntary work to volunteers – some of whom acquire skills that can help find a job in the open labor market [11].
- Engbers investigated the characteristics of the most civic cities in the US and concluded that institutional factors unite cities with the highest levels of participation. These include a strong corporate presence, mobilization mechanisms, strong community identity, public spaces, good government and investment in youth. In our case, the above mentioned factors were only partially present [12].

Many LGUs use various ICT solutions that help them in their everyday work. In the case of LGUs using ICT solutions to establish and maintain a structured dialogue, we can say that there is a lot more potential there waiting to be explored and implemented.

- Paper [13] describes various classifications, possibilities and uses of ICT in local government administration. In the structure dialogue process, we can identify ICT solutions and their advantages and disadvantages in relation to their benefits, costs, risks, and impacts.
- Using ICT, LGUs can create a culture of transparency [14].
- Planas, Soler, and Vilà proposed an assessment tool – System of Assessment Indicators for Local Government Youth Policies (http://siapjove.udg.edu/), which provides both quantitative and qualitative indicators through which youth policy managers, with relative ease, can obtain assessment reports in 12 possible youth policy areas of assessment [15].
- The emergence of social media, mobile technologies, Web 2.0 and the connected government do not play a truly significant role in the quest for e-government individually, but only in combination with other factors as discussed in [16].
- In the paper [17] Agostino analyzed 119 Italian municipalities and examined in what way do social media stimulate public engagement. YouTube is used to support public communication and Facebook to support public participation.
- The Civitas Initiative provides a set of guidelines addressed to cities that want to start building their own social media communication strategy, but could also be useful in cities that already have a social media strategy in place and would like to improve it [18].

III. METHODOLOGY FOR INCREASING THE STRUCTURED DIALOGUE LEVEL

The project methodology was created with the goal of increasing the structured dialogue level through the project "Go, go, NGO!" and was based on the Deming cycle [19]:

- PLAN: designing or revising business process components in order to improve results
- DO: implementing the plan and measuring its performance
- CHECK: assessing the measurements and reporting the results to the decision makers
- ACT: deciding which changes are needed to improve the process

TABLE I. PHASES OF THE PROPOSED METHODOLOGY FOR INCREASING THE STRUCTURED DIALOGUE LEVEL

PDCA	Goals	Methods/techniques	Outputs
Plan	Formulate the central problem Problem analysis Define goals Create potential solutions	Present state analysis Problem tree SWOT Forming strategies Brainstorming Case studies	Concrete activities that could help improve the structured dialogue
Do	Describe all defined activities	Resource allocation analysis Cost-benefit analysis Scenarios analysis Debate, discussion In-depth analysis	Detailed description of each activity implementation (costs/resources needed and benefits expected)
Check	Present analysis of all activities	Presenting analyses; Case studies	Analysis of costs, resources and benefits in all activities
Act	Select activities that will be applied	Structured dialogue: discussion (Evaluation)	BSC strategic map of goals
Plan	Implement selected activities	Creating an action plan (defining deadlines, responsibilities, budget) for each activity	Implementation plan BSC strategic map of measures
Do	Implement all activities	Implementation Creating checkpoints	Implementation results (measure values achieved following implementation)
Check	Present implementation results	Case studies	Activity implementation analysis
Act	Evaluate implementation results	Qualitative and quantitative analysis	Determing the reached structured dialogue level

In our approach, we applied two Deming cycles: first on a theoretical level (without any application in practice) and then on a practical level. Of course, the Deming cycle implies a continuous process, and when the second cycle is completed, a new one can be applied for additional improvement. Our approach uses the not-for-profit BSC as a performance management system [20], [21].

The methodology is presented in Table 1. In the beginning, we had to define a central problem, as well as other related problems. The central problem referred to the existing structured dialogue level, and the related problems were its sources (causes) or consequences (e.g. an LGU has no established youth councils; there is no support for NGOs; citizens do not understand many of the decisions made by the LGU's decision makers, etc.). The following methods and techniques can be used in defining the problems: a problem tree, case study analysis and present state analysis. Also in that phase, possible solutions need to be suggested. In the BSC, that means that we had to set strategic goals we wanted to achieve. After setting the strategic goals, we had to do a SWOT analysis for each goal and pair the identified SWOT elements to create strategies and alternatives, i.e. activities that will help increase the structured dialogue level. Other methods that can help create such activities include brainstorming and case study analysis – good practices from other local government units. As the example of one of the five LGUs participating in the project will show, some of the identified activities are connected with the implementation of ICT technologies, social networks, and web 2.0 tools.

In the DO phase, all generated activities need to be described. We needed to identify (1) the benefits of activities in relation to the post-implementation structured dialogue level and inputted (2) cost and (3) resources. To identify these elements, we used a cost-benefit and resource allocation analysis or scenario analysis (trying to predict the process of implementing each activity, as well as possible pitfalls, reactions, and results). In the CHECK phase, we presented the identified activities and cost-benefit-resource results to the decision makers. In the ACT phase, the decision makers had to select activities that would go into the implementation phase, keeping in mind their individual but also their combined effects. It is recommended to complete both the CHECK and ACT phase on the same day/at the same meeting. Personal attendance of the decision makers (not their delegates, i.e. LGU's employees) is also recommended because these are crucial moments in making the most important decisions. If possible, an additional group (i.e. people who did not generate ideas) should also take part in the process and evaluate the selected activities.

As shown in Table 1, the first PDCA cycle is purely theoretical, without any implementation. The prescribed methodology here entails dialogue steps that both the decision makers and young people (NGOs) have to take. We can say that *to increase the structured dialogue level, we have to use a methodology based on the structure dialogue mechanism*. The second PDCA cycle takes place on a more practical level.

In the PLAN (2) phase, we needed to create implementation plans (action plans) for all activities selected in the ACT (1) phase. The result was a BSC strategic map of goals. Deadlines, responsibilities, resources and other important elements had to be defined for each goal in the map. (It was at this stage of our project that the heads of the LGUs signed documents containing that information.) During the implementation phase, i.e. the DO phase, selected activities were executed and monitored. We also had to create a BSC strategic map of measures. For each measure, we defined four target values: U, u, 1 and L [22]. Values between U and u suggest *good goal achievement*, values between u and 1 show *satisfactory goal achievement* and values between 1 and L indicate *poor goal achievement*. The person in charge of the activity oversaw its implementation. The

CHECK phase starts when the last action plan activity finishes. If there are many activities, it is recommended to define a checkpoint before the final check. In this phase of our project, the analysis of each activity implementation was done – we had to check whether the BSC strategic map goals had been achieved. We also drew conclusions on the measured values in the map of measures and goal efficacy (achievements).

In the case of more complex problems, software use is recommended. In our case, calculations were made in Excel. However, Dialog strategy is a more appropriate software because it enables dynamic monitoring of goal achievements through scorecard graphs [22].

The proposed methodology was tested during the course of the project "Go, go, NGO!" project and gave very good results. More methods per methodology phase can be counterproductive in the application domain of both LGUs and the young people.

IV. CASE STUDY: PROJECT "GO, GO, NGO!"

The following part of the paper presents the application of the methodology in the case study of the project "Go, go, NGO!". As mentioned earlier, project covered five LGUs, but this presentation will deal with the results (DO (2)) of only one of them. Similar results were achieved in the other four LGUs.

The project was structured in the form of six three-day meetings. At each meeting some methodology phase(s) was (were) carried out. Approximately 50 participants were present at the meetings and assignments were executed in groups. The decision makers did not attend all meetings – other LGU's employees took their place at some. After completing each assignment, groups exchanged their results.

At the first meeting, the following activities were implemented: theoretical presentation on the structured dialogue; presentation on results of an income survey previously filled out by project participants (both the young NGO representatives and decision makers from LGUs); presentations on good practices in relation to the structured dialogue and case studies; brainstorming on possible activities that could be implemented in the LGU to increase the structured dialogue level; learning about the *problem tree* method: theory and examples; making a problem tree for three problems: *low structured dialogue level*, *weak support of NGOs by LGUs* and *weak motivation of the young people for active involvement in the community social life*; brainstorming on how to influence problem causes in the created problem trees; learning about *SWOT* in theory and practice; making SWOT analysis for two strategic goals: *increasing the LGUs' support of NGOs* and *making LGUs totally transparent and motivating the young people to be active participants in the community social life*; making strategies (grouping SWOT elements and creating logical activities); presenting examples of a good structured dialogue from a partner institution. The resulting activities stemming from the previously mentioned ones were candidates for implementation in the LGU.

At the second meeting, each activity was analyzed and described in detail. The goal of the meeting was to create an in-depth analysis of each idea. Participants completed resource analysis and cost-benefit analysis for each idea. They also carried out a scenario analysis – they tried to implement ideas on a theoretical, debate level and then to identify possible problems during the actual implementation. Young participants offered arguments and promoted benefits for each idea they wanted to implement. LGUs also weighed in with their perspectives, opinions, and experiences. Both evaluated each idea on a scale of 1 – 10, where 1 and 10 denoted low and high contribution to the structured dialogue level in the LGU, respectively. (Evaluations are depicted in Table 2, columns 2 and 3.) Then, the structured dialogue was presented at a meta-level – activities that would help increase the structured dialogue level in the LGU were decided via structured dialogue mechanism (discussions, debates, case studies, personal reflections). Logically, one generated idea was to promote this project and foster similar ones.

At the third meeting, all of the results were presented to the decision makers and they had to decide which will be selected for implementation, bearing in mind costs and resources needed, but also the benefits that will each LGU reap through implementing each action. The young people presented activities and their arguments. In our case, not all of the proposed actions were accepted because, even though some of them do not require a lot of resources individually, their joint implementation was simply not possible.

LGU that is the case of our study selected the following activities: (1) organization of panel discussions and public debates dealing with the functioning of the LGU or the upcoming decisions; (2) influencing youth activities implemented by the county because until now they were for the most part located in other LGUs in the county; (3) introduction of an *LGU open day*; (4) supporting the state change of the law dealing with the establishment of youth councils (under the current law, youth councils members are primarily selected by LGU's council members, not by the young people whom they should represent); (5) activities directed toward educating elementary and secondary school students on LGUs and youth councils; (6) electing the children's mayor; (7) inviting applications for the LGU's youth council; (8) supporting NGOs by enabling their meetings and other appropriate indoor activities that are under the LGU's jurisdiction. Finally, the following activities included ICT solutions:

- (9) Opening a Facebook profile and maintaining continuous communication with citizens. Considering that many of them, especially the young people, use Facebook on a daily basis, this should ensure better communication and a direct dialogue.
- (10) Recording LGU's council sessions. Every citizen who wants to know how certain council session evolved, could contact the LGU and get an audio copy of the session;

- (11) Head of the LGU makes a guest appearance every month on the local radio. However, due to the broadcast time, not many citizens can actually hear the show. The idea is to, in agreement with the local radio, record the show, upload it to a hosting service (such as YouTube) and make it available to a larger audience at any time and place.

Also, one of the selected activities was to (12) support NGOs via public tender (competition) for the allocation of funds for NGO projects. The current practice is that in most cases it is the mayor who approves or denies funds following individual NGO's requests; the process is not always fair nor transparent. Three items could play a role in the implementation of ICT solutions:

- Internet articles and blogs dealing with the NGO's activity will be accepted as proof of the applied NGO's attainment (for now, only paper references are accepted).
- All information about the tender (competition) will be available online, as well as all the reports on how the money is spent (the list of NGOs that applied, got funding and the amount of funding).
- In future activities, an appropriate system will be implemented to help organize tender documentation, applications, results, etc. So far, Moodle seem to be a successful solution, because it enables user registration (NGO's members and experts who can evaluate project proposals); discussions; uploading documents by the LGU; uploading applications by the NGOs; evaluating applications by evaluators using transparent criteria that can easily be implemented via rubrics option in Moodle; publishing announcements, results and other relevant information; creating surveys, etc.

The listed activities were grouped with the BSC perspective and BSC strategic map of goals shown in Figure 1.

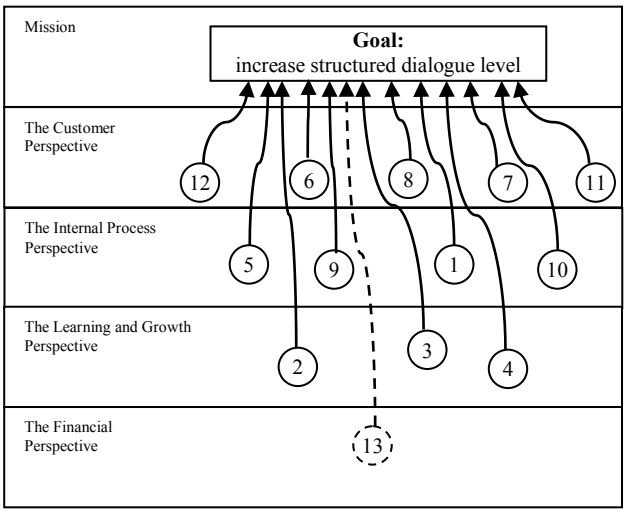


Figure 1. The BSC strategic map of goals (map includes goal 13: writing new project proposals, similar to "Go, go, NGO!")

At the fourth meeting, as recommended in the methodology, additional evaluation of the selected activities (from the previous meeting) was done. Representatives from an LGU that did not participate in the project gave their opinions and recommendations on the selected ideas. That information proved valuable for the following phase and meetings (creation of an action plan). Besides that, project participants learned about new case studies on implementing structured dialogue.

At the fifth meeting, a BSC strategic map of measures was defined: people in charge, deadlines, resources needed and implementation description. That data was included in the *Decision about implementations of selected project activities*, signed by the mayor and presented to every NGO in the LGU.

The time period between the fifth and sixth meeting was reserved for the implementation of every selected activity (DO (2)). After an activity had been implemented, the structured dialogue level was recalculated (based on the BSC strategic map of measures). Table 2 shows the implementation of the BSC strategic map of measures in MS Excel, where the structured dialogue level was calculated. The possible contribution to the structured dialogue for each activity was evaluated at the second meeting. Also, the average contribution of the NGO's and LGU's grade was calculated (column 4)). Those values became weights in measuring the structured dialogue level. During the structure level measuring, we had to evaluate the completeness of each activity implementation. We used a 0-10 scale for that – 0 meant that certain activity had not been implemented, and 10 meant that the activity had been implemented in full. Then, we calculated the structured dialogue level for each activity by multiplying the weight (the possible contribution of each activity in improving the structured dialogue level) and the grade describing the implementation completeness. Columns 5 and 6 contain data on the structured dialogue level before the project had started in the selected LGU. Columns 7 and 8 contain data showing what the structured dialogue level would be if all selected activities were fully implemented.

TABLE II. IMPLEMENTATION OF BSC STRATEGIC MAP OF MEASURES

A. No	Contribution to structured dialogue			Implementation of activities					
				Before project		Maximum		Achieved	
	NGO	LGU	AVG	Grade	Result	Grade	Result	Grade	Result
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	9	6	7.5	2	15	10	75	3	22.5
2	5	2	3.5	0	0	10	35	10	35
3	5	4	4.5	0	0	10	45	0	0
4	2	1	1.5	2	3	10	15	3	4.5
5	9	7	8	0	0	10	80	3	24
6	10	8	9	0	0	10	90	5	45
7	5	4	4.5	5	22.5	10	45	3	13.5
8	8	5	6.5	8	52	10	65	10	65
9	7	5	6	0	0	10	60	8	48
10	7	5	6	0	0	10	60	10	60
11	10	10	10	0	0	10	100	0	0
12	10	10	10	0	0	10	100	5	50
TOTAL					92.5		770		367.5

The sixth meeting is planned at a later date than the publication of this paper. According to plans, at that meeting, the final evaluation of implementation completeness will be done, phases CHECK (2) and ACT (2). In Table 2, last two columns contain data on the current state of each activity implementation.

If we observe only activities 8-12, dealing with the use of ICT solutions, we can conclude that their implementation can contribute a maximum of 320 points to the structured dialogue level, i.e. 41.55% of total structured dialogue level points. This only shows how big of a role ICT plays in the structured dialogue level.

V. CONLUSION

The goal of this project was to increase the structured dialogue level in small government units (LGUs): municipalities and small cities. For this cause, we proposed a methodology and applied/tested it during the course of the "Go, go, NGO!" project.

In order to increase communication with NGO sector and citizens, LGUs opened Facebook profiles. That enabled two-way communication and better visibility of information that were originally published only on web page. 100% of young participants from project find this way of communication more practical for them because so far they didn't have a habit visiting web page of LGU, and now, when information are linked to the Facebook, they got information on social network which they use on daily basis. Even though some of the information that are published are not understandable for them, young participants became aware of complexity of LGU's business processes.

When recordings of LGU's council sessions become available for the public, two main benefits of implementing this activity were identified. One is related to better preparedness of council members – now when everything is recorded and available to the public, inactive council members (who don't give any ideas as a solution to problems that are discussed) are identified. Second benefit is related to involvement of citizens in decision making processes. Similar benefits are achieved when talking about recording LGU's head appearance on local radio (when the recording became available to the public).

ICT technologies enhanced the process of support NGOs via public tender (competition) for the allocation of funds for NGO projects in all phases: dissemination, receiving project applications and evaluation of application. In the same time, conducting this activity by using ICT helped both, NGOs and LGUs to understand the applying process to EU funds.

ACKNOWLEDGMENT

This work has been partly supported by European Commission under the project "Go, go, NGO!" (Erasmus+ program, project number: 2014-2-HR01-KA347-012471), implemented by Culture and Art Society Belec (KUD Belec, Croatia) and "Alexandru Stefulescu" Gorj County Museum (Romania).

REFERENCES

[1] S. Commission, "Strukturirani dijalog u Hrvatskoj | European Youth Portal." [Online]. Available: http://europa.eu/youth/node/20097_ro.

[2] European youth forum, "Position paper on structured dialogue," 2015.

[3] Zlatar Youth Association, "Being young in EU - How to start?," 2013.

[4] D. Potočnik, "Strukturirani dijalog s mladima u Republici Hrvatskoj," 2011.

[5] M. regionalnoga razvoja i fondova E. Unije, "Vrijednosti indeksa razvijenosti i pokazatelja za izračun indeksa razvijenosti 2013.," 2013. [Online]. Available: https://razvoj.gov.hr/o-ministarstvu/djelokrug-1939/regionalni-razvoj/indeks-razvijenosti/vrijednosti-indeksa-razvijenosti-i-pokazatelja-za-izracun-indeksa-razvijenosti-2013/3214. [Accessed: 18-Jan-2016].

[6] Salto Youth, "Young people with fewer opportunities," 2016. [Online]. Available: https://www.salto-youth.net/tools/otlas-partner-finding/help/young-people-with-fewer-opportunities/. [Accessed: 18-Jan-2016].

[7] R. S. Kaplan, "Conceptual Foundations of the Balanced Scorecard," *Handbooks Manag. Account. Res.*, vol. 3, pp. 1253–1269, 2009.

[8] Y. Laouris, A. Erel, M. Michaelides, M. Damdelen, T. Taraszow, I. Dagli, R. Laouri, and A. Christakis, "Exploring Options for Enhancement of Social Dialogue Between the Turkish and Greek Communities in Cyprus Using the Structured Dialogic Design Process," *Syst. Pract. Action Res.*, vol. 22, no. 5, pp. 361–381, Oct. 2009.

[9] E. Choudhury, "Attracting and managing volunteers in local government," *J. Manag. Dev.*, vol. 29, no. 6, pp. 592–603, Jun. 2010.

[10] C. D. O'Connor, "Engaging Young People? The Experiences, Challenges, and Successes of Canadian Youth Advisory Councils," 2013, pp. 73–96.

[11] T. Babila Sama, "Role of NGOs in the implementation of active labour market policies," *Int. J. Leadersh. Public Serv.*, vol. 8, no. 3, pp. 121–143, Aug. 2012.

[12] T. A. Engbers, "Building community? The characteristics of America's most civic cities," *J. Public Aff.*, , Apr. 2015.

[13] U. Sivarajah, Z. Irani, and V. Weerakkody, "Evaluating the use and impact of Web 2.0 technologies in local government," *Gov. Inf. Q.*, vol. 32, no. 4, pp. 473–487, Oct. 2015.

[14] J. C. Bertot, P. T. Jaeger, and J. M. Grimes, "Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies," *Gov. Inf. Q.*, vol. 27, no. 3, pp. 264–271, Jul. 2010.

[15] A. Planas, P. Soler, and M. Vilà, "Assessing youth policies. A system of indicators for local government," *Eval. Program Plann.*, vol. 45, pp. 22–28, Aug. 2014.

[16] E. Claver-Cortes, S. de Juana-Espinosa, and J. Valdés-Conca, "Emerging and Traditional ICT as Critical Success Factors for Local Governments," in *Emerging Mobile and Web 2.0 Technologies for Connected E-Government*, IGI Global, 2014, pp. 258–279.

[17] D. Agostino, "Using social media to engage citizens: A study of Italian municipalities," *Public Relat. Rev.*, vol. 39, no. 3, pp. 232–234, Sep. 2013.

[18] CIVITAS WIKI, "The use of social media to involve citizens in urban mobility projects and city planning," 2015.

[19] P. Averson, "The Deming Cycle," *Balanced Scorecard Institute*. .

[20] M. Martello, M. J. Fischer, and J. G. Watson, "Implementing a Balanced Scorecard in a Not-For-Profit Organization," *J. Bus. Econ. Res.*, vol. 6, no. 9, pp. 67–80, 2008.

[21] Z. Dobrovic, M. Tomicic, and N. Vrcek, "Towards the Effective E-Government Implementation of Balanced Scorecard in Public Sector," *Intelekt. Ekon.*, vol. 8011, no. 1, pp. 7–17, 2008.

[22] J. Brumec, M. Tomičić, and S. Brumec, "Konstrukcija mjernih instrumenata za Balanced Scorecard," *Proc. 18th Conf. Methods Tools Inf. Bus. Syst. Dev.*, pp. 21–30, 2006.

Using ICT Tools for Decision Making Support in Local Government Units

Nikola Kadoić* and Irena Kedmenec*
* Faculty of Organization and Informatics, Pavlinska 2, Varazdin
nkadoic@foi.hr, ikedmenec@foi.hr

Abstract - Decision making process is an everyday activity. However, special attention has to be payed when making decisions in public sector since those decisions result in long-term and costly consequences. Thus, ICT could improve the decision-making in public sector. The examples of ICT-supported decision making process in Croatian public sector are rare. The paper gives an overview of the successfully implemented ICT-supported decision making process in the Croatian local government units. The process was related to the project *Ivanscica for better tomorrow* which was supported by the European Commission. The aim of the project was to build the dialogue between the two local government units, City of Zlatar and City of Ivanec, in order to define strategic development activities for the mountain Ivanscica. Considering that the mountain is a part of two different local government units, when deciding about strategic directions and activities, it was decided to use ICT-supported AHP method. Calculations were made in both Excel and Expert Choice Trial. This application of ICT tools with multi criteria decision making method can also be a good starting point for new similar applications and possibly for the development of the methodology for decision making processes in government units.

I. INTRODUCTION

The project *Ivanscica for better tomorrow* was created and implemented under the Youth in action program (European Commision program four youth between 2007 and 2013) by Zlatar Youth Association (non-governmental organization, NGO) after young people from both Zlatar and Ivanec participated in training organized and conducted by Agency of Mobility and EU programs *Join now* (Croatian: *Uključi se!*). In the past, there were examples of conflicts between the two cities regarding the ownership of some land and buildings on the top of the mountain Ivanscica. Because of those conflicts, the area at the top of mountain was not developed in terms of tourism, even though there are potentials for that. Youth representatives from two cities decided to encourage solving this problem through the mentioned project. Through three three-day meeting during 4 months they analyzed the problem, identified touristic potentials of Ivanscica and suggested the first steps as the solution to that problem. The solution was presented to the local government units, City of Zlatar and City of Ivanec who had to decide about the acceptance of solution and implement it.

In the process of solving the problem project participants decided to use professional methods, such as SWOT analysis, creating strategies by combining SWOT elements, brainstorming techniques, scenario analysis and decision making methods. Before that, the heads of local government units have not used these techniques when making decisions.

The paper presents the application of two decision making methods supported by the ICT tools. Those methods are: AHP method (applied by using Expert Choice Trial) and ranking based on grades of participants (applied by using MS Excel). That was a novel approach to solving the problem in local public administration, since in Croatia there are no examples of ICT-supported decision making in small cities and municipalities.

II. STATE OF THE ART

The results of literature review (both, scientific and professional) showed that the examples of using ICT-supported decision making processes in local government units are rare. Professors Enrique Mu and Howard A. Stern in their paper presented the application of analytic hierarchy process (AHP) prioritization methodology, which used various criteria obtained from the stakeholders to develop a decision-making framework to select the most suitable cloud solution. The case allows a review of best practices in the cloud solution selection process, using theoretical lenses from the fields of IT alignment, decision-making, and stakeholder management [1]. Sayyadi and Awasthi applied the AHP on the problem of selecting the best location for pedestrian zones. The selection of best location from a list of potential stations involves consideration of different technical, economic, environmental, and social factors [2]. They used software called AHPPROJECT (now known as MakeItRational).

Authors of paper [3] dealt with the topic of creating successful public-private partnership model. This topic is very complex because public-private partnership (PPP) model includes different key stakeholders which have conflicted interests. Authors proposed the use of Bayesian network techniques in dealing with the problem. In the paper [4] the risk allocation in public–private partnership projects is discussed. Authors suggested an enhanced multi-objective optimization approach in order to achieve the fair allocation of risk between the PPP projects. The approach is based on knapsack method which is described and solved by applying the genetic algorithm.

Participatory Democracy is a relatively new phenomenon and it is related to including citizens in

decision making process about public issues. In paper [5], some examples of such decisions are presented. Communication between citizens and public administration can be implemented by using online tools. Chun and Cho presented the usage of an online tool, Cyber Policy Forum by the Seoul Metropolitan City Government (SMG) in Korea, used for citizen participation in the policy making process. They also suggested some additional features of the software in order to achieve a true citizen-government partnership in policy decision making [6]. Similarly, Insua, Kersten Rios and Grima discussed several mechanisms for participatory democracy and provided a framework for decision support in this area and described decision support functions that could be implemented in such a framework (which has to be implemented as a Web-based group decision support systems) [7].

Authors of paper [8] presented methodology which can be implemented in small and medium local governments units in order to deal with sustainability of energy action plans. Methodology was based on multi criteria decision making approach and Electre III method. It was implemented in the case of municipality in Lombardi region in Italy for the most appropriate strategy (set of activities) to deal with reducing CO_2 emissions.

When discussing complex decision problems in the area of public sector in Croatia, there is a lack of complex scientific studies like the ones presented in previous paragraphs of this chapter. This does not mean that ICT is not used in decision making in public sector because we can find some less complex but important examples where ICT is used in some phase of decision making process. The most often examples are the uses of ICT solutions in order to collect citizens' opinions when introducing some new law. Croatian government implemented web portal esavjetovanja.gov.hr and so far 400 public consultations were held by using this platform (starting from December 2014) and 19 public consultations are currently open [9].

III. CASE STUDY: IVANSCICA FOR BETTER TOMMOROW

The mountain Ivanscica is the natural border between the two cities (Zlatar and Ivanec) and the two counties (Krapina-Zagorje county and Varaždin county). More precisely, the border runs through the top of the Ivanscica where a mountain lodge is. Both cities claim the lodge and the top of the mountain. The space at the top of the mountain is undeveloped and unused because of the lack of desire from both sides to plan and realize potential joint activities in the area. There are no trails for hikers and bikers; there is neither skiing nor flying airfield nor other features that are common to the mountains, especially the ones that are in the cities. Unused natural potential results in the absence of any activities for the local population and unrealized possibilities for the tourism development that might generate income for both cities. Young people from the area, well aware of this problem and its cause, wanted to gather, analyze the problem and propose a solution in the form of specific guidelines on how to manage this area and use it for the benefit of the two communities. Thus, the project *Ivanscica for better tomorrow* was started.

The project *Ivanscica for better tomorrow* was a regional meeting of young people from two cities (Zlatar and Ivanec) and two counties (Krapina-Zagorje county and Varazdin county). It was funded by the European Union.

In the next part of the paper we present two decision making problems which were investigated under the project *Ivanscica for better tomorrow*. For each of them we will: describe the context of the problem, alternatives, criteria, decision making method and software that were used together with the results of decision making process.

A. Choosing the cooperation model between the two cities

The two local government units, City of Zlatar and City of Ivanec, are placed in two different counties in Croatia: one south of the mountain Ivancica and the other north of the same mountain. The border between those two cities passes through the top of the mountain. When comparing Ivanscica with other mountains in Croatia in terms of touristic offer and development it can be said that Ivanscica is totally undeveloped, although the development potential is undoubtedly present. The two cities desire to develop a touristic offer at the top of the mountain Ivanscica. They cannot do it separately (but can together) because of the ownership of the space at the top of mountain (which is shared), and because each city does not have the capacities and resources needed to develop that area alone.

Thus, the project *Ivanscica for better tomorrow* wanted to create foundations which would (or could) result in the ultimate goal of future developed touristic offer. First, a form of the cooperation between the two cities needed to be defined. It is important to note that the usage of AHP method and ICT tools was accepted by both local government units as the best way for the decision making process.

After several focus groups, three alternatives were created. They represented a way how the two cities could cooperate in the future trying to achieve the ultimate goal. The alternatives were:

1. The establishment of the Foundation *Ivanscica*,
2. Town twinning,
3. Cooperation Agreement between the two cities.

The criteria and sub-criteria that were used to describe and evaluate alternatives were as follows:

- Finance: the finance needed to implement an alternative and the finance needed to manage an alternative (in terms of staff costs, overheads and similar);
- Legal complexity: the legal complexity of implementing an alternative and legal complexity of managing an alternative;

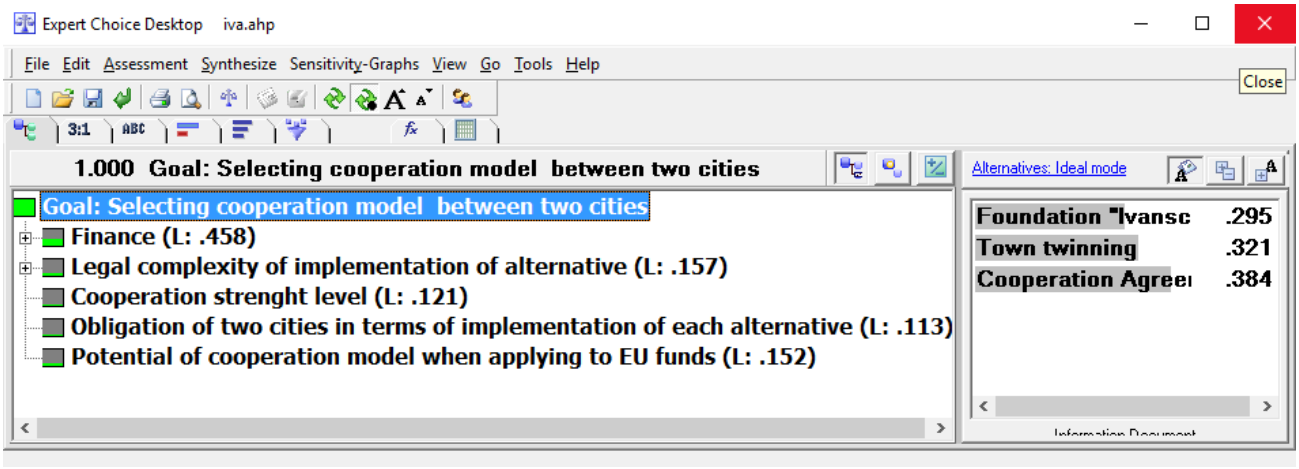


Figure 1. AHP model in Expert Choice Trial (goal, criteria, alternatives)

- Cooperation strength level of certain alternative;
- Obligations of two cities in terms of practical implementation of each alternative
- Potential of alternative when applying to EU funds

The general idea behind the establishing some form of cooperation (alternatives) is starting the collaboration between the two cities at both formal and practical level. The chosen form of cooperation should be managed by both parties i.e., representatives from both cities. Main activities that would be done under the chosen alternative were also defined in the project. The two cities would have the obligation to invest funds to the bank account of chosen form of cooperation (size would be agreed each year separately). The money would be used for functioning of alternative and for public tendering for small projects which support the cooperation between the two cities. Tenders would be called for the projects of institutions, organizations, schools and associations from both cities. The rule would be that each application has to have two applicants, one from Ivanec, and the other from Zlatar. General motivation for doing those activities is creating a set of different projects, initiatives and activities between the two cities which would be helpful when applying for the EU funds (EU tenders evaluate project applications with more points if there are some previous forms of cooperation between the applicants).

After alternatives and criteria were defined and each alternative was described in terms of criteria, the AHP method was applied by using Expert Choice Trial software. The Analytic Hierarchy Process or shorter the AHP approach was founded in 1980 by Thomas Saaty. The AHP is one of the most popular methods of professional scenario analysis and decision-making based on the evaluation of the hierarchies that include objectives, scenarios, criteria and alternatives. It is one of the most proven methods of decision-making by creating and evaluating the hierarchy problem [10]. It is a mathematical model of multi-criteria decision making that uses the decomposition of a complex unstructured situation into simpler parts – the hierarchical system. Using subjective pairwise comparisons numerical values are assigned to individual components, showing their relative importance [11].

AHP provides flexibility to the decision making process and helps decision makers to set priorities and make the best decision, taking into account both qualitative and quantitative aspects of the decision. The

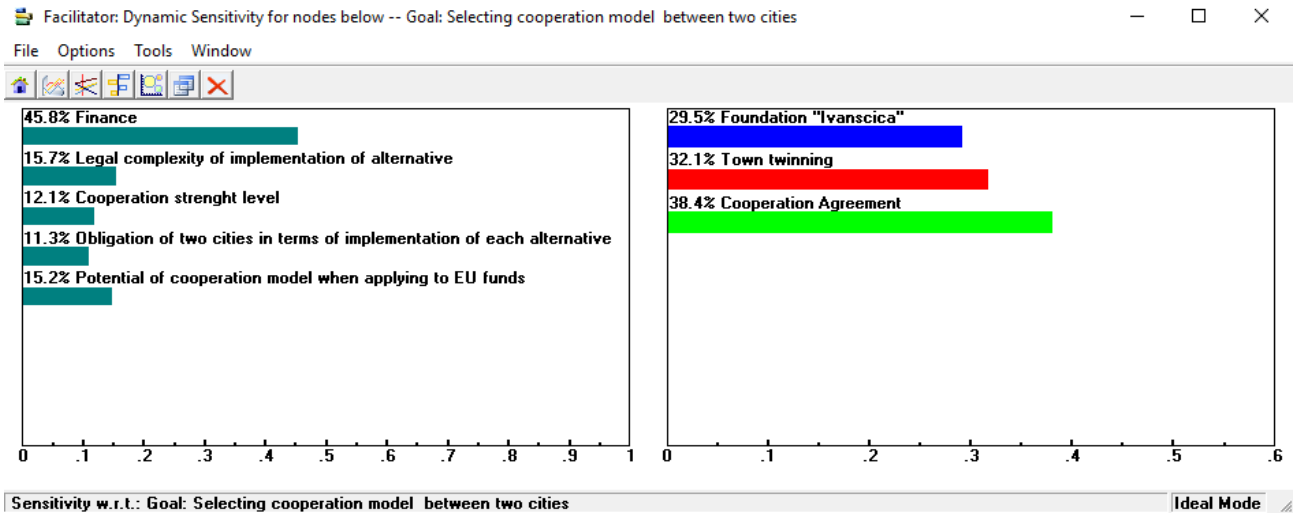


Figure 2. Dynamic sensitiviy analysis - screenshot

value of this method is that it synthesizes information from decision-makers and other stakeholders who have knowledge about the problem, to identify the problem and to harmonize the views of its structure [12]. The AHP structures the problem in levels by defining criteria, sub-criteria and alternatives. By breaking the problem into levels the decision-maker can focus on smaller sets of decisions. Because of these characteristics, AHP method was applied to the described problem.

After creating the AHP hierarchy tree in Expert Choice, and doing the pairwise comparisons it was concluded that the highest global priority has the alternative 3. Figure 1 shows the implementation of hierarchical tree in Expert choice as well as calculated weights and global priorities of alternatives. In Figure 2, dynamic sensitivity analysis is presented. Weight of each criterion (left side) was increased and decreased for 5% and the changes in global priorities of alternatives were observed (right side). After all changes of weights were made, the highest global priority of alternative 3 was confirmed.

B. Choosing the activity for the first cooperation between the two cities

The second decision making problem was related to choosing the activity that would be implemented as the first cooperation between the two cities on the Ivanscica mountain. That implementation should have served as an example to all possible applicants when applying small projects to the tender of Zlatar&Ivanec.

At the beginning of the project during the European Youth Week (26th May 2013 – 2nd June 2013), the youth got familiar with the fundamental values of the EU, priorities and guidelines of the EU so they could be applied in the solving process of the described problem. Additionally with acquiring the knowledge essential to the theme of the problem, young people learned about the strategy development tools and techniques for creative thinking (the problem tree, SWOT).

In the first meeting, the participants prepared materials on the potential of Ivanscica. The following topics were included: natural geographic characteristics of the area (climate, vegetation, geological development, main speleological objects, protected natural resources in the area), socio-geographic characteristics of the area (the existing infrastructure on the mountain and in its surroundings, art and legends inspired by the mountain Ivanscica, the existing events taking place at Ivanscica, a short review of 31 existing NGOs in the area) and the development potential of Ivanscica. Based on this input, the participants performed the SWOT analysis. SWOT analysis is an analytical method which is used to identify and categorize significant internal (Strengths and Weaknesses) and external (Opportunities and Threats) factors faced either in a particular arena, such as an organisation, or a territory, such as a region, nation, or city (European Commission, Joint Research Centre, 2013). In this case, SWOT analysis was performed for Ivanscica's territory. The participants worked together and identified strengths, weaknesses, opportunities and threats.

The identified strengths of Ivanscica's territory were its natural potential, existing events, NGOs' activities, rich tangible and intangible cultural heritage, proximity of urban areas – cities with greater number of citizens compared to rural areas, potentially rich gastronomic offer, kindness and hospitality of the local population, initiative for cooperation of the two cities (Ivanec and Zlatar) and hydrologic resources. On the other side, weaknesses included the lack of money, weak local transport connections, the lack of accommodation and catering facilities, the lack of promotional materials, inertness of people, lack of education and interest, and inadequate infrastructure. Opportunities were listed as follows: funds from the European Union, implementation of the EU and Croatian strategies, economic recovery, and underdevelopment of the area. The identified threats included poor responsiveness of people on different social initiatives, restrictions due to regulations, political disagreements, mentality (envy of people), resistance of a certain number of people, the land register, and the old age of population.

Strategic logic requires that the future pattern of actions to be taken should match strengths with opportunities, ward off threats and seek to overcome weaknesses [13]. Thus, the SWOT analysis was a good starting point for the further discussions on possible activities that could support the development of the Ivanscica's territory.

After the SWOT analysis was made, in the second meeting the participants generated ideas for possible activities that may take place on the mountain. Participants proposed their ideas, but also commented on other's ideas. The ideas that were presented were the following: Hiking trails (activity encompasses marking and mapping of the hiking trails and the digitalization of the maps and their publication on the web), Educational workshops (education and experience exchange in writing project proposals intended for NGOs whose area of interest is Ivanscica), Touristic map (making and digitalization of the touristic offer map with periodic update), Summer on the Ivanscica (a major manifestation in July or August in the duration of two weeks), Crne Mlake (the use of legend about gathering of witches at the place called Crne Mlake on the mountain in order to create an interesting and amusing hiking route that could be presented for the Halloween), Photo hunting (the encounter of photo amateurs and photography enthusiasts who would take photos of the nature – the best photos would be chosen and put together for an exhibition), Wild grown herbs (a workshop on wild grown herbs, the production of creams and ointments of ruffled plants), Eco action (garbage collection on the Ivanscica area that would include all the NGOs and all people of good will who want to join the action), Traditional sports tournament on the top of the Ivanscica (tournament that would encompass a variety of traditional games such as sack race and long jump), Website (creating a website with all the information about the Ivanscica in general and about current events with regular updates), School in nature (organization of a school in nature program for primary or secondary school students), Resting places (building wooden benches, waste bins, fences and observation

points for hikers), Paintball terrain (arranging the field taking advantage of the existing military facility and procure the equipment for paintball, Soccer tournament (arranging the football field and organizing a football tournament during the spring and summer), Adventure on Ivanscica (manifestation that would include visiting the fortresses on Ivanscica).

The first filter for the activity selection was the time that was necessary to prepare the activity since according to the project plan *Ivanscica for better tomorrow* the activity had to be realized in August 2013 and had to fit into the project budget. Thus, from the proposed activities, those that could be realized in the time frame of one month were chosen. These activities were: (1) hiking trails, (2) educational workshops, (3) photo hunting, (4) eco action, (5) traditional sports tournament on the top of the Ivanscica.

Project participants had to select one activity which would be implemented. The method used for this selection was *ranking based on grades of participants* [14]. Implementation of this method was done in MS Excel. Each of young project participants evaluated each of proposed idea and gave a grade between 1 and 10 where higher grade meant higher wish for implementation.

After evaluating/grading each activity the SUM of all given grades per each participant was calculated. After that each participant's grade was divided with the sum of all given grades of that participant. Finally, global priorities of each activity were calculated (average of values from previous step for each activity).

	1	2	3	4	5	S	1	2	3	4	5
Participant 1	5	6	7	5	9	32	0,16	0,19	0,22	0,16	0,28
Participant 2	6	8	9	4	8	35	0,17	0,23	0,26	0,11	0,23
Participant 3	5	5	8	6	7	31	0,16	0,16	0,26	0,19	0,23
Participant 4	6	7	8	8	9	38	0,16	0,18	0,21	0,21	0,24
Participant 5	7	8	6	7	7	35	0,20	0,23	0,17	0,20	0,20
Participant 6	7	4	6	8	9	34	0,21	0,12	0,18	0,24	0,26
Participant 7	7	6	5	8	10	36	0,19	0,17	0,14	0,22	0,28
Participant 8	6	7	10	8	9	40	0,15	0,18	0,25	0,20	0,23
Participant 9	5	7	6	8	9	35	0,14	0,20	0,17	0,23	0,26
Participant 10	10	7	8	6	9	40	0,25	0,18	0,20	0,15	0,23
Participant 11	7	6	9	8	10	40	0,18	0,15	0,23	0,20	0,25
Participant 12	7	6	8	9	9	39	0,18	0,15	0,21	0,23	0,23
Participant 13	8	6	7	5	9	35	0,23	0,17	0,20	0,14	0,26
Participant 14	7	6	10	8	9	40	0,18	0,15	0,25	0,20	0,23
Participant 15	7	6	8	5	10	36	0,19	0,17	0,22	0,14	0,28
Participant 16	7	5	6	8	9	35	0,20	0,14	0,17	0,23	0,26
Global priorities of activities							0,18	0,17	0,21	0,19	**0,24**

Figure 3. Choosing 3rd meeting activity by using ranking based on grades of participants – implementation in Excel

The best alterative was a tournament in traditional sports (highest global priority, 0,24). The third meeting was held on the 01 September 2013 at the top of Ivanscica. The partners on the project organized the tournament in traditional sports that attracted 300 participants. The sports tournament included the traditional sports games: tug of war (pooling the rope), long jump, sack race and skiing on grass. The program of the event beside the main sports part also included the informative part and entertainment.

At the beginning the participants were introduced to the project *Ivanscica for better tomorrow* during an interesting presentation about the project progress and results. After the tournament, the participants could enjoy in the two short plays performed by the cultural and artistic association Staznjevec. At the end of the day the impressions were positive. This meeting was a good opportunity for the two cities to get to know each other better. By organizing this activity the youth wanted to show to the associations and institutions that they are also capable of organizing events in the area of Ivanscica and that they should apply for the funds that the two cities will provide in the future.

The main result of the project was the adoption of The Act for Supporting the Cooperation between the City of Zlatar and the City of Ivanec. By this act a fund for the development of Ivanscica should have been established. The both cities would have given funds for which the associations and institutions from the both cities would have been able to apply. The idea was to realize projects that would develop the area of Ivanscica. The preconditions to apply for the funds were to find a partner from the other city (so both a partner from the City of Ivanec and a partner from the City of Zlatar would be included) and to design a project that would contribute to the development of the Ivanscica. In addition to the establishment of the fund, the act should have included a toolbox of activities, i.e., potential project topics for those organizations that do not have their own suggestions. These ideas are those that were generated as the result of the second meeting.

Finally, the project ended with deliverables that were accepted by both cities. In the post-project period the Act was not completely implemented as it was initially planned: there are no public tenders for small projects for organizations, associations and institutions. Reasons for that were not given by responsible representatives, but the reason for that might be the fact that the activities which should be supported by the two cities are too small in scope and thus might seem less important to the local government units' representatives. Maybe their perspective will change in the future. However, the main idea of signing the Act is implemented, because the cooperation between the two cities, which was initiated by the project, continued after the project was completed.

IV. CONCLUSION

The project *Ivanscica for better tomorrow* was designed as cooperation between the two cities, City of Zlatar and City of Ivanec, that would lead to the eventual development of the natural border between those two cities – the mountain Ivanscica.

The project was important for young people to develop structured dialogue with decision-makers and to actively participate in regional and local community. This was an opportunity for various organizations of civil society, especially the youth, to demonstrate their skills and

contribute to their community and also an opportunity for decision-makers to better meet the needs of its young citizens, to work together and jointly achieve the maximum benefit of their communities.

This paper presented two examples of decision making methods supported by ICT tools in solving important problems in LGUs. The usage of those methods as well as the belonging software support was accepted by LGUs, even though in the past they have not used similar solutions.

ACKNOWLEDGMENT

This paper was created as a result of project *Ivanscica for better tomorrow* (project number HR-51-E2-2013-R1) which was partly funded with support of European Commission, Youth in action program. Youth in action program supported different activities and initiatives that were based on the concept of informal learning.

As a software support for informal learning in this project software Expert Choice Trial was used.

REFERENCES

[1] E. Mu and H. A. Stern, "The City of Pittsburgh goes to the cloud: a case study of cloud solution strategic selection and deployment," *J. Inf. Technol. Teach. Cases*, vol. 4, no. 2, pp. 70–85, Jan. 2015.

[2] G. Sayyadi and A. Awasthi, "AHP-Based Approach for Location Planning of Pedestrian Zones: Application in Montreal, Canada," *J. Transp. Eng.*, vol. 139, no. 2, p. pp 239–246, 2013.

[3] J. Xie and S. T. Ng, "Multiobjective Bayesian Network Model for Public-Private Partnership Decision Support," no. September, pp. 1069–1081, 2013.

[4] V. Alireza, Y. Mohammadreza, R. M. Zin, and N. Noor, "An Enhanced Multiobjective Optimization Approach for Risk Allocation in Public Private Partnership Projects : A Case Study of Malaysia An Enhanced Multiobjective Optimization Approach for Risk Allocation in Public Private Partnership Projects : A Case Stu," vol. 177, no. November 2013, pp. 1–46, 2014.

[5] M. Danielson, L. Ekenberg, A. Ekengren, T. Hökby, and J. Lidén, "Decision process support for participatory democracy," *J. Multi-Criteria Decis. Anal.*, vol. 15, no. 1–2, pp. 15–30, Jan. 2008.

[6] S. A. Chun and J. S. Cho, "E-participation and transparent policy decision making," *Inf. Polity*, vol. 17, no. 2, pp. 129–145, 2012.

[7] D. Rios Insua, G. E. Kersten, J. Rios, and C. Grima, "Towards decision support for participatory democracy," *Inf. Syst. E-bus. Manag.*, vol. 6, no. 2, pp. 161–191, 2008.

[8] G. Dall'O', M. F. Norese, A. Galante, and C. Novello, "A multi-criteria methodology to support public administration decision making concerning sustainable energy action plans," *Energies*, vol. 6, no. 8, pp. 4308–4330, 2013.

[9] Croatian Government, "E-Savjetovanja," 2015. .

[10] N. Begičević and S. Vidović, "Grupno odlučivanje AHP metodom podržano sa Team EC 2000," *Bus. Process Conf. Zagreb*, 2006.

[11] H. Tomaškova and M. Gavalec, "Preference Matrices in Tropical Algebra," pp. 1–7, 2010.

[12] N. Tomić-Plazibat, "Višekriterijalna analiza u investicijskom odlučivanju," Sveučilište u Zagrebu, 1994.

[13] European Commission and Joint Research Centre, "SWOT Analysis," *European Commision*, 2013. [Online]. Available: http://forlearn.jrc.ec.europa.eu/guide/4_methodology/meth_swot-analysis.htm.

[14] D. Munđar and T. Hunjak, "Kvantitativni menadžment: Odlučivanje i rangiranje," 2012.

The Conceptual Risk Management Model - A Case Study of Varazdin County

R. Kelemen*, M. Biškup** and N. Begičević Ređep***
* Varazdin County/County Governor's Office, Varazdin, Croatia
robert.kelemen@vzz.hr
**matija.biskup@gmail.com
*** University of Zagreb, Faculty of organization and informatics, Varazdin, Croatia
nbegicev@foi.hr

Abstract – Risk Management has become a very important issue in the corporate management in recent years. Similarly, Public Sector recognized the importance of Risk Management but the implementation of Risk Management principles in Croatia is mainly focused on financial aspects with subsequent checks. This paper presents the Conceptual Enterprise Risk Management Model developed in Varazdin County, using the Method engineering principles. The Method engineering approach used in this model describes the steps and methods which have to be taken, the precise order and time in the Risk Management Process in public sector. Developed Conceptual Enterprise Risk Management Model has been tested on the Project "Development of broadband infrastructure in the areas of Varazdin County without sufficient commercial interest for investment".

Keywords: Risk Management, Public sector, Government, Conceptual Model, Method Engineering

I. INTRODUCTION

The Risk Management is becoming very important issue in all levels of government. Some risk management requirements are implemented in various laws. Despite all that requirements the implementation of risk management is not satisfactory. Why is that happening? There are few reasons. In various laws only the financial aspect is considered. The controls are performed at the end of the process. The process approach is neglected. Further, there is no problem understanding and systematic approach.

Considering all that issues, the authors present possible solution for implementing the risk management process in organization, specifying the methods, sequence and data flows.

The main contribution of this paper is the Conceptual Enterprise Risk Management Model which can be implemented and used for risk management in processes and projects as well.

This conceptual model has been implemented in Varazdin County's Quality management system and tested in the first stage of the Project "Development of broadband infrastructure in the areas of Varazdin County without sufficient commercial interest for investment".

This paper consists of five main sections and the conclusion. The State of the art section describes the problem domain and points out some major issues concerning Risk Management in public sector. The second section briefly describes the international standard ISO 3100:2009 which is used as the base for the Conceptual Enterprise Risk Management Model development. The whole concept of the new model development is complex so it requires the knowledge of some methods suitable for this task as well as the method engineering method. The third section describes the methods implemented in the conceptual model. The following section describes the logic, sequence and connection of the Conceptual Enterprise Risk Management Model parts. The Case Study of Varazdin County is described in the fifth section. Finally, paper ends with the conclusion.

II. PREVIOUS RESEARCH

Despite the fact that Risk Management is not a new concept in government sector and it has been recognized as an important issue, the implementation of risk management in business processes in government is often deficient. The problem occurs within organizations which persist in traditional and outdated practice of managing risks within functional silos and stovepipes [7]. The principle of functional silos is even more highlighted in traditionally hierarchical organizations like governments. The few main issues of project management in ICT projects in all levels of the Croatian governments have been detected [10]. The main issues are that the risk management activities are not implemented and the control activities are performed mainly at the end of the project cycle by the State Audit office, when it is not possible to make corrections [10]. The more efficient mechanism has been introduced in public sector in 2012. when the new law on Public Internal Financial Control (PIFC) has been introduced [1]. This Law introduces Risk Management principles, but the main problem is that it has been strongly focused on financial risks. The authors [10] pointed out that Public procurement Law obligates [12] public client to monitor the contract fulfillment and emphasizes the issue of managing IT projects. Here, it can be noticed that it is not necessary to stick just to the IT projects, the good solution is to implement risk management in organizations and consequently in the project management.

According to the research [8] on Croatian non-financial companies, they stick primarily with simple risk management instruments like natural hedging. Among the ones that use derivatives, forwards and swaps are by majority the most important instruments. The majority of analyzed companies do not have a documented risk management policy neither do they use any type of simulation techniques [8]. Some of the reasons for not using risk management are: insufficient supply of risk management instruments offered by domestic financial companies, the high cost of implementing and maintaining risk management programs and difficulties in pricing and valuing derivative instruments [8].

Enterprise Risk Management (ERM) has been recognized as the process for making this integration work. ERM is defined as "a process, effected by an entity's … management and other personnel, applied in a strategy setting and across the enterprise, designed to identify potential events that may affect the entity, and manage risk to be within its risk appetite, to provide reasonable assurance regarding the achievement of entity objectives" [5].

III. THE INTERNATIONAL STANDARD ISO 31000:2009

The international standard ISO 31000:2009 - Risk Management can be used by any organization regardless of its size, activity or sector [2]. This framework is intended to assist the organization to integrate risk management into its overall management system [2]. That is very important, because the organizations can use it instead of developing their own. For the purpose of developing Conceptual Risk Model, the risk management process of this framework is used and therefore it is necessary to describe it briefly.

Using ISO 31000 organizations increase likelihood of achieving objectives, improve the identification of opportunities and threats and effectively allocate and use resources for risk treatment. Using this standard, organizations can compare their risk management practices with an internationally recognized benchmark, providing sound principles for effective management and corporate governance. ISO 31000 cannot be used for certification purposes, but it provides guidance for internal or external auditors [2].

The framework assists in managing risks effectively through the application of the risk management process (Fig. 1) and ensures that information about risk derived from the risk management process is adequately reported and used as a basis for decision making. The framework consists of following components: Mandate and commitment, Design of framework for managing risk, Implementing risk management, Monitoring and review of the framework and Continual improvement of the framework [2].

Risk management process comprises following activities: Communication and consultation, Establishing the context, Risk assessment (which consists of Risk identification, Risk analysis and Risk evaluation), Risk treatment and Monitoring and review (Fig. 1) [2].

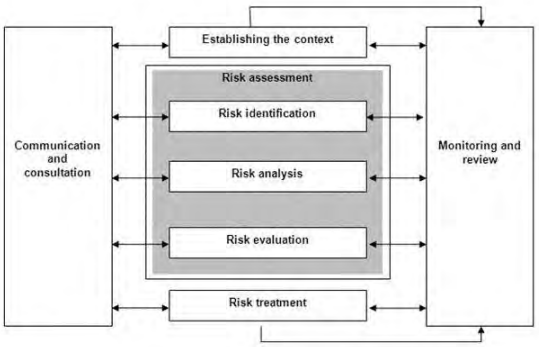


Figure 1. Risk management process [2]

IV. THE METHODOLOGY FOR CONCEPTUAL RISK MANAGEMENT MODEL DEVELOPMENT

Implementing risk management process in organization is a complex task and requires knowledge of different methods. Since, the new model has been developed; the knowledge about method engineering approach (metamodeling method) is required. To successfully implement Enterprise Risk Management (ERM) in any framework, it is necessary to identify and apply appropriate methods which will provide the data and results required for every stage of the process. A good combination and sequence of methods is a prerequisite for ERM.

Regardless of some suggestions of authors [4] which have developed different model of Risk Management process (RMP), the ISO 31000:2009 framework (Fig. 1) has been selected for the new model development.

This section briefly describes the methods used for the conceptual model development.

A. The metamodeling method

The metamodeling method has been used to explore and formalize a new model and to ensure that the requirements of risk management process are implemented.

Since one of main components of the risk management process is Communication and Consultation, the principles of Information Systems (IS) development can be used for Risk management process. According to [9], there are two research domains in Information Systems (IS) development nowadays: Problem System domain (PS) and Problem Solving System domain (PSS). Problem System relates to organization that we want to develop IS for. Problem solving system consists of components that enable IS designers to identify, design and perform the changes within the PS, which will pave the way to future development of organization and its IS. The most important part of the PSS is a set of methods they can use to develop new or improve existing IS. Activities that designers perform in PS using methods from PSS are called, in general, modeling [9].

However, when the nature of change that is going to be performed within the PS is complex, it can happen that the available set of methods within the PSS is not sufficient for performing the mentioned organizational or IS change in PS. Then the focus of research transits from

the PS to PSS. Processes that take place therein are called metamodeling, method development, or method engineering [9]. The primary goal of metamodeling is to explore and formalize new method that will be successfully applied for performing necessary changes in PS [9].

Authors [3] in 2010 presented the formal description of method engineering which can be used to describe each part or element of almost any IS method. Therefore, it can be applied to get information about the structure and the underlying assembling technique [3]. This knowledge can help to develop new methods and to reconstruct existing methods in order to adapt them to the project situation at hand [3].

B. The SWOT Analysis

The SWOT analysis method is suggested for context analysis in risk management process. The SWOT analysis is a structured analysis matrix for categorizing the factors of organizations functioning [6]. In the matrix, strengths, weaknesses, opportunities and threats are analyzed, including both internal factors controlled by an organization (e.g., marketing, finances, etc.) and external factors (e.g., political and economic factors, competition, etc.) [6]. The SWOT analysis can be used multiple times in the model.

C. The Work Breakdown Structure

A Work Breakdown Structure (WBS) is a project deliverable that organizes the team's work into manageable sections [13]. It visually defines the scope into manageable parts that a project team can understand with each level providing further definition and details. A WBS starts with the project as the top level deliverable and each functional deliverable is further decomposed into sub-deliverables [13]. These sub-deliverables are then further decomposed until a single person can be assigned. At this level, sub-deliverables contain work packages that are assigned to organization's department or unit [13]. The work package defines the work, duration and costs for the tasks. It is important that work packages are independent and not duplicated across the WBS. By using WBS, a project schedule and budget can be quickly calculated by allocating time and cost estimates to specific sections of the WBS [13]. An essential part of WBS is the 100% Rule. Using the 100% Rule, the top level contains everything, meaning that is 100% done. Each level then breaks down and shows effort that is needed to complete it. However, the 100% Rule also applies at every level below that, meaning that every level also has 100% [13].

D. The Risk Breakdown Structure

A Risk Breakdown Structure (RBS) is similar to the Work Breakdown Structure (WBS), as it is hierarchical representation that starts from higher levels and goes down to lower levels [13]. In RBS a structure of the risks that impact project or activity is created [13]. By using the RBS, project risks are grouped and organized in categories. Each category is then broken down into levels, with every level detailing the source of risks for the project [13]. Going deeper into each level, potential risk triggers can also be identified. This way it is easier to determine the root cause of risks, understand risk exposure to a project and identify risk dependencies [14].

Since the implementation of Enterprise Risk Management within the organization (PS) is dynamic and complex, it requires research in PSS. That research should result in the design of the new method of the risk management which can be implemented in PS. The focus of this research is the risk management process and its dynamics.

V. THE CONCEPTUAL ENTERPRISE RISK MANAGEMENT MODEL

The conceptual enterprise risk management model has been developed in accordance with the risk management process (RMP) (Fig. 1). For every stage of RMP, the appropriate method, described in previous section, has been selected, analyzed and represented as entity-relationship model. Relationships between objects ensure that the communication and consultation requirement, as well as monitoring and review requirement of (RMP) have been fulfilled. The following subsections describe each stage.

A. Stage 1 - Establishing the context

The first stage of the RMP is to establish the context. This stage requires the understanding of the organization and its context. Consequently, the first entity of the new model is ORGANIZATION. This entity will contain information about the organization and its parts – departments (Fig.2 – The organization entity).

As stated earlier, the SWOT analysis method is suggested for context analysis. Each area is explored from the perspectives of internal strengths and weaknesses, as well as external opportunities and threats. Elements that present strengths, weaknesses, opportunities or threats in each area of assessment should be identified and given certain weight with which they contribute in the overall readiness assessment. These elements will be stored in SWOT_NAME, SWOT_ELEMENT and SWOT_AREA entities (Fig. 3). SWOT analysis can be performed multiple times in the organization. Therefore, it is connected with organization and risk treatment plan entities using foreign keys.

To understand the organization or the project, it is necessary to discover what the processes are, and to decompose them into smaller parts - activities. For every activity and process it is necessary to discover whose responsibility it is. This information will be stored in the WBS, RESPONSIBILITY and EMPLOYEE entities. In some cases one person can have numerous activities. Consequently, that information is stored in RESPONSIBILITY entity and the keys of WBS and EMPLOYEE become foreign keys in that entity. Primary

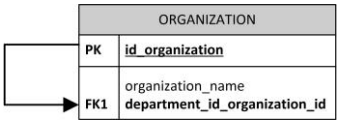


Figure 2. The Organization entity

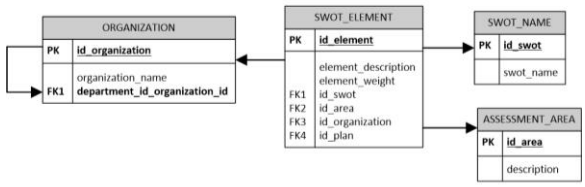


Figure 3. SWOT and organization entities

key of ORGANIZATION becomes a foreign key in EMPLOYEE entity. A unary relationship in WBS entity is created as the process can have its sub processes and activities (Fig 4.)

B. Stage 2- The risk assessment

The second step in RMP is the Risk Assessment with sub steps: Risk Identification, Risk analysis and Risk Evaluation. To perform this step it is necessary to define some threat treatment rules and to develop the Risk Matrix. That means that the rule name, number of risk tolerance levels, risk model and model variables has to be set up. This data will be stored in RISK_MATRIX, RISK_MODEL and RM_VARIABLES entities. After that, in order to develop the risk matrix, the criteria for the risk assessment have to be defined. ASPECT and CRITERIA entities will store the information on risk aspects and criteria rank with criteria description. The primary key of RISK_MATRIX entity is foreign key in CRITERIA entity (Fig. 5). When all the rules are set up the risk matrix for every aspect is calculated. The responsibility of the top management is to accept or modify the suggested Risk matrix. The risk matrix is crucial for the Risk evaluation sub step.

The Risk identification sub step could begin with threat source identification. The SWOT will provide the necessary inputs for threat source identification and classification. Information about threat sources will be stored in THREAT_SOURCE entity and that entity is connected with SWOT_ELEMENT entity using foreign key. The hierarchical representation of risks is represented using Risk Breakdown Structure (RBS) (Fig. 6). RBS contains the risk that could impact the project or the process. RBS entity stores information on risk breakdown structure, including unary relationship, because it has a hierarchical structure. Naturally, every risk has to be connected with risk source so the RBS entity has foreign

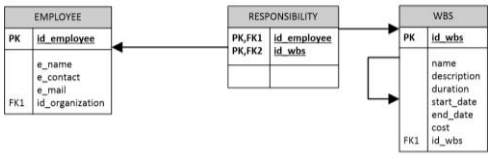


Figure 4. WBS entities

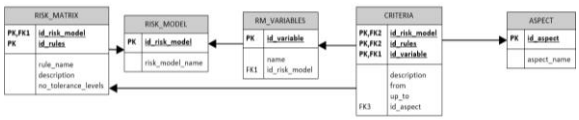


Figure 5. Threat treatment rules entities

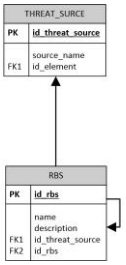


Figure 6. RBS entities

key which is the primary key of THREAT_SOURCE entity (Fig 6.).

That brings us to the next sub step – the Risk Evaluation. Every activity has to be related with one or more risks. Potentially, in some cases it is possible that more than one activity can be related to more than one risk, the new WBS-RBS entity is introduced and stores data on that connection. After establishing connection, risk evaluation has to be carried out. The Risk evaluation means that for every activity with every identified risk for every predefined aspect the value of variable has to be estimated. That information will also be stored in WBS-RBS entity, along with connections with RBS, WBS, ASPECT and RM-VARIABLES (Fig. 7). The Risk Evaluation is a very important issue, because after risk calculation, it is the basis for decision making on which risks have to be treated, monitored or accepted.

C. Stage 3 – The risk treatment

The risk treatment is the next step in the Risk Management process. The top management or the project manager has to make a decision on which risks have to be treated on the grounds that they are unacceptable. For those risks the Risk Treatment Plan has to be developed. The Risk Treatment Plan should include the activities defined to reduce risks. For each activity it is necessary to *define a responsible person, planned and actual costs, start* date, end date. All that data will be stored in RISK_TREATMENT_PLAN and RISK_TREATMENT_TEAM entities. These entities are connected with WBS-RBS and EMPLOYEES entities.

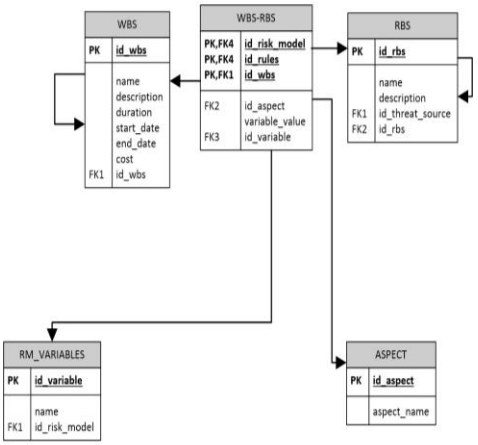


Figure 7. Risk Evaluation entities

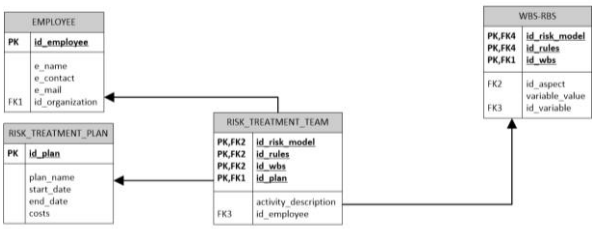


Figure 8. Risk Treatment entities

The treatment plan is connected with the SWOT_ELEMENT entity because in some other iteration it will be analyzed by SWOT method to evaluate the corrective actions in the context of reducing identified risks (Fig. 8.).

The complete presentation of all entities and their relationships represents The Conceptual Enterprise Risk Management Model (Fig. 9). The model has been developed using consistent application of metamodeling principles.

VI. The Case Study of Varazdin County

The following part of the paper describes the implementation of the conceptual enterprise risk management model suggested in previous chapter. The model has been tested on the project "Development of broadband infrastructure in the areas of Varazdin County without sufficient commercial interest for investment" with focus on meeting the requirements of Digital Agenda for Europe 2020.That was the great opportunity to test the model on the existing project in early stage of implementation.

The main constraint was the possibility to perform risk management only with the first few stages of the overall project were the overall insecurity is very high, but on the other hand it is the good example of iterative nature of the conceptual risk management implementation model.

To be able to successfully implement all 14 identified activities of the conceptual model, the workflow model have been developed (Fig. 10). The workflow model represents the activity sequence, competences and data flows among activities and it is the way of the model

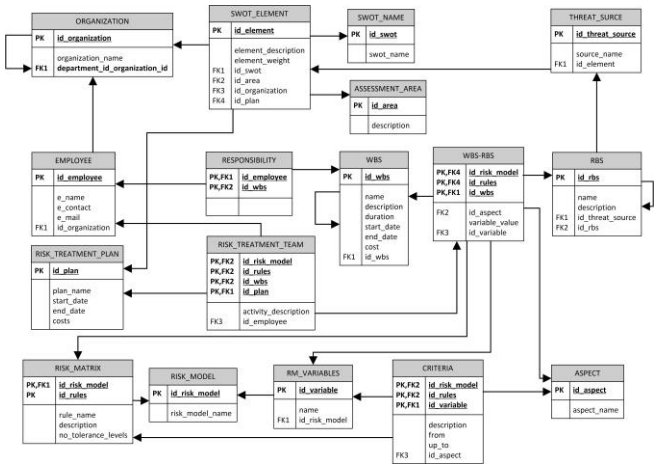


Figure 9. The Conceptual Enterprise Risk Management Model

implementation. The activities are connected with 17 data flows.

Since, the implementation of the suggested model is performed on project, the process starts with the Define the project activity, and consequently acquiring the project team. The project team's first task is to perform the SWOT analysis which results with the context assessment, Risk source identification and SWOT analysis data. Based on outputs of SWOT analysis, activities Develop WBS, Develop RBS and Risk treatment Rule Development are performed. Inputs for the Develop WBS activity are Project Charter and SWOT analysis. The activity results with WBS register flow which is the input of Responsibility Association activity. This activity suggests the sponsor of the project how to associate the WBS with the responsible person for every activity. The sponsor then makes decision and designates the activity owners. The activity owners list is an input of the Risk Evaluation Activity.

Another output of Perform SWOT analysis activity is the Risk source identification which is used as an input of Develop RBS activity. The activity develops hierarchical RBS structure. The third output of the Perform SWOT analysis is the Context Assessment flow. This flow is the input of the Risk Treatment Rule development Activity. That activity is very important because it determines the basis for evaluation of the whole system. In this activity the Team suggests the Risk model, rule names, tolerance to risk level (in this example the 3 level of risk tolerance, green, yellow and red is used) and aspects of risk for the risk evaluation followed by the definition of criteria evaluation.

This case study uses the short risk model for implementation with variables Impact (I) and Probability (P). The risk (R) is then calculated as R=(P*I). The used aspects are Quality, Time and Expense. Every aspect has to be described concerning the Impact, Probability and Tolerance criteria. The implemented model defines ranges for an Impact criterion from 1 to 5 where 1 is insignificant, 2 is small, 3 medium, 4 high and 5 – very high. Similar ranges have been defined for Probability, where 1 means unlikely, 2- possible, 3 – likely, 4 – very likely and 5 – inevitably. The Tolerance criteria have range from 1 to 3 where 1 means low risk, 2 moderate risk and 3 high risk. Having all that rules defined, the Risk Matrix is calculated for every aspect. The decision of acceptance rules and risk matrix must be made by the authorized person e.g. project sponsor and it is made in the Designate Risk Matrix activity. Output of that activity is Risk Matrix flow.

The Risk Evaluation activity has 5 input flows, WBS, RBS, Activity owners list, Risk Matrix and Treated Risk. Based on Risk Matrix every activity form WBS with designated owner is connected with one or more threats from RBS. Having established the connected structure with WBS and RBS, it is necessary to evaluate every activity with every connected risk concerning every aspect providing the values for the Impact and Probability variables. This activity should result with Evaluated Risk List flow e.g. Risk Register. The project sponsor must then decide which risks are acceptable, unacceptable,

transferable or monitored. Naturally, unacceptable risk has to be treated. For every unacceptable risk Evaluated risk flow is input to Risk Treatment Activity which must develop the risk treatment plan. The risk treatment plan flow is then input for Risk treatment implementation activity which should reduce the risk. Treated risk flow is then input for the new Risk evaluation and this is the good example of feedback. Every evaluated risk is input of the Risk Control activity. The output of that activity is the major feedback to the activity Perform SWOT analysis. This is the example of iterative nature of risk management.

Prior to the development of the conceptual risk management model, the process risk evaluation or project risk evaluation was performed sporadically, without prescribed methodology, it wasn't obligatory and wasn't continuously monitored. The aspects, criteria and the criteria description were unclear and it was impossible to calculate risk because for example the impact was evaluated alphanumeric and Probability numeric. It was even possible to propose the project without risk evaluation. In fact, the process of risk management hasn't been really established. Naturally, the subsequent controls were performed, but in fact it is too little and too late.

Considering that problem, using the Conceptual Risk Management Model in sequence as described in Fig 10., the overall unregulated system becomes more stable and organized. The methodology was established and now it is possible to have firm foundation for further development.

Comparing situation with project management before, and now, based on collected risk management data used during the project, the benefits are evident:

- there is prescribed methodology with prescribed feedbacks,
- every project/process must have number of documented information: detailed WBS, RBS, defined Risk treatment rules (Risk model, countable and described variables, aspects, defined criterions), Risk Matrix, Risk Evaluation, Evaluated Risk Lists, Risk Treatment Plan, implementation and Risk Control,
- the main decisions are made by project sponsor,
- having documented information and data it is possible to measure, monitor and control,
- Risk are identified at the beginning of the project in the planning stage and repeated,
- Risk management is iterative process,
- Having developed entity relationship model for the Conceptual Risk Management Model it is possible to develop the risk management software.

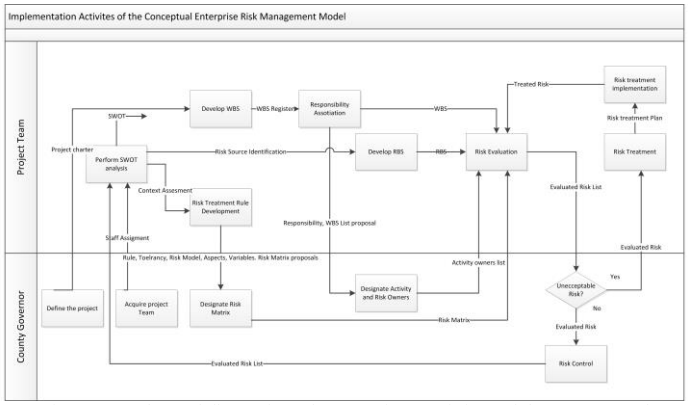


Figure 10. The wokflow of Implementation Activites of the Conceptual Enterprise Risk Management Model

VII. Conclusion

The risk management is becoming more and more important in government agencies. This importance is emphasized by different laws. The problem with that kind of risk management implementation is that it takes into consideration only financial risks, which is not enough for successful risk management and it is not a systematic approach.

Considering all that problems, and available solutions, this paper proposes the Conceptual Enterprise Risk Management Model. The new model includes the data perspective of risk management process, SWOT analysis, work breakdown structure, risk breakdown structure, risk model and risk treatment model with precise relationships which ensures communication between entities. The engineering method is used to develop the new model.

The new model meets all necessary prerequisites (overall results of methodology are represented with a single metamodel, the conclusive set of methods is presented with precise order of implementation [11]) to be considered as a new method.

To verify this new model, it has been implemented in Varazdin County's Quality management system. At the same time, that implementation fulfills all the law requirements and requirements of the new ISO 9001:2015 standard.

The model has been tested in the first stage of the project "Development of broadband infrastructure in the areas of Varazdin County without sufficient commercial interest for investment". Using the new model, some evident benefits have been noticed. The risks were identified and evaluated at the beginning of the project. The model reduces "unpleasant surprises" and costs and, at the same time, supports better communication, control and transparency. Consequently, the chances for achieving project goals are higher.

References

[1] ... Zakon o sustavu unutarnjih financijskih kontrola u javnom sektoru [Internet]. [Cited 2015 apr 14]; Available from: http://narodne-novine.nn.hr/clanci/sluzbeni/128931.html

[2] ... HRN ISO 31000:2012, Upravljanje rizicima- Načela i smjenrice [Internet]. 2012 jun [cited 2016 feb 7]; Available from: http://31.45.242.218/HZN/Todb.nsf/wFrameset2?OpenFrameSet&Frame=Down&Src=%2FHZN%2FTodb.nsf%2Fcd07510acb630f47c1256d2c006ec863%2F76489495956a70c1c1257a940041bd08%3FOpenDocument%26AutoFramed

[3] Ali Sunyaev MH. Method Engineering: A Formal Description. 2010; 645–654.

[4] Bešker M, Ćorić I, Gašpar D. Risk Assessment and Risk Treatment using MRA Methodology and „ERM Venture“ software tool [Internet]. 2015 [cited 2016 feb 6]; Available from: http://erm-venture.com/wp-content/uploads/2015/03/ERM-Venture.pdf

[5] COSO. Enterprise Risk Management- Integrated Framework: Executive Summary [Internet]. Committee of Sponsoring Organizations of the Treadway Commission; 2004. Available from: http://www.aicpa.org

[6] Dobrović Ž, Penić S. Managing the Private Higher Learning Institutions with the Support of Information Technology: Strategic Framework [Internet]. 2008 [Cited 2016 feb 7]. Available from: http://bib.irb.hr/prikazi-rad?rad=488514

[7] Hardy K. Managing Risk in Government: An Introduction to Enteprise Risk Management [Internet]. 2010 [cited 2016 jan 9]; Available from: https://www.rims.org/resources/ERM/Documents/Risk%20in%20Government.pdf

[8] Miloš Sprčić D, Tekavčić M, Šević Ž. CORPORATE RISK MANAGEMENT PRACTICES IN CROATIAN COMPANIES. Ekon. Pregl. 2008 jul 22;59(7-8):344–369.

[9] Rossi M, Ramesh B, Lyytinen K, Tolvanen J-P. Managing Evolutionary Method Engineering by Method Rationale. J. Assoc. Inf. Syst. 2004 sep 9;5(9):356–391.

[10] Vojković G, Cerin B. Upravljanje rizicima IT projekata u drzavnoj i javnoj upravi [Internet]. U: MIPRO 2012, Proceedings of the 35. International Convention. Rijeka: MIPRO - Croatian Society for Information and Communication Technology, Electronics and Microelectronics; 2012 [cited 2016 jan 9]. p. 1899–1903.Available from: https://bib.irb.hr/datoteka/582629.Upravljanje_rizicima_IT-projekata_u_drzavnoj_i_javnoj_upravi.pdf

[11] Željko Dobrović. INFORMATION SYSTEM'S DESIGN AND DEVELOPMENT METHODS FOR OBJECT SYSTEMS WITH TIME CHANGEABLE FUNCTIONS. 1998;

[12] Zakon o javnoj nabavi - Zakon.hr [Internet]. [Cited 2016 feb 5]; Available from: http://www.zakon.hr/z/223/Zakon-o-javnoj-nabavi

[13] What is a Work Breakdown Structure - Workbreakdownstructure.com [Internet]. [Cited 2016 feb 7]; Available from: http://www.workbreakdownstructure.com/

[14] Organizing Risks with a Risk Breakdown Structure (RBS) [Internet]. [Cited 2016 feb 7]; Available from: http://www.brighthubpm.com/risk-management/51997-using-a-risk-breakdown-structure/

Electronic Commerce in Croatia and a Comparison of Open Source Tools for the Development of Electronic Commerce

J. Tomljanović*, T. Turina**, E. Krelja Kurelović***
Polytechnic of Rijeka
Vukovarska 58, Rijeka, Croatia
* jasminka.tomljanovic@veleri.hr
** tturina@veleri.hr
*** elena@veleri.hr

Abstract – Electronic commerce (also known as e-commerce) has experienced a constant growth and development in recent years. Thus, customers and sellers are turning to this type of business in hope that it improves their sales. In order to create such a website, it is important to have proper tools and skills. A person may opt for two options and these are: commercial or open source. The objective of this study is to explore the selected open source tools for creating an e-commerce site. The purpose and goal is to create an e-commerce site using all of these tools that will be compared in order to make a comparative analysis of the offered possibilities. The following three tools were selected: AbanteCart, PrestaShop and OsCommerce. The study describes in short the concept of e-Business and e-commerce, examines the growth and the development of e-commerce in Croatia, shows the most visited domestic e-commerce sites, and in more detail analyzes the possibilities of each of the tools. Finally, a comparison of the tools has been made and the conclusion about which of these tools is the most suitable for creating the best initial e-commerce site has been drawn.

I. INTRODUCTION – ELECTRONIC BUSINESS

Electronic business is a new way of doing business, which offers a lot of special benefits and is a business system or set of actions that seeks to take full advantage of the progress of information and communication technologies to improve business processes. Electronic business is any business that relies on the information system. The information system can be thought of as a database that is accessible on local or global networks. The success of electronic business depends on the correct application of the procedures. Basically, there has to be a good strategy to consistently execute procedures and have logical and functional structure. Secondary factors relate to the characteristics of an individual such as talent, creativity and leadership. [8] The results of the application of electronic business enable the company to take its place on the world's largest market, the Internet, on time and in the right way.

Theory of electronic business recongnizes a few basic business models. They are defined based on the parties involved in the business process and the nature of their business relationship. A generally accepted theoretical division of business models includes: [8]

1. Electronic business between businesses (B2B - Eng. Business to Business), which includes business cooperation among companies that share resources in the form of products, services or information.
2. Electronic business between companies and clients (B2C - Eng. Business to Consumer) is a direct business cooperation between the company and the client, where the client company sells products, services or information.
3. Electronic business between customer (C2C - Eng. Consumer to Consumer) means business cooperation among customers who share resources in the form of products, services or information. This model is realized via an online broker.
4. Electronic business between businesses and employees (B2E - Eng. Business to Employee) is a system that defines procedures within the company concerning the relationship of the company and employees.
5. The electronic service of the state administration (Eng. eGovernment) is a form of electronic business between the state administration and clients. Clients of the state administration can be citizens, companies and other bodies of state administration.

6. B2B2C is a newer model that combines the use of B2B model that supports the company's business model acording to the B2C model.
7. C2B2C involves performing transactions between consumers using online company as a broker.

II. ELECTRONIC COMMERCE

The definition of electronic commerce is not unique since different sources give different definitions. Panian has so far offered the best definition that describes electronic commerce and he defines it as the process of buying, selling or exchanging products, services or information through publicly accessible computer network, the Internet, and offers reduced costs and time of transactions. According to Watson, activities resemble electronic trading activities which are carried out during the on-line sale of own goods and services. [2]

Electronic retailer mainly produces nothing, but offers consumers the services or products that have already been produced by another. In electronic commerce one can trade all that someone has produced or will produce, therefore, electronic traders offer to sell goods and services from a number of sources, often competing, but the relations with suppliers are significantly different from the relations of other sellers who produce their own goods. Ecommerce websites will have to support a wider variety of types of communication than previous, with such a level of functionality that a trader can respond quickly to changes in the market of sales and market purchases. Electronic retailers create electronic market and, therefore, their concern is focused on the maintenance and support of one or more markets, which is not the case with the vendor's own products or services. [6]

The concept of electronic commerce can be seen from these four perspectives: [1], [5]

1. Communication perspective: electronic commerce enables the delivery of information, products, services or means of payment through a public telephone line, publicly accessible computer networks or other electronic means.
2. Business processes perspective: electronic commerce is the application of new technologies to automate business transactions and business improvement.
3. Perspective of service delivery: electronic commerce is the means by which companies, users and management are trying to reduce the cost of service delivery while increasing the level of quality of goods and increase the speed and their delivery.
4. Virtual perspective: electronic commerce offers the possibility of buying and selling of products, services and information through the Internet and its services.

Electronic commerce refers only to the trading activities on the Internet and can not be equated with the broader concept of electronic business related to intensive development of all business activities electronically. [1].

A. Electronic commerce development

Electronic commerce has expanded in the early nineties and today the Internet is an essential channel of commercial transactions as well as the global market of hundreds of millions of buyers and sellers, and a place for all kinds of transactions and activities. Websites are becoming more intuitive by nature; they are managed by platforms for collaboration and communication that exceed one-way communication flow. Faster and cheaper communication is at the center with end users, which are then offered special forms for ordering products that electronically initiate business events. The use of advanced web tools and techniques irreversibly changed relatively outdated websites and electronic commerce sites, and affects the development of new sales channels and the expansion of relationships with end users. New technologies have led many traders to set up or move their business to the Internet in hopes of becoming wealthy quickly. The strategy of electronic commerce must be appropriate to the industry, innovative and needs to create competitive advantages and new values. The speed and extent of acceptance of electronic commerce depends on the country and sector in which it is applied. [1]

B. Payment systems in elecrtonic commerce

Today's development of electronic commerce has created a need for new money - electronic money and a new payment method, i.e. electronic payment because the paper money is inappropriate with electronic trading. Electronic payment is a separate part of electronic commerce, and there are four main methods of payment in electronic commerce: [4]

- Debit and credit cards
- Electronic (digital) money
- Smart cards and electronic wallet
- Micropayments

A credit card is an instrument of cashless payment transactions and means of short-term lending and has a transactional function. Payment by credit card is the most common form of payment in electronic commerce. [4]

Smart cards allow users to store electronic money in memory-processor chips on the actual cards in the form of digital messages, and in the function of electronic money or information for the digital verification with electronic payments. Functionally similar, but when it comes to performance, very different on-line payments technologies, are so called electronic wallets. Their software is programmed in the retailer computer server, allowing users to track information on commitments and the deliveries of goods. Due to the fact that merchants have to pay a fee for each transaction made with credit or debit card,

when it comes to small amounts of payment, it could happen that a fee exceeds an amount. To avoid such losses electronic retailers can offer customers a micropayments system for amounts less than 10 US dollars. [1]

C. *Electronic commerce in Croatia*

According to the Index of development of the digital economy and society, which was recently published by the European Commission, Croatia occupies 24th place in the digital and technical development, while it is in third place in Europe with 25 per cent in the number of small and medium-sized companies engaged in Internet commerce. Regarding the use of the Internet, Croatia occupies 21st place with 65 per cent of the active population, while the EU average is 75 percent. In the digital skills Croatia is below average with 39 per cent and the EU average is 60 percent. [7]

If Croatia is compared with its neighboring countries, the Internet breakthrough is very similar to that in Bosnia and Herzegovina - 63%, Serbia - 65%, Macedonia - 67%. Although a neighboring country, Slovenia is far ahead of us in the number of Internet users – there almost 80% of the population over 15 years uses the Internet, in Slovakia - 70%, in Hungary 74%, while, for example, in Russia, the Internet uses a smaller part of the population - 58%. Internet use varies significantly considering the age of the population: among the youngest (15-24 years), only 3% of them do not use the Internet, while at the age of 25-34 years only 7%. With the increase in the number of years, the number of Internet users significantly decreased. Thus, among the oldest (over 65 years), there are only 10% of those who use the Internet. [11]

According to research conducted in 2014, Croatia has continued its growth by 3% a year, although in comparison with the neighboring country of Slovenia, it is still lagging behind. A detailed outline can be seen in figure I where darker colors indicate the number of Internet users and brighter color indicate the number of Internet buyers. [9], [12]

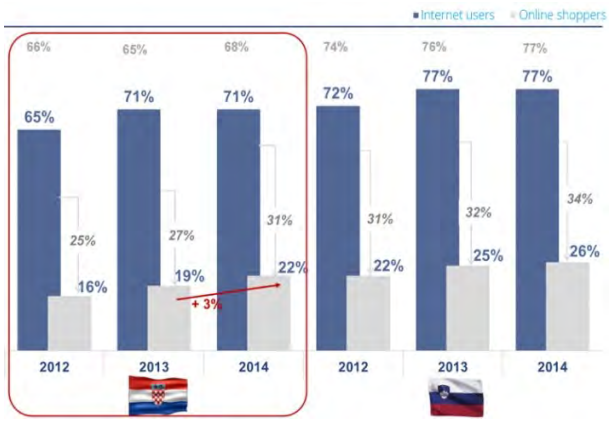


Figure I: The growth of online shopping in Croatia and Slovenia

III. THE MOST VISITED DOMESTIC E-COMMERCE SHOPS

At the conference Web trgovac (in Eng. Web Retailer), which was organized by the leading portal for price comparison Jeftinije.hr in collaboration with the research house Valicon, the best Croatian web shops were proclaimed. The company eKupi.hr was declared the best web merchant of the year , and it also also won the in the category Internet shopping mall. The other winners were announced in eight categories, that are: [10]

- printink.hr in the category Customer reviews,
- Sonusart.hr in the category Audio&Video,
- Silux.hr in the category Auto&Moto,
- Emmezeta.hr in the categorx Home &Garden,
- Gizzmo.hr in the category IT,
- Alternativa-webshop.com in the category Beauty, and
- Dobra-macka.com in the category fashion.

The main criteria for the announcement of the winners was the opinion of customers collected by research based on Internet surveys that Valicon conducted on more than 29,000 Croatian Internet users who made a purchase in at least one domestic online store in the last year. [10]

IV. THE TOOLS FOR THE DEVELOPMENT OF ELECTRONIC COMMERCE SITE

To install electronic commerce tools, the required specifications are:

- a domain,
- a web server: Apache1.3, Apache 2.x, Microsoft IIS and similar,
- PHP 5.2+,
- MySQL 5.0+ with a created database
- SSh (engl. *SecureShell*) ili FTP access

Three open source tools were selected for comparison: AbanteCart, Prestashop and OsCommerce. Each installed on a separate web server with all the required specifications.

A. *AbanteCart*

After you log into the user interface, a simple outline of a dashboard that offers a general overview of sales, products, etc. is given and a statistical display of sales. The main menu consists of a few options, and these are the following:

- The catalog – which manages products, product categories, manufacturers, reviews and other attributes,
- Sales – which manages orders, customers and user groups and coupons,

- Design – which manages store appearance and themes for decorating shops looks,
- Extensions – which manages additional modules for the store,
- System – which manages the system and localization system,
- Reports – in which is the display of the sale reports , and
- Help – where there is AbanteCart's documentation.

B. *PrestaShop*

When logging into the PrestaShop's administrator page, the dashboard is displayed the dashboard, which shows statistical information about the website, such as: the number of visitors, orders, the value of the basket, the total profits of the page and etc. The tools menu consists of the options:

- The catalog – which manages products. It offeres the following options: products, categories, supervision, product attributes, product features, manufacturers, suppliers, tags and attachments,
- Orders – which manages orders. The administrator is offered the option to print an order, see an account, see delivery or perform a partial refund to.the customer,
- Customers – where it manages customers. It offers search options over customers, editing and adding customers as well as managing addresses, manage groups of customers and support customers where one can respond to customer inquiries,
- Basket rules – which manages the basket, adding the rules that apply when ordering, catalog price rules where you can edit the discounts and benefits that can then be applied to the products,
- The modules – which control modules and a catalog of modules. The administrator can see all the modules that are installed, information about which modules are disabled, and what modules can be updated,
- Delivery - which manages suppliers and shipping settings, where it is possible to place the cost of operation, determine the conditions of free shipping, and default posting suppliers,
- Localization - which manages language settings. This option offers adding another language to a store to support multilingualism and it is possible to add, edit and delete those languages. Other options localization offeres are zones, state, county, currencies, taxes, tax rules and translations where you can add, edit or delete all the items listed,
- Settings - which controls the settings of orders, products, customers, topics etc.,
- Advanced settings - which manages the configuration of the site,
- Administration - which regulates the General Administration settings, notification of orders, customers and messages, employees, profiles that can be allocated etc., and
- Statistics – which gives access to all the statistical data of the site.

C. *OsCommerce*

After the administrator first logs in the user interface on display is a graphical representation of earnings and rating per site. It also offers a view of orders, clients, displays the administrator login, the last news related to the tool and the latest additions. The main menu consists of the following options:

- The catalog – which manages product categories, products and their attributes, and suppliers,
- Configuration – which manages administrators, trade preferences, memory, backup, products and delivery settings,
- Customers –which manages customers,
- Localization – which manages currencies, languages and order status,
- Location / Taxes – which manages countries, zones and taxes,
- The modules – which manages the content within installed modules. The most important modules in this option are additions for the payment and delivery methods. In this option one can adjust all modules that are going be active on the control panel,
- Orders – which manages orders. Under this option there is only a list of the order left in the system. The statuses of the orders can be altered here,
- Reports – these manage reports of the orders made, reports of purchased products and reports for all products that have ever been in the system, and
- Tools – manages a variety of other settings tools

V. AN ANALYSIS AND A COMPARISON OF THE TOOLS

There are three factors when evaluating the quality of electronic commerce website [3] and they are closely represented in Table I:

1. Usability – it refers to characteristics that enable electronic commerce to be useful in a variety of situations, not only through its own development process, but the use and maintenance, and is based on the reliability of

the web site and consists of the following two aspects;

2. Conceptual reliability – it deals with the possibilities of electronic commerce to satisfactorily implement what was specified and designed.
3. Representative reliability – it refers to the representative possibilities of electronic commerce that affect the understanding and handling of the store through its life cycle.

Table I shows the main three factors and their sub-factors that more closely describe the functionality and a brief description of all the sub-factors is offered. According to research conducted in 2010 in University of Jordan[3], each of these sub-factors of quality has its own attributes that further assist in assessing the quality of websites. These factors are listed in Table II where, for every factor there are corresponding sub-factor and a group of attributes for each sub-factor.

TABLE I: THE MAIN FACTORS AND SUB-FACTORS WHEN EVALUATING THE WEBSITE QUALITY

Usability	
Efficiency	The timeliness of which the website responds to the user.
User-friendliness	The user interface capabilities to which the website provides a supportive experience to the user.
Navigability	The browsing extensibility which the website's software allows.
Maintainability	The reduced effort which the website's software requires for it's upkeep, enhancing its ability to be kept up to date and usable.
Involvement capacity	The measure of which the website can adapt and attract each user's individuality.
Conceptual reliability	
Functionality	The extent of the operational aspecs of the website software and its fitness of use.
Security	The extent of safety assured against malicious or accidental intrusion of unauthorized users when using the website.
Reliability	The extent of which the website remains available and working.
Integrity	The reliability, consistency and correctness of stored data.
Trustworthiness	The extent to which the user percieves the website to behave consistently, reliability and correctly, building a trusting relationship.
Content adequacy	The extent to which the information presented is contextually applicable to the user and sufficient for the user's needs
Scalability	The websites readiness to meet rising demands in user and usage.
Availability	The extent of website accesability to users through different browsers in different times.
Representation reliability	
Readability	The appropriate application of the written language within the website.
Standards conformance	The extent of consistency applied within the user interface of the website.
Ease of manipulation	The extent of help provided to operate the website and the software underneath it.

Using the above factors a detailed analysis of tools can be carried out and with data processing, it is then possible to determine which of those tools that have been compared is the most appropriate one to use. A comparison of the tools is based on the existence of the above factors and on their proper functioning in order to provide the users with maximum support and the use of the tools. An online store was built in each of the tools as a basis for an assessment. Each feature was used and tested through the building process. The online store was tested throug the user point of view as well as the administrative point of view, including registration, site navigation and the simplicity of placing an order for a specific product.

Final points of the subfactors and their attributes are shown in Table III, for each of the tools.

TABLE II: WEBSITE QUALITY SUB-FACTORS OF AND THEIR ATTRIBUTES

USABILITY	
Sub-factor	Attributes
Efficiency	Purchase process performance, Page generation speed
User-friendliness	Understandability, Product information availability, Interactivity, Learn-ability, Response time uniformity, Forms of payment availability, Product comparison, "Shopping cart" metaphor, Printing facilities, Download facilities
Navigability	Absence of navigation errors, Minimal path and shortcut facility, Navigation structure taxonomy, Link visibility, Link visualisation consistence, Alternative paths, Navigational prediction, User level adaptability, Mobile devices accesability
Maintainability	Stability, Testability, Analyzability, Changeability
Involvement capacity	Attractiveness, Aestetic attributes, Client profile identification, Simulation, Additional services availability
CONCEPTUAL RELIABILITY	
Sub-factor	Attributes
Functionality	Accuracy, Client support, Information on product delivery, Suitability, Flexibility
Security	Payment systems security, Vulnerability, Privacy, Access control, Confidentiality, Site authentication
Reliability	Recoverability, Maturity, Fault tolerance
Integrity	Data integrity, Data entry signalizing, Robustness
Trustworthiness	Correctness, Completeness, Necessity
Content adequacy	Updated content, Correctness, Intelligibility, Compatibility with real store, Concise content, Completeness, User oriented
Scalability	Multiprocessor handling, Farming capabilities
Availability	24/7/365 readiness, Partial availability, Browser version compatobility, Cross Browser Support
REPRESENTATION RELIABILITY	
Sub-factor	Attributes
Readability	Language correctness, Style uniformity, Clarity, Conciseness, Terminology uniformity
Standards conformance	Interface standards, Programming standards, Navigation standards
Ease of manipulation	Up-To-Date, Ability to trace

TABLE III: OVERVIEW OF THE POINTS OF THE WEBSITE QUALITY SUB-FACTORS

Factors and sub-factors	AC	PS	OsCom
USABILITY			
Efficiency = 10	9	8	8
User-friendliness = 50	45	44	39
Navigability = 45	40	36	36
Maintainability = 20	18	16	12
Involvement capacity = 25	21	23	12
CONCEPTUAL RELIABILITY			
Functionality = 25	22	23	19
Security = 30	26	26	23
Reliability = 15	14	12	11
Integrity = 15	14	13	13
Trustworthiness = 15	14	13	10
Content adequacy = 35	33	32	26
Scalability = 10	9	9	7
Availability = 20	16	15	13
REPRESENTATION RELIABILITY			
Readability = 25	22	22	16
Standards conformance = 15	13	14	11
Ease of manipulation= 10	9	9	9
TOTAL = 365	325	314	265

A point from 1 to 5 was given for each of the attributes, 1 meaning poor performance and 5 meaning full performance. Points were then summed and a final point was created. The points were given based on a peronal experience when using and testing of the tools. It is safe to conclude that the worst-placed tool is OsCommerce due to inadequate upgrades through versions. As for the validity of the tools, it performs and executes what it is designed to do, but any attempt to adapt the tools to the modern times, causes errors in the system. Contrary to this tool, there are two newer tools; PrestaShop who took the second place and AbanteCart who received the most points in the table. These tools are similar in functionality, but AbanteCart is more intuitive, it is easier to make the site and it provides administrative user easy navigation through the system. All ratings are based on personal experience when using the tools and testing their capabilities.

VI. CONCLUSION

Electronic commerce is becoming more widespread notion in goods and services trade. In Croatia, there is a growing demand for electronic shops and more customers are turning to the Internet therefore more and more stores direct their business accordingly. A list of the best rated electronic commerce, which are evaluated by the users by research initiated in 2014., is the best evidence of the popularity of electronic commerce in Croatia.

By creating an electronic commerce shop in all the three tools and comparing these, it was concluded that the tools AbanteCart and PrestaShop are better than the outdated OsCommerce, which results from insufficient upgrade integrations. That makes it a weaker option in relation to the more modern competitors. What is inadequte in all the three tolls is the lack of possibility to adapt to the Croatian language. AbanteCart in its additions has no options for the use of the Croatian language, PrestaShop supports it although some terms are incorrectly translated and incorporated, while OsCommerce supports additions for language, but an administrator has to search a large number of pages to find one, and even if the option is found it can be defective and can disrupt the functioning of the tool. Furthermore, one of the drawbacks common to all the tools is the inability to adjust to the user's preferences, and personalize the sites to the user's preferences. The tools do not have the option to memorise unique user visits and retain specific products or product categories so that the tool can intuitively offer similar products on their next visit.

REFERENCES

[1] Babić R., et.al., Dosezi elektroničke trgovine u Hrvatskoj i svijetu, Sveučilište u Zadru, Zadar, 2011.

[2] Panian, Ž.,. Izazovi elektroničkog poslovanja, Narodne novine d.d., Zagreb,2002

[3] Rababah, O. M., Masoud, F. A., KeyFactors for Developing a Succesful E-commerceWebsite, The University of Jordan, IBIMA Publishing, Jordan, 2010

[4] Ružić, D., E-Marketing, Ekonomski fakultet u Osijeku, Osijek, 2003

[5] Spremić, M., Menadžment i elektroničko poslovanje, Narodne novine d.d., Zagreb, 2004

[6] Srića V., Muller J., Put k elektroničkom poslovanju, Sinergija, Zagreb, 2001

[7] Bačelić Grgić, I., Hrvatska na trećemu mjestu u Europi u internetskoj trgovini, HGK, http://www.hgk.hr/djelatnost/gosp_trgovina/konferencija-povodom-obiljezavanja-svjetskog-dana-zastite-potrosaca-na-temu-internet-trgovine-i-potrosaca

[8] Gavrilović B., Osnovni modeli elektronskog poslovanja, Kompjuter biblioeka, http://vesti.kombib.rs/osnovni_modeli_elektronskog_poslovanja.html

[9] IzBizCrunch.com, Ima li smisla otvarati Internet trgovine u Hrvatskoj?, http://www.itbizcrunch.com/index.php/analize/item/225-ima-li-smisla-otvarati-online-trgovine-u-hrvatskoj

[10] Jeftinije.hr, Najbolje hrvatske web trgovine, http://www.jeftinije.hr/Novost/2798/proglasene-najbolje-hrvatske-web-trgovine

[11] Progressive.com, U RH sve više korisnika Interneta kupuje online, http://progressive.com.hr/demo/4817-gfk-u-rh-sve-vie-korisnika-interneta-kupuje-online.html

[12] StaffEcommerce News, 76% of Croatian on-ine buyers shopa broad, http://ecommercenews.eu/76-of-croatian-online-buyers-shop-abroad/

The social marketing as prerequisite for the competitiveness of South-East European companies

Iva Bubanja, Mr.sc.
Belgrade Business School, Belgrade, Serbia
E-mail: ivabubanja@gmail.com,

Abstract - The decreasing costs of network access, as well as, the growth of its accessibility to individuals raised the internet popularity. Internet and cyber market are redefining the relationship between buyers and sellers, enabling the development of interactive relations between companies and customers. This paper examines the use of social marketing by SEE companies in order to find out whether they are aware of all the benefits offered by social media for their further development and for making their products and services more familiar to potential customers. In the focus of analyses are social marketing players, social marketing process, as well as, available social media statistics of South East European companies.

I. INTRODUCTION

Information - communications technologies (ICT) and the Internet are in the center of technological changes that affect contemporary world economy, causing a massive change in the economic system and affecting the work of international companies and domestic enterprises. ICT have the characteristics of general purpose technologies or enabling technologies because they are widely used (or embedded) in different domains, and have a profound effect on the entire economy.

All countries, weather developed or developing are trying to tap advantages offered by technological progress based on ICT in business innovations, improved products, more efficient processes and new organizational structures. The final expected result is higher economic growth rates and better life for all citizens [1].

The new business environment has become a challenge for governments and companies in developed countries, because they have to adapt and rapid accept the new tools and technologies with the objective of maintaining market position and distancing its competitors. For developing countries and transition economies, such as South East European countries (SEE), these adjustments tasks are even greater. In parallel with the process of transition and economic reforms for revitalization of the domestic economy and service resources, companies should also include ICT and they need to introduce modern methods of online communication and work in their management processes.

For this analysis, SEE covers Croatia, FYR Macedonia, Montenegro, and Serbia. Despite their many differences, these countries share certain common characteristics: they are all transition countries and except for Republic of Croatia who has become full EU member state since 2013, all other analyzed countries have been taking part in an ongoing process of the enlargement of the European Union.

II. ICT AND E-BUSINESS FOR A SUSTAINABLE ECONOMY IN SEE COUNTRIES

After the Euro zone sovereign debt crisis deteriorated in the summer of 2011, growth prospects have again weakened and the economic outlook has worsened for SEE countries in transition. As a consequence, SEE governments have been faced with stagnant or decreasing growth rates and rising unemployment. Exports and capital inflows declined, causing difficult fiscal challenges and rising levels of public debt [2]. It was evident that SEE countries need the changed growth model based on ICT potential that can enable further reform, modernization and innovation in order to boost growth rates towards their long-term potential.

In order to face decreasing production and export challenges caused by inside or outside factors, SEE companies can use ICTs to reduce business costs, improve internal management, and expand access to new technologies and information on market opportunities. ICTs can also help optimize supply chains, making it easier to get goods and services to the market.

Empirical evidence shows that, on average, firms that use ICTs grow faster, invest more, and are more productive and profitable than those that do not. According to a World Bank study, sales growth and profitability are 3.4 and 5.1 percentage points higher, respectively, among firms that use ICTs effectively in their businesses [3].

Companies in SEE counties can use ICT and e-business for two basic objectives: to reduce costs, and to increase their markets and revenues. The focus and priorities differs between sectors, but can also change depending on the company strategy and the business environment. Many companies currently focus on the cost-cutting potential of ICT (efficient processes, e-procurement), due to the difficult economic environment. When the economy recovers, the strategic focus could shift towards marketing and growth objectives such as optimizing the supply chain, improving products and customer service, reaching new customers or increasing the market share [1]. The focus of this analysis is how

new ICT technologies can improve marketing activities of order to make them more competitive on European and world market.

III. INTERNET AS A PLACE OF MODERN DEMAND

With the decreasing costs of network access and increase of its availability to individuals, the popularity of the Internet increase and it is seen as a place where contemporary customers can freely seek for information or implement a business transaction. At the very beginning of the commercial use, in 1995, Internet was used globally by about 14 million people. In 2000 there were about 400 million users, and in the 2015, the network has 3.2 billion users [4]. At European level, which is for the SEE companies one of the most important markets, the number of Internet users grew from 105 million in 2000 to 507 million in 2015 (Table 1)[4].

Constant rise of network users on a global level proves Internet as a market with huge potential demand that enterprises cannot ignore especially given the forecasts that by 2020 the Internet will have about 4 billion users in the world [5].

TABLE I. INTERNET USERS AND INTERNET PENETRATION RATES

World regions	Internet users in million		Internet penetration rate	
	2010.	2015.	2010.	2015.
World	**2.054.2**	**3.079.3**	**30,0**	**42,4**
Europe	475.1	582.4	58,4	79,4
Croatia	2,244	2,780	54	70
Macedonia	1,057	1,408	52	70
Montenegro	0,280	0,364	37,5	67,2
Serbia	2,900	4,705	40,9	65,8

Source: [4].

Use of Internet in SEE countries can be commented in accordance with targets stetting by Digital Agenda for Europe. With regard to internet usage the Digital Agenda sets the target to increase regular internet usage from 60% to 75% by 2015, and from 41% to 60% among disadvantaged people [6]. In all analyzed countries Internet usage by individuals is close to 70%, but still below EU average penetration of 79 in 2015 [4].

IV. USE OF INTERNET AND ICT IN SEE COMPANIES

There is a rising awareness among SEE companies that ICTs and the Internet are a fundamental economic infrastructure. The benefits of ICTs are amplified by their proper use in companies, and the innovations that they drive. ICT investments spur competitiveness and productivity at the firm level, in particular when combined with investment in skills, organizational change (and industry restructuring), innovation and new firm creation [7].

Usage of ICT is an important part of contemporary business conduct in SEE companies according to statistics (Table 2). Nearly all enterprises in Serbia and Montenegro with at least 10 persons employed had internet access in 2015. Percentages are little lower in Croatia (90) and Macedonia (93). Many enterprises may consider it important to be visible on the internet and close to three quarters (75%) of enterprises in SEE region had a website in 2015. The percentage of enterprises having web site is something lower in Macedonia (53) [8].

TABLE II. USE OF INTERNET

	HR	MK	ME	RS	EU28
Enterprises with Internet access					
2010	95	84	95	96,8	94
2014	96	93	98	100	97
2015	90	93	99	99,1	97
Enterprises with WEB sites					
2010	61	43	-	67,5	67
2014	66	53	*73.3*	74	74
2015	71	52	74,3	75,2	75

Legend: HR-Croatia, MK-Macedonia, ME-Montenegro, SR-Serbia. **Source:** [8] and for Serbia [9] and [10].

Despite having WEB sites, SEE companies still are modest in use of Internet for purchasing or selling online. Around a quarter of companies in Croatia, Serbia and Montenegro receive orders online, and Serbia is leader concerning the percentage of companies having purchased online in 2014 (40, 3%) [8].

Current technological progress like WEB 2.0 services, virtualization and cloud computing, dynamic rise of mobile technology and broadband connections create new opportunities that SEE companies cannot ignore if they want to increase their competitiveness and position their products and services on European and world market.

V. SOCIAL MEDIA AND SOCIAL NETWORKS IN SEE COMPANIES

The recent explosion in social networking and the related evolution of new forms of business, operational, scientific, and other relationships point in even more promising directions and face SEE companies with new challenges.

Social networking and other forms of user-generated content are the most dynamic area of Internet development today, at least as far as end-users are concerned. The changes in the relationship between the production and consumption of content which are associated with them have already had profound implications. Web 2.0 has enabled a widespread and continuing shift in the balance of communications and content consumption away from a more traditional pattern (such as newspapers, broadcasting and mainstream websites), in which content was delivered to consumers, towards a new pattern in which a high proportion of content is created by consumers and exchanged among consumers [11].

A new dynamic to the growth of social networks is added by the increasing availability of Internet, and so of Web 2.0 interaction on mobile phones. User-generated content takes many forms, such as: online chat and instant messaging (IM) services, voice over Internet services (VoI), audio and video sharing websites, blogs, wiki-based sites and wiki software, other important areas of user-and network-generated content include personal sales and auction sites, online gaming and peer-to-peer file sharing [12].

Over 3.17 billion people used social media networks across the world in 2015. Some of the most popular

networks are: Twitter with 290 million active users, followed with 300 million Instagram users, Linkedin monitors 347 million people, while Facebook is still dominant social network with over 1.49 billion active users and it takes over 70% of the traffic on social networks (Figure 1) [13].

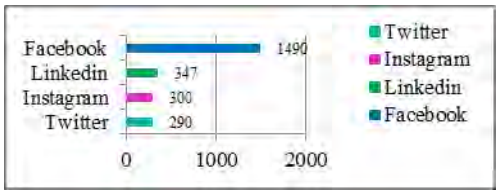


Figure 1. Social media users in million 2015 (worldwide)

Social penetration rate in 2015 was 26% at the global level, and a typical user of social networks, the potential buyer, spend about 2.5 hours a day using social networks. As the age category of people using social media is widening, applicability for a much broader product range, and thus a broader range of companies, becomes of interest. Therefore, it is important for companies to become visible on social media and portals [14].

Having this in mind, companies in SEE countries should pay more attention to the use of social internet media in order to enable their influence on business processes and better business results. Analyzed data show that companies form this region are still modest users of social networks. Statistics are following for 2015 (Figure 2):

1. Croatia: 64% of internet users were participating in social networks. Social internet media usage was not widespread and was used by 38% of enterprises [15];
2. Montenegro: 76% of internet users were participating in social networks. Enterprises used social networks (43,2%) and had blogs and micro blog (13,6%) [16];
3. Macedonia: 83,6% of internet users were participating in social networks. Half (49.8%) of the enterprises used social media. i.e. had a user profile, an account or a user license for using certain social media. Around 40.2% had links or references to their social media profiles [17];
4. Serbia: 75,6% of internet users were participating in social networks. Enterprises used social networks (28,6%) and had blogs (8,8%) [10].

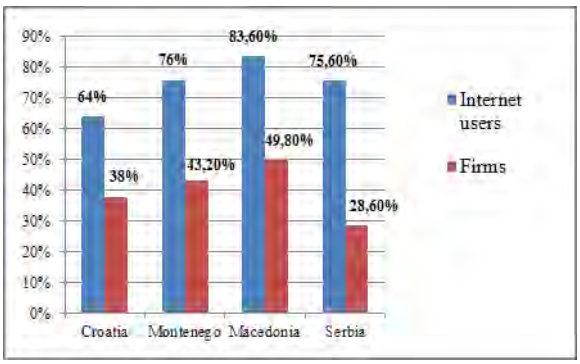


Figure 2. Internet users and firms participating in social networks in 2015 (%)

This growing global trend of social media can help SEE companies to expand their business and better position their product and services on the market what should not be ignored.

VI. WHY ARE SOCIAL MEDIA AND SOCIAL NETWORKS IMPORTANT FOR SEE COMPANIES

Only with the presence on social networks and with the use of Web 2.0 internet services, SEE companies can properly present their offer, make contacts with their buyers and reach required competitiveness level on the global market. Social media platforms allow companies to perform various business functions through social media, including marketing, market research, and online retailing. As a result, more companies start to use them in their marketing strategies.

New customers on an international market will buy a company's services or products only if they are aware of their existence. Setting up a company account at a social networking platform and using it to perform marketing functions is usually free of charge and allows company to disseminate messages to large audiences against low costs. Moreover, social media enables new forms of customer segmentation and targeting. Traditional marketing instruments (e.g. newspapers, television and radio ads) are usually country specific, while social media facilitates targeting are based on parameters that are supranational (followed topics, likes, or page memberships). This results in more effective marketing campaigns, only reaching out to customers that are actually interested [18].

Apart from this business-to-consumer (B2C) communication, social networks also facilitate consumer-to-consumer (C2C) communication, which can be compared to what, in traditional marketing, is called word-of-mouth. However, by instantly leveraging a consumer's online network (e.g. when a user shares a company message/video), the potential speed and outreach is much greater.

Social networks also perform a market-research function. Consumer-to-business (C2B) and C2C communication provides a company with valuable insights into an international market's response to certain products and services. This information is freely accessible and can be gathered against relatively low cost. It allows a company to tailor its marketing message and products to better fit customer perception/preference.

Finally, latest technological developments also allow social media to be used as actual sales channels (as one of the case companies illustrates). Social media platforms like Facebook (enabled through certain technology) can perform a similar function as a web shop, allowing customers to directly purchase products and services based on what they read or see.

These three different business functions (or channels) that social media perform are valuable instruments for accessing new international markets, especially considering the non-proximity requirement that is associated with using social media.

The fact that most social-media users have multiple accounts at different social networks, and use those networks for a variety of purposes, implies that SEE companies have to carefully pick the network they want to use.

VII. POTENTIALS OF SOCIAL MEDIA

A concept of social media is connected to the idea of allowing many internet users to share, collaborate, and update the web content. This ideology is rooted in community – users may engage, collaborate and share with others in real time (in the case of virtual and mobile technologies) without constraint of time or geography. O'Reilly [19] references the internet as a conduit for participation, a consumer-driven medium with global reach and the content becomes more visible and more powerful with volume.

Social marketing is one the fastest growing forms of online marketing. The emphasis in social marketing has shifted from exposing consumers to messages, towards engaging them in conversation about your company, or your products and services. Social marketing and advertising is not simply a new ad channel but collection of technology - based tools for communicating with shoppers.

Social marketing differs markedly from traditional online marketing. The objectives of traditional online marketing are to put your business message in front of as many visitors as possible and hopefully encourage them to come to your web site to buy products and services, or to find out more information. The more impressions you get, and the more unique visitors to your site, the better. Traditional online marketing never expected to listen to customers, much less have a conversation with them, any more than TV advertisers expected to hear from viewers [20].

In social marketing, the objective is to encourage your potential customers to become fans of your company's products and services and engage with your business by entering into a conversation with it. Your further objective is to encourage your business's fans to share their enthusiasm with their friends, and in so doing create a community of fans online. Ultimately, the point is to drive sales, and to do this by increasing your "share of online conversations." There is some reason to believe that social marketing is more cost effective than traditional marketing although this is still being explored [20].

The main benefits of social media strategies for increasing competitiveness include the huge exposure that company can generate, the relatively low costs associated with it, the possibilities for customer segmentation and targeting it provides and the market insight that can be gained from analyzing consumers' online behavior and interacting with them through social networks.

VIII. CONCLUSION

The growing digital market represents a favorable business environment for companies and allows them to carefully make their own online communication strategy in order to improve their business position and become more competitive. Internet provided privileges as a medium of communication and as a channel of transmission of advertising messages. It is linked primarily to the possibility for the company's further development and for their products to become more recognizable to consumers.

Nowadays social networks have a profound impact on how business operate, communicate and serve their customers. The most visible business firm use of social networks is as a marketing tool. The most popular social network tool for large corporate business is Linkedin, with 95% of the Fortune 500 having a Linkedin company page. Pinterest has the fastest growing rate among the Fortune 500, increasing from 9% in 2013 to 36% in 2014 [20].

For these reasons companies in analyzed SEE countries should pay more attention to marketing activities on Internet in order to get closer to their potential customers and present them their products and services. As proved by analyzed statistics Internet users in SEE region use social networks more for private purposes (around 70%) than in business (around 40%). Studies have shown that majority of SEE companies use mostly one of the social networks in their business, which is still a low percentage. In business purposes the most commonly used social network are Facebook and YouTube, followed by LinkedIn, and at least Twitter.

The performed analysis also shows that SEE companies used social media mostly for advertising and exchanging the information with customer via blogs or social networks. They were predominantly used in small enterprises, while medium-sized and large enterprises rarely used social internet media for business purposes.

For SEE companies less visible marketing use of networks is as a powerful listening tool that has strengthened the role of customers and customer feedback systems inside a business.

It is also important to stress that social networks are where corporate reputations are formed and SEE firms today should take very seriously the topic of online reputation as evidenced by network posts, commentary, chat sessions and Likes. In this sense, social network sites become an extension of corporate customer relationship management systems and can be important contribution to creation SEE company's competitiveness level.

REFERENCES

[1] European Commission, e-Business Watch, ICT and business for an Innovative and Sustainable Economy, Brussels, 2010.

[2] EBRD, Transition Report 2014, London, 2014.

[3] Pelissie du Rausa, M. Manyika, J. et al, Internet matters: The Net`s sweeping impact on growth, jobs and prosperity, McKinsey Global Institute, May, 2011.

[4] http://www.internetlivestats.com/

[5] ITU, ICT Facts and Figures, 2015.

[6] EC, A Digital Agenda for Europe, com (2010) 245 final, Brussels, 2010.

[7] OECD, Information Technology Outlook 2010, Paris, 2010.

[8] http://appsso.eurostat.ec.europa.eu/nui/show.do - last visited 17.1.2016.

[9] RZS, Upotreba informaciono-komunikacionih tehnologija u Republici Srbiji 2014, Beograd, 2014.

[10] RZS, Upotreba informaciono-komunikacionih tehnologija u Republici Srbiji 2015, Beograd, 2015.

[11] Bubanja, I., The importance of network presence for domestic company revitalization, Proceedings SPIN, Belgrade, 2015.

[12] UN, Implementing WSIS Outcomes: Experience to Date and Prospects for the Future, New York and Geneva, 2011.

[13] Statista - The Statistics Portal, Leading social networks worldwide as of August 2015: http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ - last visited 20.1.2016.

[14] We are social, Social, Digital & Mobile Worldwide Report, 2014: http://wearesocial.net/blog/2014/01/social-digital-mobile-worldwide-2014/- last visited 20.1. 2016.

[15] http://www.dzs.hr/ - last visited 25.1.2016.

[16] http://www.monstat.org/cg/ - last visited 25.1.2016

[17] http://www.stat.gov.mk/- last visited 25.1.2016

[18] Business Innovation Observatory, Innovative Business Models for Competitiveness, Social media for internationalization, EU, February, 2014.

[19] O`Relly, T., What is Web 2.0? Design patterns and business models for the next generation of software, web article, 30 September, O` Relly Publishing, Sabagtopol, CA, 2015.

[20] Laudon, K.C. and Traver, C.G., E-commerce 2015, Pearsons, Boston, 2015.For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

Homeostasis and Collaborative Decision Making for Smart and Cognitive Cities

Klasinc, Jan* **
* Croatian Institute of Public Administration
** Vertex projekt, ltd
e-mail:ajklasinc@gmail.com

Abstract – Tbe aim of the paper is to propose a new model for understanding smart and cognitive cities. While the term smart encompasses a lot of different ideas, cognitive when applied to cities means that a city is able to learn, sense and adapt based on experiences and changes in the environment. In theory, the interaction between the city and its inhabitants should work both ways - cognitive cities change the lives of their inhabitants and their behaviour as much as the inhabitants change the city and its processes by interacting with them. Desired states of wellbeing could be measured by Human Development Index (HDI) applied on a local scale as City Development Index (CDI), including health, education and wealth, but in complex environment such as a smart city wellbeing can take different and often unpredictable forms, Given the specific nature of smart cities that are ICT enabled, we would also like to propose an extension of this development model in order to include other development related concepts, such as social capital and mobility, quality of life and knowledge based capital, which could, despite some limitations of measurement in theory create reversible effects that allow sustainability.

I. INTRODUCTION

Smart city is a concept that is still not well defined, but it is generally understood to mean an urban area that uses information and communication technologies (ICT) to optimize the "efficiency and effectiveness of useful and necessary city processes, activities and services by joining up diverse elements and actors into a seamlessly interactive intelligent system."[2] According to Smart city readiness guide, "a smart city uses information and communications technology (ICT) to enhance its livability, workability and sustainability. In simplest terms, there are three parts to that job: collecting, communicating and "crunching." First, a smart city collects information about itself through sensors, other devices and existing systems. Next, it communicates that data using wired or wireless networks. Third, it "crunches" (analyzes) that data to understand what's happening now and what's likely to happen next."[3]Decision making in such systems should obviously be firmly based on knowledge management and collaborative thinking, increasing the knowledge based capital (KBC) of cities as a basis for overall development. EU report on smart cities (2007)[2] recognises six dimensions of smart cities: people, government, mobility, economy, environment and living, which means that smart cities enhance not only the collective intelligence of their citizens and communities,as well as their well-being and quality of life.[5] An analogy with living organisms that feed and grow, interacting with their environment could be useful. So in this paper we would like to suggest a model for smart city development that is based on open systems theory in which maintaining homeostasis, similar to the homeostasis in organisms, is necessary in order to preserve its health, preventing illness, and/or even death. Complex networked systems in changing environments such as cities are more likely to be nonlinear than linear, meaning also that they are more likely to be open than closed, which somewhat justifies our choice, although some opposite examples can be cited. In this paper we argue on the basis of socio-economic analysis that decision making in cognitive cities should be based on collaborative processes which involve an important degree of imprecision and complexity. Furthermore, we also analyse and evaluate available multi-criteria decision making models (MCDM)[7] such as Multi-attribute utility theory (MAUT)[4] and Analytic hierarchy process (AHP) [19] fuzzy nonlinear environments[6] involving Internet of Things (IoT), Internet of Services (IoS) and Internet of Persons (IoP) (possibly also citizens' pesonal biometric data, with due respect of privacy issues) as possible decision making solutions for cognitive cities[5].

II. METHODOLOGY

The equation that defines the thermodynamic framework for life and evolution is the equation for Gibbs free energy (Gibbs, J.W., 1873), namely:

$$G = H - TS \quad (1)$$

$$H = U + pV \quad (2)$$

where H is enthalpy of the system (or its internal energy at constant pressure), T is temperature, and S is entropy, or the degree of disorder of the system. Gibbs free energy is defined as a thermodinamic potential that measures the maximum of reversible work that may be performed by a thermodynamic system at a constant temperature and pressure. In the physical world, a chemical reaction will occur if the total change in the entropy of the whole world that would be caused by reaction is nonnegative. If temperature and pressure are held constant, the Gibbs free energy is the (negative) proxy for the change in total entropy of the universe. It can be found that systems with higher Gibbs free energy have better possibilites to evolve and maintain

homeostasis. There have already been useful applications of entropy in other fields than thermodynamics, such as information theory (Shannon), where it is used as 'information entropy', in statistical mechanics, and in social science. Social entropy is defined as a measure of natural decay within a social system – a decomposition of social structure, as much of the energy consumed by a social organisation is spent to maintain its structure, e.g. through legal institutions, education, or TV programming. Social entropy implies the tendency of social networks and society in general to break down over time, moving from cooperation and advancement towards conflict and chaos. As cooperation and conflict are two main modes in which sentient social beings can interact, the theory implies that conflict is the preferred means of interaction, whereas cooperation requires effort. ICTs are just another way of counteracting social entropy. If we apply the analogy to cognitive cities, the equation can be transformed to the equation describing "social capital" of cities, which is the amount of social 'energy' or 'freedom of choice' – ingrained in structural capital and other forms of social capital - that allows it to evolve and transform in changing environments, while maintaining internal stability. In general, knowledge based capital (KBC) can be divided into social capital and human capital[12]:

$$KBC = SC + HC, \text{ or } SC = KBC - HC \quad (3)$$

In this formula, we can see the analogy with Gibbs free energy, where SC is the 'free energy' of the system, that may be maintained under some constant 'pressure' and constant 'volume', e.g. when there is no migration or economic pressures, KBC is the 'internal energy' of the system, and HC is the 'social entropy' under constant 'temperature', or rather under stable Cost of living. In order to adapt to circumstances, cities have to employ 'disruptive innovation', which implies some entropic behaviour, in which negative change in social capital is offset by technological means, allowing the knowledge based capital to remain constant, as well as the Cost of living (CL). What has to change then is the internal social entropy of the system, that has to be reduced in order to maintain stability of the system, or its homeostasis. If the cost of living is, in fact, increased by the investment into new services, and the social capital is decreased (which is the measure of social entropy needed for disruptive innovations), than they have to be offset by some change in internal entropy of the system that makes it more organised or less chaotic, or, in general, improving Quality of Life (QL). The equation than becomes:

$$SC = KBC - CL/QL$$

The concept of 'social capital' is not entirely clearly defined in literature, but it is taken to mean the sum of norms and rules in a society that allow it to function and develop. It is the key to achieving social mobility. The problem with this concept is that some forms ot social capital that are acquired by tradition can be destructive for the society as a whole (e.g. Organised crime, mafia), in a similar fashion as a cancer is destructive to an organism. In order to overcome this problem, we take into account only positive forms of social capital that are measured in terms of free exchange of services and patents in a society, in this case a smart of cognitive city. If all citizens are allowed to be producers and consumers of services in a city in a joint collaborative effort, that the number of software or mobile applications created may be taken as an objective measure of social capital in a smart city.

A typical problem of cities is the problem of social mobility or migrants entering a city. Social capital would mean that the trust and other values that constitute it are not harmed, or if they are, by the reduction of KBC and increase of total CL for all citizens, that the reduction of internal entropy of the system can effectively replace some loss of social capital due to immigration, making cities more inclusive. However, this need not be sufficient, and more investment into KBC and networking should be made in order to make cities more sustainable, and this may also include investments into collaborative decision making.

III. Measuring knowledge based capital, social capital and social entropy of cities

As for measuring this effect, knowledge based capital (KBC) that plays the role of H (enthalpy) could, for instance, be measured by the number of highly educated and trained citizens and/or some other measure of knowledge based capital, such as participation of different sectors of economy in the overall economy of a city (e.g. electric equipment and software), and the term that gives the product of temperature and entropy (TS) may be translated into the ratio of Cost-of-living and Quality-of-living, given by human development index (HDI) for cities, or city development index (CDI). In this model, higher development of the city or higher CDI would translate into greater degree of order and reduction of disorder (entropy) in the city, whereas the 'temperature' that has to be maintained stable in an organism would translate simply into the total Cost-of-living for all citizens of a city. It can be easily proven that cities with higher cost of living reduce the degree of freedom and possibility for evolution of their citizens, while this can be somewhat offset by the higher city development or the reduction of 'entropy' of the system. In such a system, if we consider immigration of uneducated immigrants, this would not increase the first term (KBC), but would increase the Cost of living (CL) and reduce the CDI, which would create more disorder and reduce the freedom to evolve and create in changing circumstances (social capital of the city).

In fact, it is the negative change in social capital that allows for disruptive innovations (evolution and transformation) to happen, just like negative change in Gibbs free energy in living organisims, which in turn reduces the income of citizens, but also decreases the Cost of living and increases the Quality of life. In the long run they may also reduce the knowledge based capital, as the citizens become 'lazy' and rely only on changes brought about by technology. This is a well known reproach to some 'smart cities', such as Songdo, that they are in fact 'dumb smart cities', or that turning cities into smart cities reduces the 'inner life' or the need to adapt by intuition or by learning the secret processes of cities by experience. The amount of negative changes to social capital, although increasing Quality of life and reducing cost of life, may

also mean reducing KBC of cities, which has to be replaced by investment into KBC (education, training or social networking) in order to replace the lost social capital. In this sense, the 'reaction' created by change may become 'reversible', in the sense that social capital is never lost and city becomes sustainable. We may use income (I) of citizens as a proxy for SC, investments into education and networking as a proxy for change in KBC, and CDI for QL so the equation then becomes:

$$\Delta I = \Delta IKBC - \Delta CL/\Delta CDI \quad (4)$$

If $\Delta I = \Delta IKBC + \Delta CL + x$ then $\Delta CL\ (\Delta CDI-1) = x$

then either $\Delta CDI = 1$ or $\Delta CL = 0$ (5)

or ΔCDI is greater than 1 for every x greater than 0 (6)

It means that either all new income should be invested into education or social networking, without increasing the Cost of living (CL), but increasing possibly the Quality of life (QL), or that not all increase should go education and networking, that the Cost of living should increase, possibly with some other funds being used for other purposes, but also the Quality of life (QL). Depending on the amount of increase of Quality of life, the amount of 'free money' x is determined, and in cities with large increase in CDI, that amount is much larger than the increase in Cost of life. This means that cities which invest more in human capital through thier Quality of life achieved through better 'smart services' would get more highly educated immigration, and there would be no need to invest all surplus income into education and networking, either by citizens themselves, or through taxes (for education and smart city infrastructure).

In analysing which multicriteria decision making models should be used in cognitive cities in order to maximize social capital, or rather minimize the loss to social capital due to disruptive innovation and increase of entropy of human capital, we search for those models that would tend to maximize CDI and Quality of Life (QL), while minimizing the increase in Cost of life (CL). But we should also look for such models that tend to increase KBC through collaborative decision making processes.

IV. Empirical data for cities

City development index has been developed on the basis of human development index for countries.[1]

Table 1 – City development Index for various cities,1998

	CDI	City product	Infrastructure	Waste	Health	Education
Stockholm	97,4	93,5	99,5	100	94	99,8
Melbourne	95,5	90	99,8	100	93,7	94,1
Singapore	94,5	91,6	99,5	100	92,7	88,6
Hong Kong	92	89,4	99,3	99	90,9	81,3
Moscow	89,9	81	98,7	86,8	83,8	99,3
Seoul	86	65,3	98,4	100	88,7	77,7
Rio de J.	79,4	82,3	86,2	62,6	81,9	84,3
Sofia	79,1	70,9	93,7	58,5	86,2	86,3
Hanoi	74,2	59,6	72	90	80,6	69
Havana	71	65	74,8	50	80,7	84,7
Jakarta	69,2	66,2	57,3	46,7	80,2	95,7
Ulan Bator	68,4	53,7	59	90	72,5	66,7
Lahore	61,1	71,1	78,5	50	64,9	40,8
Colombo	58,4	46,9	68,6	45	86,2	45,3
Bangalore	58	51,1	82,7	31,3	76,5	48,5
Dhaka	48,4	55,6	45,2	27,5	64,6	48,7
Vientiane	47,1	44	58	0	62,3	71,3
Accra	46,6	49,4	50	0	71,4	62
Phnom Penh	43,5	40,2	33	27	47,2	69,9
Port Moresby	39,3	69	18,1	10	59,1	40,2
Lagos	29,3	42,1	29,5	2	44	29,1
Niamey	21,7	40	22	0	78,3	14,9

Source: UN-HABITAT http://urbandata.unhabitat.org/

As for cost of living, the differences among cities are listed in several categories, such as food, housing, clothing, transportation, personal care, entertainment. We can find the information at www.expatistan.com. For instance, the cost of living in London is 23% greater than in Singapore, although London is not as 'smart' as Singapore, and does not provide the same range of 'smart' services. If we take Stockholm as a benchmark, we can find following information about other cities in our list, and than compare the ration of Cost of life to Quality of life expressed by CDI, as well as KBC by the indeks for education, and average income (I) with Stockholm benchmark city as a proxy for predicted Social Capital (SC) of cities is calculated according to the formula as a difference between CL/CDI and KBC which is than added to the benchmark income of 100 in Stockholm. The average salaries in Stockholm, Sweden range from 20000 to 50000 SEK, on the average 35000 SEK, or about 4000 USD. On the other hand, average monthly salary in Korea is 5,058,122 KRW and 1 USD is 1234,64 KRW, so it is slightly above 4000 USD, or about 4100 USD, which could be true for Seoul as well, so this seems to agree with our hypothesis – although the cost of living is lower in Seoul than in Stockholm, its CDI is lower as well, including CDI for education component, which corresponds to the amount of KBC in Seoul and it yields roughly eaqual income. It could be said that taxation system in Sweden is more effective in maintaining good education system than Korean system that leaves more money to citizens to spend on education, and it shows somewhat better social capital than in Korea and Seoul, which accounts for somewhat better results for Stockholm than for Seoul when it comes to KBC, CDI, education and income, but not the CL, so the Korean system (Seoul) is more efficient, reducing the cost of life in order to maintain stable CL/QL ratio (or human

capital) with income that is roughly the same or even greater than in Stockholm, showing greater Social capital of Seoul. This means that Seoul is able to achieve more results in city development with fewer resources (lower taxes), which makes it smarter and more sustainable with respect to Stockholm, while being also able to maintain stable Social capital (SC) reflected in income that matches that of Stockholm. This leaves more money for Seoul citizens to spend on entertainment, culture, private business or other possibilities that may also increase their subjective quality of life, which is not measured by CDI, and may in the end account for some kind of accumulation of KBC that is not recorded by the CDI indeks, but which allows for reversibility of Social capital destruction and creation through innovation. This may be true, as South Korea is second highest ranked country when it comes to innovation according to Bloomberg, with overall score 2,26, innovation input score of 1,76, innovation performance score of 2,55, GDP of 1,3 trillion USD and 74,6 billion USD of FDI. No 1 in innovation is a smart city – Singapore with 142,4 billion USD in FDI.

Table 2 – Cost of Living (CL) in cities – with Stockholm as a benchmark

	CL (%)	Housing	Transport.	Pers. Care	CL/ CDI	KBC	I pred.
Stockholm	100	100	100	100	100	100	100
Melbourne	101	111	84	97	106	95	111
Singapore	111	128	115	96	118	89	129
Hong Kong	120	177	81	104	130	82	148
Moscow	56	52	29	56	62	100	62
Seoul	77	82	62	76	87	78	109
Rio de J.	56	52	49	75	71	85	86
Sofia	41	27	42	57	52	87	65
Hanoi	39	31	45	44	53	70	83
Havana	88	85	129	66	124	85	139
Jakarta	51	47	44	36	74	96	78
Ulan Bator	46	39	28	64	67	67	100
Lahore	37	31	26	37	61	41	120
Colombo	36	32	34	42	62	46	116
Bangalore	38	36	34	39	65	49	116
Dhaka	55	27	49	43	114	49	165
Vientiane	66	56	87	69	140	72	168
Accra	70	69	54	85	150	62	188
PhnomPenh	58	47	69	68	133	70	163
P. Moresby	124	200	48	109	316	41	275
Lagos	83	87	54	90	283	30	253
Niamey	73	62	96	62	336	15	321

Source: www.expatistan.com – calculation by the author

It becomes quickly evident that only the first seven cities could meet the criteria for 'smart' cities, whereas the figures of expected income become ridiculous for the lower portion. The key measure of CL/QL or CL/CDI yields the cities that are best ranked according to smart city readiness, or the most sustainable cognitive cities. If compared with income in those cities, we may see that there the income I corresponds to CL, but as it is highly diminished by CDI in HC term, the other term – KBC – is significantly greater. The reduction in income to make cities smarter produce some positive changes in KBC and negative changes in CL. In such a way, it is possible to achieve disruptive innovations in cognitive cities without increasing social entropy manifested in greater cost of life and lower quality of life, or reduced KBC, which makes them sustainable in a certain homeostatic way similar to living organisms in nature. However, the key to this 'reversibility' or resilience of social capital of cognitive cities is to enable simultaneous increase in KBC and QL through collaborative decision making. More recent data from 2012 (Appendix 1) seem to confirm this conclusion.

V. Different collaborative decision making models and their application in cognitive cities

In cognitive cities, the models of production of services tend to be based on collaboration, rather than separation of business, labour and market. Such models include cocreation or other forms of collaborative work, where all citizens can at the same time be designers, producers or consumers of services.[10] In order to facilitate such a model, collaborative decision making should be assisted by corresponding decision making models, and the goal is not just to allow for disruptive innovation emerging from collaboration, but also for simultaneous cocreation of intellectual (or knowledge based) capital – KBC – or learning through doing in collaboration. This allows cognitive cities to make up for lost income due to disruption, which also reduces human capital through the reduction of investment (or cost of living) and increasing of quality of life, through better services. It has been confirmed in some instances (e.g. use of smartphones, search engines) that humans tend to rely excessively on technology instead of their own thinking if it is available and tend to be active and creative when problems in the environment demand it and when they are able to solve them, especially when they are stimulated through competition to earn better salaries and cover the costs of living of a better social class. Therefore, in cognitive cities, this loss in competition (or conflict) and thinking capacity which constitutes human capital is theoretically offset through greater KBC creation through collaborative efforts or greater cocreation through cooperation of citizens. This makes cognitive cities less dependent on individual human capital, but more dependent on collaborative networks, that allow human capital to be transformed into KBC, that in turn may recreate social capital in order to make the process reversible or permit disruptive innovation to continue indefinitely while maintaining homeostasis, reducing internal social entropy and

constantly adapting to external circumstances. In such a case, MAUT[9] or AHP models for collaborative decision making should be used, but in fuzzy[17][15] environments of cognitive cities we also have to address imprecision and uncertainty.[8] The Analytical Hierarchy Process (AHP)[19] represents the hierarchical organization of given information, and it enables a hierarchical granulation[13][14][16][18], whereas "FAHP makes it possible to work with uncertain information and facilitates the interactive decision-making process. As usually no exact measurements are used in an interaction among multiple stakeholders, applying fuzzy logic is a way to address vagueness. Instead of searching for the best solution, it is often better to search for good enough"[11] solutions that fit the needs, which enables improved collaborative citizen management (E-governance) and decision making in cognitive cities.[20]

VI. Conclusion

In this paper a new model for analysis of smart or cognitive city development is proposed, which relies on the concept(s) of homeostasis and collaborative decision making in fuzzy (vague or uncertain) environment. As in biology homeostasis and possibility of evolution of living creatures are defined by free Gibbs energy and internal characteristics of organisms that allow it to maintain stable functions, such as temperature, we have proposed a formula that is analogous to the formula for free Gibbs energy to define Social capital of cities (SC) as the difference between total knowledge based capital (KBI) of a city minus the cost of living divided by the city development index (CDI). In this formula, cost of living plays the role of temperature that should be maintained stable, and city development index (CDI) is negatively proportional to the 'entropy' of the system, which can be measured as lower cohesion (income, education and other inequalities), negative fertility, and migration due to poor quality of life. The quality of life in this model is achieved through collaborative efforts that are the essence of cognitive cities. These efforts allow for greater degree of freedom of individual citizens, manifested in greater number of disruptive innovations, meaning that possibilites for evolution are greater, whereas on the other hand they maintain the social cohesion (homeostasis) necessary for stable development, which can be measured through 'cost of living' and 'quality of life' ratio. Data for different cities would seem to confirm the model, as cities with higher quality of life and lower cost of living have greater possibilites for evolution and attracting capital (investments), and their citizens have higher incomes. Therefore, it would seem that the best multicriteria decision making models for cognitive cities are those that create more knowledge base capital and allow for greater quality of life while keeping costs of living as low as possible, such as for example MAUT or AHP in fuzzy environments. With such models for collaborative decision making, it is predicted that the social capital of cities will be the greatest, which will allow for greater degree for freedom and evolution. The key to this 'reversibility' or resilience of social capital of cognitive cities which makes them similar to living biological organisms in nature is apparantly to enable simultaneous increase in KBC and QL through collaborative decision making in which citizens are at the same time producers and users of 'smart' services. Although this approach has some serious limitations in terms of quantifying actual data, it may be regarded as a useful model for understanding processes in cognitive cities and their dependence on collaborative decision making.

References

[1] http://www.un.org/ga/Istanbul+5/116.pdf The City Development Index

[2] Eraranta, S, Staffans, A: From Situation Awareness to Smart City Planning and Decision Making http://web.mit.edu/cron/project/CUPUM2015/proceedings/Content/pss/197_eraranta_h.pdf

[3] Smart Cities Readiness Guide http://smartcitiescouncil.com/system/files/main/premium_resources/Readiness-Guide-V2-8-24-2015.pdf?file=1&type=node&id=2532

[4] Baykasoglu, A, Development of Novel Multi-Attribute Decision Making Model via Fuzzy Cognitive Maps and Hierarchical Fuzzy TOPSIS, Journal of Informationa Sciences, Volume 301, Issue C, April 2015, pp 75-98

[5] Khansari, N, Impacting Sustainable Behaviour and Practice in Smart Cities

[6] van Ravenzvaaij, D, Moore, C.P., Lee, M.D, Newell, B.R., A Hierarchical Bayesian Modeling Approach to Searching and Stopping in Multi-Attribute Judgment, Cogn Sci. 2014 Sep-Oct;38(7):1384-405. doi: 10.1111/cogs.12119. Epub 2014

[7] Multi Criteria Decision Analysis for Strategic Decision Making, www.springer.com

[8] Opricovic, S, Gwo-Hshung T, Defuzzification within a Multicriteria Decision Model, *Int. J. Unc. Fuzz. Knowl. Based Syst.* **11**, 635 (2003).

[9] Gwo-Hshiung, T, Jih-Jeng, H., Multi Attribute Decision Making: Methods and Applications, Taylor and Francis Group, 2011

[10] Mostashari, A, Arnold, F, Maurer, M, Wade, J., Citizens as Sensors, the Cognitive City Paradigm, IEEE Explore, 2011

[11] Kaltenrieder, P, Portmann, E, , Finger, M, Applying the Fuzzy Analytical Hierarchy Process in Smart and Cognitive Cities https://diuf.unifr.ch/main/is/sites/diuf.unifr.ch.main.is/files/documents/publications/Applying%20the%20Fuzzy%20Analytical%20Hierarchy%20Process.pdf

[12] Schiuma, G., Lerro, A,Knowledge-based capital in building regional innovation capacity, Journal of Knowledge Management, ISSN 1367- 3270, Vol. 12 Iss: 5, pp.121 – 136, 2008, http://www.emeraldinsight.com/doi/abs/10.1108/13673270810902984

[13] Yao, Y. Y. 2000. Granular computing: basic issues and possible solutions. In Proceedings of the 5th Joint Conference on Information Sciences 1, 186-189.

[14] Yao, Y. Y. 2005. Perspectives of granular computing. In 2005 IEEE International Conference on Granular Computing 1, IEEE, 85-90.

[15] Zadeh, L. A. 1965. Fuzzy Sets. Info. and Control 8(3), 338-353.

[16] Zadeh, L. A. 1979. Fuzzy sets and information granulation, Advances. In Fuzzy Set Theory and Applications, M. Gupta, R. K. Ragade, R. R

[17] Zadeh, L. A. 1988. Fuzzy logic. Computer, 21(4), 83-93.

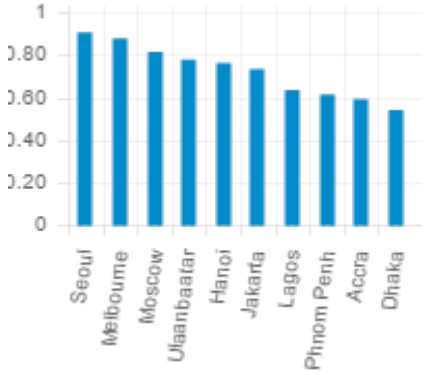
[18] Zadeh, L. A. 1998. Some reflections on soft computing, granular computing and their roles in the conception, design and utilization of information/intelligent systems. Soft Computing-A fusion of foundations, methodologies and applications, 2(1), 23-25.

[19]] Vaidya, O. S. and Kumar, S. 2006. Analytic hierarchy process: An overview of applications. European Journal of Operational Research 169, 1-29.

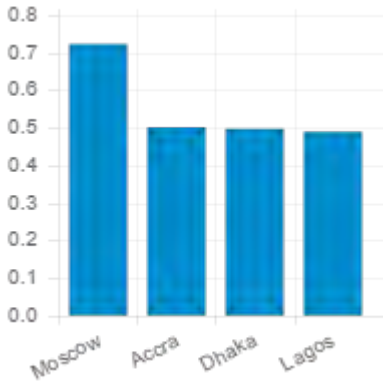
[20] Mostashari, A., Arnold, F., Mansouri, M., and Finger, M. 2011. Cognitive Cities and Intelligent Urban Governance. Network Industries Quarterly, 13(3).

Appendix 1 – Available data for various cities, 2012

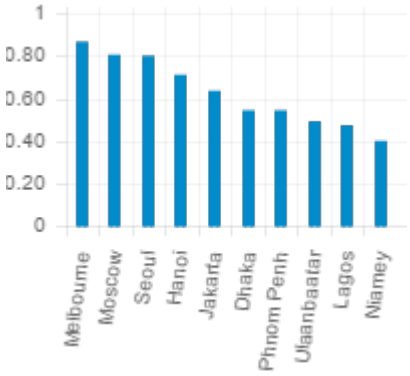
Quality of life



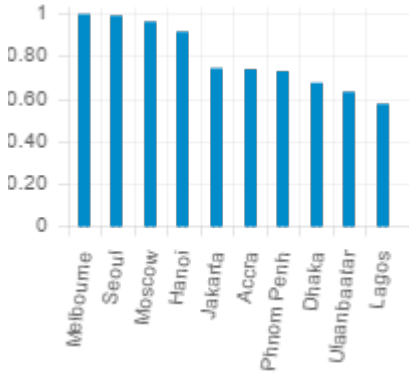
City prosperity index with six dimensions



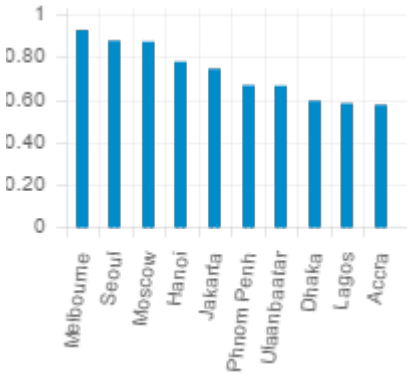
Productivity index



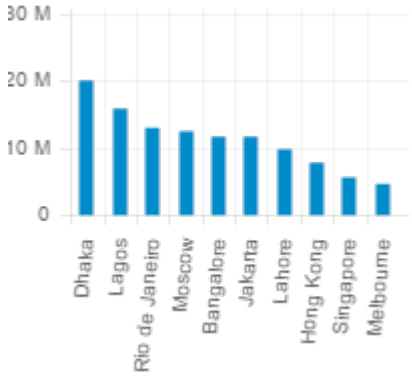
Infrastructure development



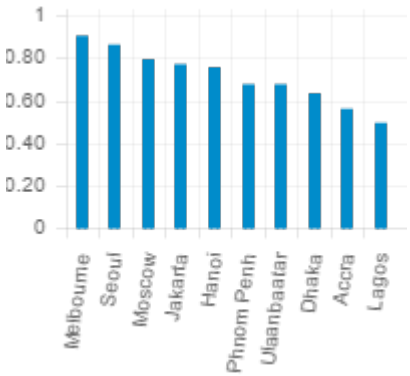
City prosperity index with four dimensions



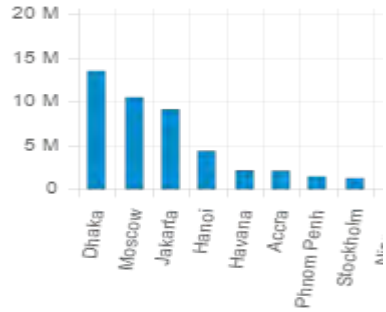
Urban population



City prosperity index with five dimensions



Populattion of capital cities



Source: http://urbandata.unhabitat.org/

Can the Bank Payment Obligation Replace the International Documentary Letter of Credit?

R. Bergami
Victoria University, Melbourne, Australia & University of South Bohemia in Ceske Budejovice, Czech Republic
Roberto.Bergami@vu.edu.au

One of the most fundamental considerations for export transactions is payment. It is commonly accepted Documentary Letters of Credit (DLC) are the payment of choice for high-risk/high-value transactions. In a DLC transaction the issuing bank provides an irrevocable payment undertaking that is only triggered when the exporter presents 100% compliant documents. DLC are governed by the UCP 600 rules, universally followed by all banks. Documentary non-compliance, estimated at 30% on first presentation, potentially nullifies the payment undertaking.
Even though DLC transaction data are generated using general, or specific, software solutions, paper-based documents are still routinely presented to banks, as earlier attempts to facilitate electronic presentations have failed, largely due to the inflexible rules (eUCP) surrounding such presentations.
The Society for Worldwide Interbank Financial Telecommunication (SWIFT) has devised a new electronic payment option – the Bank Payment Obligation (BPO), operating within SWIFT's Trade Services Utility platform.
Although SWIFT claims BPO solves documentary non-compliance, and potentially replaces DLC, on closer examination, the claim appears not to be justified. The conclusion is that, in reality, banks will not readily enter into BPO arrangements, especially with smaller firms, due to risk exposure considerations.

I. Introduction and Background

The notion of payment security in international transactions is one of the paramount considerations for both exporters and importers. Exporters need to be reassured timely payment is received for sold goods. Importers need to ensure they only pay on timely receipt of goods matching those ordered. It is not difficult to imagine reluctance from both exporters and importers to deal with less secure forms of payment under a number of risk conditions that take into consideration the value of the transaction, the payment period, the relative perceptions of country and customer/supplier risk and prior trading history between relevant parties.

It is commonly accepted that four traditional methods of payment are routinely used in international trade transactions. To set the background for this paper, a summary of each is provided below with an indication of the most appropriate risk environment for their usage [1].

1. Prepayment, or Clean Payment in Advance. The importer has complete trust in the exporter and pays for the good prior to shipping. Suitable only for small value transactions where there is little, if any, prior trading history. Typically this payment method is used in the beginning, before moving to the other options below;
2. Documentary Letters of Credit (DLC), one of the foci of this paper. This is an arrangement whereby the buyer's credit risk is replaced by that of their bank, the Issuing Bank (IB), typically the importer's bank, pursuant to the transaction governing Uniform Customs and Practice for Documentary Credits rules, commonly referred to as UCP 600, effective 1 July 2007 [2]. The IB provides an irrevocable payment undertaking to the exporter. The DLC for security, counter-fraud and fund transfer reasons is advised through a local bank, typically the exporter's bank. The IB's payment undertaking may only be called upon where the exporter is able to present all required paper-based documents, in one presentation, that are 100% compliant with the DLC requirements, in timely fashion. DLC are suitable for high value, high risk transactions, as banks provide third-party independent oversight on documents – this will be discussed in greater details later in the paper. DLC values average between USD 500,000 and USD 750,000 [3]. Indeed, the perception that DLC are viewed as a good trade finance option is evidenced by recent estimates claiming this method of payment accounts for approximately 43% of world trade, with a failure (default or non-payment rate) averaging less than 1% [3]. Without doubt DLC are the most complex of traditional payment arrangements, as shown in Figure 1;
3. Bill of Exchange, or Draft. An arrangement whereby the exporter sends the goods and subsequently request payment by lodging documents via the banking system. This is suitable for mature relationships where customer/country risk is perceived as being relatively low. This method of payment accounts for about 16% of trade [3]; and
4. Open Account, or Clean Payment. Accounting for about 16% of payments, this is suited to very mature relationships, a high degree of trust, an

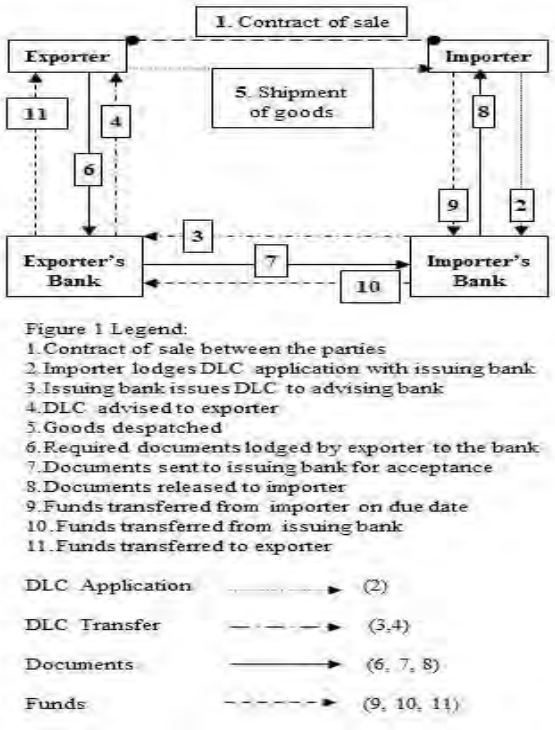


Figure 1. Typical Deferred Payment DLC Cycle [4]

interdependent economic relationship between seller and buyer, capped transactional exposure, and very low perceived customer/country risk.

The Bank Payment Obligation (BPO) the other foci of this paper, is a relatively new introduction that relies on e-commerce notions and is touted by its supporters to be superior to the traditional DLC and arguing it will soon replace the old style payment (the DLC). How true this claim is will be explored in this paper.

The next section provides an outline of the main benefits and pitfalls of DLC, citing only the most relevant literature as appropriate, and including earlier efforts to electrify all, or part, of its operations. This is followed by an explanation of the BPO, a method of payment about which there is a dearth of academic literature, because it is a recent introduction. This is followed by a discussion of whether the BPO may replace the DLC, before reaching the conclusion.

It should be noted the terms exporter/seller and importer/buyer are used interchangeably in this paper.

II. DLC AND BPO

DLC have been around in one form or another for millennia, depending on which school of thought one believes, and the definitions given to them. There is evidence of international trade and finance in writings in clay tables dating back to 3000 BC [5]. The Italians, who were influential bankers of Europe around the later Middle Ages and Renaissance periods developed relatively modern banking techniques, including currency stability in Europe with the introduction of the gold florin issued in Florence in 1252, and foreign bills of exchange (probably to reduce the risk and weight of carrying money *in specie* [6]) and their liability. Indeed, the notion of "*chi accetta paghi*" (the acceptor must pay) remains well entrenched in current transactions involving the bill of exchange. As international trade flows grew over time, the bill of exchange lost its ability to satisfy the demand for trade finance and it was firstly replaced by the banker's credit [7], then the DLC that became prominent in the last century, but particularly since WWII. The need to rebuild economic activities coupled with the lack of trust between trading parties whose countries may have been warring only a short time before, led exporters and importers to seek alternative solutions that provided a degree of 'independent financial trust'. The DLC has proven to be the payment method of choice in high-risk/high-value propositions.

The DLC is not perfect. It is, as Figure 1 shows, quite complex. The rules surrounding its use are also not particularly well understood by exporters, in particular, as there is a high discrepancy rate on documents, yet there is a low default rate. This may seem like a paradox, therefore, it is important to appreciate the role banks play in the transaction, and how the rules operate, to fully understand the DLC cycle.

To recall, the DLC replaces the credit risk of a buyer with that of their bank - the IB. The decision to issue a DLC is one of commercial practice. The IB typically requires some form of security be provided by the importer. This type of security is usually in the form of liquid asset collateral. The amount of security varies from transaction to transaction, the value involved, the credit worthiness of the importer, the IB's practices, and banking regulations. For example, In Germany it is known banks commonly require 30% security, whereas in China it may be as low as 5%, but in Australia it is 100% [8]. The issue of security will become important when the paper later considers BOP operations.

Article 4 of the UCP 600 provides a clear separation between the contract of sale and DLC operations, by what is commonly referred to as the 'independence principle'. This is done because banks are not a party to the contract of sale, but are definitely involved in the DLC cycle. By issuing the DLC in favor of the exporter, the IB effectively 'underwrites' the debt incurred on purchase, but it does so conditionally. The proviso for payment is that data on documents required by the DLC must be 100% compliant and satisfy the standards for examination of documents outlined in Article 14. This is commonly referred to as the 'doctrine of strict compliance'.

Article 5 of the UCP 600 further stipulates certain parameters relating to the role of bankers, stating in DLC transactions that banks deal with documents, and not goods, services or performance to which the documents may relate. The default document definition is that these are conventional, paper based documents. Article 5 assumes increasing importance when its implications are fully considered. Given banks are limited to dealing with documents only, two possibilities arise in a DLC transaction. One possibility is documents are flawed, but the goods match those ordered in all respects, therefore,

there is a risk of non-payment. The other possibility is the goods are flawed, however, the documentation is 100% compliant and the payment must be effected, unless fraud can be proven, whereby a fraud exception rule applies preventing payment, or a court injunction exists preventing such payment. These conditions explain, at least in part, the role of the banker and the UCP 600 limitations imposed on them. This ultimately means the parties involved in a DLC transaction exercise an implicit degree of trust towards each other, and that trust is 'facilitated' by the bank as the 'independent party'.

The main problem in DLC transactions is the accuracy of documentary data, from both the exporter's and banker's perspectives. As shown in Figure 1, steps 5 and 6, once the DLC has been received by the exporter, the goods are shipped. The documents are subsequently lodged with the exporter's bank, within a specified time period from shipment (presentation period). The role of the exporter's bank is to check for documentary compliance within a maximum period of five clean bank working days, pursuant to Article 14b, and advise the exporter of discrepancies or acceptance of documents. Where discrepancies exist, the exporter is entitled to retake possession of the documents, correct them if possible, and represent same for another check. This process can be repeated indefinitely within the allowable presentation period.

There are no accurate statistics on the magnitude and significance of the documentary discrepancy problem at a global level. The International Chamber of Commerce (ICC) in 2002 estimated a 70% discrepancy rate [9], however, in 2010 it admitted earlier estimates were not based on credible data due to a lack of a comprehensive set of statistics on trade finance and could, therefore not be relied upon [10]. There is still a general lack of information about discrepancy rates, as little detailed empirical research has taken place in this area. An Australian study estimated approximately 26% of DLC transactions are subject to documentary data discrepancies on first presentation, with an estimated annual correction cost between AUD 9.36 Million and AUD 14.69 Million to industry in lost productivity. Given Australia represents approximately 1.5% of world trade and DLC represent 43% of world trade, if the Australian study estimates are accepted as representative, we may estimate the mid-point average annual global cost of DLC errors to be approximately AUD 345 Million.

It is important to identify the underlying causes for paper based documentary discrepancies. Based on the Australian study, it is claimed documents are generated manually in 50% of transactions (mainly small enterprises) and it is not difficult to imagine human error being a significant contributor to discrepancies in the absence of automated systems. Human error may be further implied on examination of the type of discrepancies and causative factors that render the DLC documentation non-compliant, as shown in Table 1 [8].

From Table 1, documentary discrepancies items a, b, c and e are routinely completely generated by the exporter in-house, consequently, there is little opportunity for external causative factors in discrepancy rates. Items d, f, g and h are completed based on information supplied by the exporter to third parties, consequently, it may be possible for errors to be generated externally, however common practice is for those third parties to provide draft copies of documents to the exporter for approval prior to issue. This may be less so the case for items g and h, although the requirement for these types of documents remain comparatively low due to their nature. In considering discrepancy factors, it is the responsibility of the exporter to ensure the transaction is executed in a timely fashion. It may be delays are due to external factors such as late delivery of raw materials from another supplier, as an example. However, changes to shipment dates can always be negotiated to ensure compliance. As discrepancy factors are beyond the scope of this paper they will not be considered further.

TABLE I. DLC DISCREPANCIES

Discrepancies	**Percent**
Documentary discrepancies	
a. Errors in bill of exchange	15.9 %
b. Errors in the commercial invoice	16.7%
c. Errors in the packing list	15.9%
d. Errors in the transport document	17.4%
e. Errors in the insurance document	14.4%
f. Errors in the inspection certificate	10.6%
g. Errors in the certificate of origin	13.6%
h. Errors in government certification	8.3%
Discrepancy factors	
i. Late shipment	12.1%
j. Missed consignment in a predetermined delivery schedule	6.8%
k. Documents lodged outside allowed presentation	12.9%
l. Documents presented after Letter of Credit expiry	9.8%
m. Missing documents	11.4%
n. Incorrect shipment/partial shipment	8.3%
Total	**100%**

It is generally accepted humans are a considerable risk to any organization, simply because, unlike machines, their behavior is less homogenous and reproducible, therefore, certainty of exact process replication is typically comparatively less high than that of a machine. Consequently, automation, wherever possible is preferred, as this is likely to reduce errors, provides greater confidence of process outcomes and improves efficiency. In the case of DLC transactions this would also mean reducing financial risk for exporters through lower discrepancy rates.

Attempts to automate DLC transaction were made by the ICC in 2002, with the issue of the eUCP. The eUCp is supplement to the UCP, allowing for presentation of some or all documents in electronic form, not limited to submitting all documents in one presentation, thereby

catering for situations where there may be a blend of paper-based documents and electronic records. There is a dearth of literature on this topic, with extensive database searches locating only two articles by the same author found on this topic [11] [12].

The first article deals with the introduction of the eUCP and points to an inflexible approach to electronic records (document) presentation, where it appears the applicable rules provide a high degree of risk mitigation for banks by unreasonably shifting the burden of electronic communication and delivery of records responsibility to exporters. It is further pointed out the rules ignore the practicalities of commerce. As an example, the eUCP allows third parties to lodge electronic records directly with the bank, but demand the exporter send a message to the bank confirming presentation of all documents/records has been completed. Just how the exporter would be in a position to do so is unclear, unless the exporter was copied in on the communication – and even then there is no guarantee the third party message has in fact been delivered to the bank, or that the bank can read it. It is further questioned whether it would be desirable to have direct submission of electronic records as this would prevent the exporter from checking their data content and having potential discrepancies corrected before lodgment. It should be remembered that regardless of how documents or records are presented, the exporter retains the financial risk of non-payment until the IB is satisfied the documents comply. To overcome this situation the only solution would be for the third party document producer to send the electronic record to the seller for on-forwarding to the bank. However, this would nullify any electronic messaging efficiency gains.

The second article merely points to the fact the latest revision of the eUCP was a simple cosmetic update by referencing to the, then newly introduced, UCP 600 without changes to the rules themselves. Suffice to say, the attempt by the ICC to electrify the DLC resulted in an abysmal failure with universal lack of adoption of the eUCP by exporters and banks alike. Due to its abject failure, further consideration of the eUCP in this paper is considered futile.

The failure of the eUCP was perhaps the catalyst for the Society for Worldwide Interbank Financial Telecommunication (SWIFT) to enter the financial services market with the BPO innovation. Traditional methods of payment outlined earlier in the paper have been in existence for some time and the BPO is a rare new introduction.

The BPO is an irrevocable undertaking to pay given by the buyer's bank (Obligor Bank) to the seller's bank (Recipient Bank) that payment will be made on a specified date after a successful electronic matching of data in accordance with the ICC Uniform Rules for Bank Payment Obligations (URBPO) [13]. In a BPO the seller sends shipping documents directly to the buyer, whilst the payment is effected separately electronically. The BPO electronic data matching process is conducted within SWIFT's Trade Services Utility (TSU) using ISO 20022 messaging standards.

The typical BPO cycle is shown at Figure 2. Following conclusion of contract negotiations (step1) the buyer asks their bank (Obligor Bank – OB) to establish a BPO (step 2). A baseline (electronic data requirement) is set by the OB, based on the data from the purchase order. This is done through SWIFT's TSU. The TSU sends the BPO message to the seller's bank (Recipient Bank – RB) who in turn forwards the BPO to the seller (step 3). The seller ships the goods (step 4) and submits electronic data to the RB to satisfy baseline requirements (step 5). The RB submits the electronic data to the OB via TSU. The TSU electronically matches the data received from the RB and provides both banks with either a match or a discrepancy report. This is an automated process. Once the OB has the report, it forwards this to the buyer (step 6). Where there is a match the obligation to pay is inherent in the BPO and the RB will be able to confirm the payment date to the

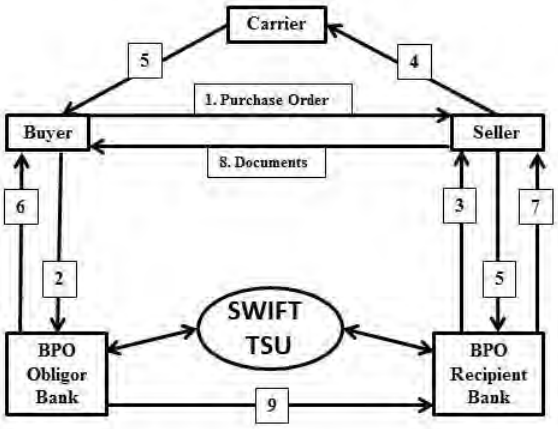


Figure 2 Legend:
1. Purchase order from buyer to seller
2. Buyer requests BPO from their bank based on Purchase Order
3. Seller's bank advises BPO
4. Seller ships goods
5. Seller submits electronic data to match BPO baseline
6. Data match report from TSU provided to buyer by their bank
7. Payment confirmation given to seller with due date
8. Seller send buyer documents
9. Payment effected on due date

Figure 2. Typical BPO Cycle [14]

seller (step7). Once payment has been confirmed, the seller sends the relevant documents directly to the buyer (step 8). Payment occurs as due (step 9).

All electronic data must be ISO 200220 compliant and is routed to the TSU via the banks. From an IT perspective the bank to corporate communication is outside the BPO and not governed by any particular standard. Banks are responsibility for ISO 20022 compliance in the TSU. From SWIFT's perspective proprietary software provides the solution and message traffic accrues a fee. There is no significant change to SWIFT's existing operations from a technical perspective. It is not anticipated the BPO will result in significant changes in international settlement and its impact on SWIFT will be minor in the short term. If BPO transactions increase, SWIFT gains additional revenue, but only where such transactions do not replace DLC, because both methods of payment work through

SWIFT. Where the BPO displaces open account or Bill of Exchange payment options, SWIFT gains additional business as these operate outside TSU.

The BPO was announced as an alternative method of payment by SWIFT in 2011 [15], but it was without rules for the first two years of its existence, experiencing poor uptake by banks and traders alike. In an attempt to give credence and entice acceptance of the BPO, SWIFT 'partnered' with the ICC, ultimately leading to the issue of the URBPO issued by the ICC in 2013[16].

The BPO is new, its usage is in the infant stage, and a significant body of literature has not yet emerged. Despite rigorous database searches on academic literature only three articles were found focusing specifically on BPO.

The first was written by the Head of Banking and Trade Solutions at SWIFT and, not surprisingly, this work extolls the virtues of BPO, but is short on providing a balanced view and discussion of BPO [17]. In that article, according to the author, as at June 2012, only 37 banks across the globe were adopting the BPO, hardly a significant number in world terms.

The second article [18], authored by two bankers, despite making a vague claim there is growing belief in the potential of BPO, also points to the challenges of document standardization, lack of BPO awareness and education, Basel III treatment and legal enforceability.

The third article [19] considers some of the differences between DLC and BPO transactions. It concludes the BPO is not currently capable of replacing the DLC, partially due to the challenges banks and traders alike may have in ensuring compatibility of data transfer. Banks are typically members of SWIFT, therefore capable of communicating ISO 20022 messages to and from TSU. The same cannot be said about buyers and sellers. Consequently, banks will need to develop and provide traders with appropriate IT solutions enabling data transfer to banks for subsequent transmission to TSU. This requires a 'mind-shift' for those used to dealing in a paper-based environment. Acceptance of the new data transmission environment is part of the transition to the BPO. It is also pointed out that under DLC transactions there is a human 'sniff test' with document, one that will be lost with automated processes [19]. Automation may be problematic or even prohibited if recommendations by regulators, such as the UK's Financial Conduct Authority, for more control over underlying transactions, are implemented [19].

Although it appears similarities exist between DLC and BPO transactions, a deeper analysis of why DLC are used is warranted, to understand where BPO may fit in traders' choice of method of payment.

The DLC has a long history of usage, is well established and well respected. Despite the 'horror stories' of discrepancies, the end result of utilizing this method of payment is a very low failure rate. Documentary security is also well established. One important consideration in DLC transaction is that being paper based, there are no legal impediments surrounding negotiability, an aspect that to date has not been generally resolved. Bills of Lading, issued for the transport of goods by sea are still largely negotiable documents, capable of being transferred from one part to another by endorsement. In an electronic format, this is not possible, other than if the transaction is conducted via special arrangements such as BOLERO or essDOCS – private electronic environments. Of course for such transactions to work, everyone has to be a member of such an environment. However, on a global basis this is currently not the case, largely due to costs, although greater adoption is being noticed. Internationally, the legal system has not yet developed far enough to enable electronic negotiable documents to be a reality.

It should be noted the BPO does not replace documentation, but merely offers an alternative to human documentation checking currently operating in DLC. However, this option may give rise to a false sense of security, particularly where immediate payment is required. An example may be useful in explaining this situation.

A seller ships goods that are not compliant with the Purchase Order, but submits nevertheless compliant electronic data that is then automatically processed through the TSU. Under these circumstances, a data match report will be generated and an obligation to pay incurred by the OB. Banks are not in a position to check any data, as they do not receive any documentation. As the payment is immediate in this scenario, the seller receives the funds very quickly. The seller may send the documents to the buyer by post, but it is not difficult to imagine these may be received after payment has been effected. On discovering the goods are non-compliant, the buyer has to seek remedy. However, as BPO Article 6 separates the contract from the BPO, and Article 7 limits banks to dealing with data only and not the documents, the buyer has no remedy against their bank. The remedial option available to the buyer is to commence financial recovery through legal means. It is well known that international litigation is very slow to finalize and extremely expensive to pursue.

Relative to the DLC, the BPO alters the balance of trust, not only for buyers, but also for banks. In DLC, document must be presented and checked for compliance and funds are subsequently paid to the seller. The buyer has confidence they will get the documents from their bank. There is certainty in this. The documents exist as the seller must surrender these before payment is made. There is no such certainty in a BPO transaction. The documents will in all likelihood exist, but their data content may not be correct and banks do not check same. Under these circumstances, the buyer must trust the seller to a much higher degree than with DLC. The OB also has to consider the risk posed by the seller (and their nation). As the OB has a transactional exposure once the BPO has been set up, they may wish to take security from the buyer to protect their liability risk. In this respect the approach would be no different to the DLC.

The claims that BPO accelerates the payment cycle are only partially true, as these depend on the agreed payment period in the contract between the seller and buyer. Where the maturity of the transaction is in the future, for example 90 days, it would not make any practical difference whether the document checks are manual or automated, as

these would be concluded well before funds transfer to the seller. Even with the comparatively slower manual DLC check, banks each have a maximum documentation checking and acceptance window period of five clear working days, pursuant to UCP 600Article 14b. Where DLC payment is immediate on presentation of documents at the counters of a bank in the seller's country, it is arguable whether the BPO would provide any quicker payment turnaround. The benefit under the DLC is of course the documents exist and have been checked.

It is also doubtful the BPO would serve the requirements of buyers in string sales, where the Bill of Lading is transferred down a chain of buyers and sub-buyers on shipment, with property following the transfer of the paper document [20]. String sales typically occur while the shipment is en-route. Given there is no certainty about the existence of the paper documents, or when these may be released to the first buyer, for subsequent hand over down the chain, this presents a problem. String sales can only work where the documents are able to be transferred relatively quickly to subsequent buyers so timely discharge can be achieved at final destination, to avoid storage charges and delayed availability to market. Certainty on the existence of documents is a paramount consideration under these circumstances.

III. CONCLUSION

The BPO presents an interesting development in international trade finance. It is certainly an innovation, one that does not occur regularly. Indeed, before the DLC became commonplace earlier last century, the Bill of Exchange had been the primary method of payment for several centuries beforehand. It is argued here the BPO provides an added alternative to the traditional suite of methods of payment, not a replacement for existing payment options. The reasons for this claim include the following considerations.

There is little adoption of the BPO to date, as a transition to data transfer is required from both a technical as well as a human perspective. The BPO is likely to be limited in application for the time being, as the OB assumes the payment risk, by providing an irrevocable undertaking to pay against matched data, without the security of physically seeing documents – this is a departure from traditional practices, one that may at least alter the risk perception in the transaction, although there is nothing stopping the OB from obtaining security and buffering its financial risk exposure. The claim BPO accelerates payment cannot be substantiated, especially where payment occurs in the future, and the BPO is not suitable in string sales.

Therefore, the BPO is not, and need not be, a replacement for the DLC, rather a widening of payment choices between traders in foreign nations.

There is scope for additional inquiry into this new method of payment, as over the fullness of time, practices may alter and adoption increase, giving rise to a better understanding of procedures and practices used. As technological changes are expected to bring greater flexibility, especially in relation to the electrification and legal recognition of negotiable documents, the BPO may well end up being a formidable opportunity for international trade transactions in the future, but not yet.

REFERENCES

[1] R. Bergami, Internationat Trade: A Practical Introduction, 4th ed., Melbourne, Australia: Eruridtions Publishing, 2003.

[2] International Chamber of Commerce, ICC Uniform Customs and Practice for Documentary Credits: UCP 600 , Paris France: ICCPublishing, 2007.

[3] International Chamber of Commerce, Rethinking Trade Finance, Paris, France: ICC Publishing, 2015.

[4] R. Bergami, 'Enter the UCP 600' presented at *"Pan-Pacific Conference XXIV"*, Dunedin & Queenstown, New Zealand, 2–4 June, 2007.

[5] N. Ferguson, "Innovations in the Approach of Merchant and Trading Banksto Financing Overseas Sales" in Current Problems of International Trade Financing, C. M. Chinkin, P.J. Davidosn and W.J. M. Ricquier, Eds. Singapore: Butterworths, 1983.

[6] A. Mugasha, The law of letters of credit and bank guarantees, Sydney, NSW, Australia: The Federation Press, 2003.

[7] Thayer, P. W., "Irrevocable credits in international commerce: their legal nature", Columbia Law Review, vol. XXXVI, no. 7, pp. 1031–1060, 1936.

[8] Bergami, R. Risk Management in Australian Manufactring Exports: the Case of Letters of Credit to ASEAN, Melbourne, Australia: Victoria University, 2011.

[9] ICC Thailand, "Examination of documents waiver of discrepancies and notice under UCP500", Bangkok, Thailand: ICC Thailand, 2002.

[10] ICC Banking Commission, "Rethinking trade finance: global survey", Paris, France: International Chamber of Commerce, 2010.

[11] R. Bergami, "eUCP: A Revolution in International Trade?", Vindobona Journal of International Commercial Law and Arbitration, vol. 8, no. 1, pp. 23–36, 2004.

[12] R. Bergami, "ISBP, Non-bank Letters of Credit and eUCP: A Commentary", Vindobona Journal of International Commercial Law and Arbitration, vol. 11, no. 2, pp. 147–157, 2007.

[13] Opus Advisory Services International Inc., Educational report: Observations on the Evolution of Trade Finance and Introduction to the Bank Payment Obligation, https://corporates.swift.com/sites/sdccor/files/trade_opus_swift_observations_2013.pdf, March 2013.

[14] Quinn, M., Gorzas, I. and Ozikis, B., "Bank Payment Obligation: Financial Instruments to Manage Risk in the Open Account Trade", Washington, DC, USA: International Finance Corporation, http://www.ifc.org/wps/wcm/connect/e2be7c004d0e681887d6c7f81ee631cc/GFM-TSC-BPM5-Quinn.pdf?MOD=AJPERES, 12 October 2012.

[15] A. Casterman, Bank Payment Obligation, Society for Worldwide Interbank Financial Telecommunication, http://www.iso.org/sites/finservworkshop_nl_2011-05/presentations/P%201150_2%20Casterman.pdf, 2011.

[16] International Chamber of Commerce, Uniform Rules for Bank Payment Obligations, Version 1.0, Paris, France: ICC Services Publication Department, 2013.

[17] A. Casterman, "Modernising Global Trade Finance Practices", Journal of Payment Strategy and Systems, Vol. 6, no. 3, pp. 225–231.

[18] C. R. W. de Meijer and M. Menon, "Bank Payment Obligation: The Missing Link?", Journal of Payment Strategy and Systems, Vol. 6, no. 3, pp. 232–245.

[19] G. l. Wynne and H. Fearn, "The Bank Payment Obligation: Will it Replace the Traditional Letter of Credit Now, or Ever?", Butterworths Journal of International Banking and Financial Law, pp. 102–104, February 2014.

[20] R. Aikens, R. Lord and M. Bools, Bills of Lading, 2nd ed., London, UK: Informa 2016.

Implementation and design of Cool'n'Project - web-based project management software

I. Špeh
Virovitica College, Virovitica, Hrvatska
ivan.speh@vsmti.hr

Abstract - In this paper, a web-based Project Management Information System (PMIS) application aimed for project tracking, monitoring and analyzing in local government units is described. This tool is used to collect, combine and distribute information about projects across different end-user devices. Projects are categorized into groups based on the project type, financing source, maintainer, deadline, priority and several other attributes. Each project is presented in a GUI with tabs that are filled with real-time data and information about project status and achievements, tasks, schedule, go-location view, gallery, phonebook and timeline. Employees of a local government unit have different roles and abilities in a specific project in which they participate. Project managers are provided with decision-making support needed in scheduling, controlling and organizing a project. During the project execution phase, the project management team collects information, files, images, reminders, contacts and contact history information into several databases. At each project phase, different types of reports and diagrams are available. This type of a Project Management Information System tool is developed by the local governments need to track a large number of projects*.

Keywords: Project Management Information System, project tracking, web-application, local government

I. INTRODUCTION

Multi-project environment features can be recognized in local government units which handle with a large number of projects. The multi-project environment is defined as a setting in which project managers and project members are participating in several projects at the same time. In organizations that are engaged in many projects simultaneously, management is faced with multiple challenges in resource planning, prioritization and monitoring [1].

Problems in managing multi-project environments are the improper implementation of the pre-project phase, infrequent project-progress monitoring, no defined owner, lack of information on projects, inadequate flow of information across the organization, inadequate project activities, unclear roles and responsibilities, improper information distribution and many others. These problems can lead to longer project execution time or even failure of the specific project. [2]

The use of Project Management Information System (PMIS) is considered advantageous to local government leadership because of the contribution regarding project success [3].

The implementation of PMIS in a multi-project environment may help accomplish a realistic project assignment, which is an effective strategy when handling multiple projects [4].

Project Management Information Systems can be defined as tools and techniques used in the management of simple or complex projects. It can be described as an electronic information system used to plan, schedule, control, report, communicate, forecast, and handle cost for most aspects of a project [6]. Therefore, PMIS should provide project managers with decision-making support for planning, organizing and controlling project [5]. For efficient work performance among team members, PMIS supports three basic functions [7]: communication – PMIS delivers related knowledge and information promptly between members of the team; collaboration – PMIS supports an active cooperative management system among the members; community – PMIS supports the accumulation of related information and data through information sharing.

During the project and following completion, large volumes of information is generated. Any project will involve huge volumes of drawings, photos, documents, e-mails, with multiple versions arising throughout the lifecycle of the project [8]. PMIS provides the framework for collecting, organizing, storing, and processing project information. It provides the basis for assessing the status of the project with respect to time, cost and performance goals and objectives.

Project management software has become a prerequisite to manage projects more efficiently and effectively, and to aid project managers in their decision-making. With an estimated $255 billion being spent on IS/IT applications annually, it is significant that Information Systems projects should effectively manage. [9].

Although projects are most often used in information technology (IT), software and development, business process reorganization and research and development [10], constantly growing the number of projects in local government units and local project management organizations makes them one of the most important factor in community development.

Among various IT solutions, the web-based PMIS has been highlighted because of its strong advantages [7]. The aim of this paper is to describe a proposal of a web-based Project Information System Application which is the product of cooperation between IT department and local

This paper is the result of implementation Cool-N-Project software in several local government units in Virovitica-Podravina County.

government. The application is currently in use at several local government units in Virovitica-Podravina county.

The aim of described web management application is to track projects with a large number of participants and low dynamic. In most cases, a local government unit has a large number of projects which have been created over the past few years. Using described web application local government employees are able to get information about current project status and other project information in just a few seconds. There are dozens similar web applications available online like Aceproject and Teamwork, LiquidPlanner [16]. Main differences between alternative applications are:

- Low cost - problem lies in the fact that similar applications are most often not affordable to local government units
- Language - low-level of foreign language skills for most employees is most often the reason why similar application are not usable.
- Different end-user devices - most of the similar applications are executable, so additional installation is required. A web browser and Internet access are only requirements for using this application.

II. ARCHITECTURE

An architecture of the web application can be described as a client-server three-tier architecture in which the graphical user interface (presentation layer), application functions and logic (application layer) and computer data storage (database layer) are developed and maintained as independent modules, on separate platforms.

Presentation layer – functionality of HTML-based graphical user interface has been extended with custom javascript/jquery functions and several plugins of which most used and most important are jquery datatable for display data in functional tables, parsley for validating data, datetimepicker, selectpicker and other bootstrap friendly jquery plugins. Using AJAX for communication with Apache web server features of a single page application are achieved. When a user action generates a client call to retrieve new data, the server only returns view fragments. This means that is possible to update parts of the application, without reloading entire viewport. At the side of design, the responsive bootstrap framework is used.

Ajax request sent by the client is received on the application layer with PHP scripts. There are three types of PHP scripts:

1) PHP script that handles users actions – used for inserting, updating and deleting data from data server
2) A PHP script that prepares data – after a request for data is received, this type of script is used for connecting and retrieving data from MySQL server which is then encoded in JSON format and sent back to client
3) configuration PHP – used for manipulating with session storage, database connection and other application configuration variables.

On the database layer, there are several MySQL database servers where information is stored and retrieved. Data in this tier is kept independent of application servers. There is the database used for configuration storage, data storage and domain with users storage. A domain stored in users database is an organization (public administration unit) with users (employees) which have access to the application and can share projects with each other.

III. RELATED WORK

In the project management literature, the definition of the project has been discussed by numbers of literature. One of the definitions defines projects as a temporary (definitive beginning and definitive end) endeavor undertaken to create a unique (projects involve doing something that has not been done before) product or service [8]. In the presented PMIS, a project is defined as a logical framework within multiple project activities can be defined. For example, generic types of project in local government unit can be an organization of city festival, restoration of city buildings, education of older city population or improving city traffic infrastructure. These projects consist of several activities which need to be accomplished. After activities are finished, the project can be closed.

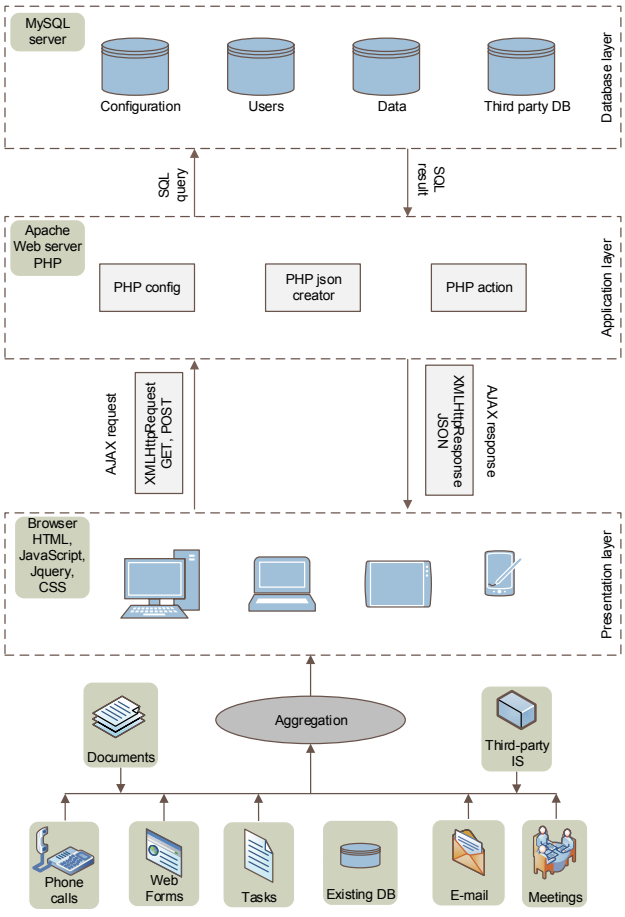


Image 1. Web application architecture

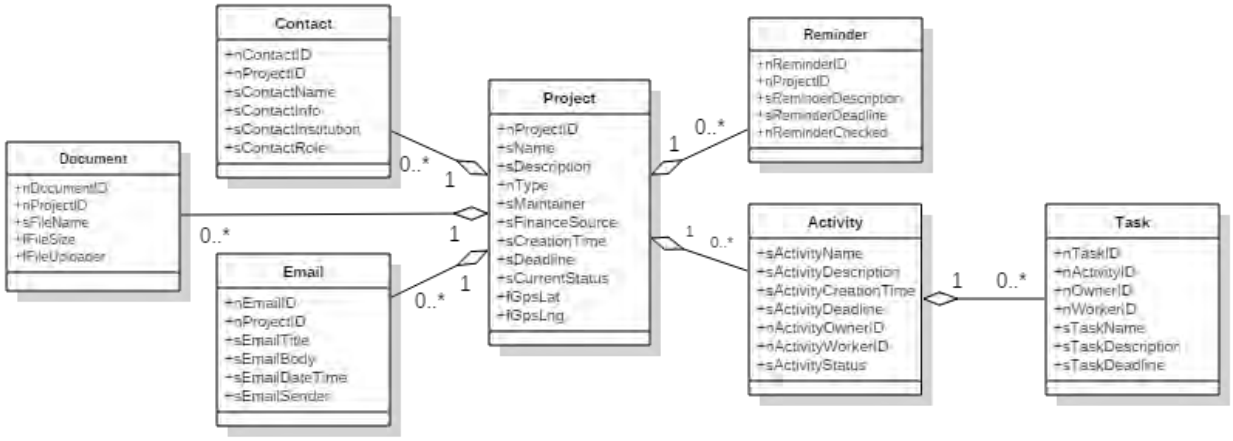


Image 2. Web application architecture

A. Project portfolio

Every project is described with several attributes which can be defined during application usage - project title and description, type, category, deadline, finance sources, maintainer, status, priority, geographic location and visibility type. The class diagram with main project attributes is shown in image 2. These attributes are defined through web form during project creation. After information is saved, the new project portfolio is available.

Project type and category – during testing application, a feature that enables filtering has been implemented in application because of the need for project list wich contains projects which have something in common. Some of the key filters are project type (local, county, regional, state..), project categories (IT sector, sport, culture, education...), project priority (low, medium, high), project value and project deadline.

Deadline - Each project might have a specific deadline, or require completion with a certain time frame. While it might be difficult to know the exact moment a project will finish, allowing additional time can give project members additional flexibility. Project deadline is defined while creating a new project through datetimepicker and can be changed in edit project form.

Every project has certain phases of development known as life-cycle phases [13]. Dividing a project into phases makes it possible to manage it in the best possible direction. There are different approaches to project phases – some claim there are 3 phases to a project while others say it's 5 [11]. Projects in presented PMIS could have following phases:

Project planning phase – project is named, defined and treated as an idea. The goal of this phase is to determine whether or not the project can be completed, is it feasible, who are possible partners, who are in the project team and what should result be.

Project execution - project teams are performing their tasks and activities defined in the planning phase. Most of the project documents and information are collected in this phase.

Project closure – a project is formally finished in this phase. After project tasks are completed and the project goal is reached, the project can be moved to archive with finished phase turned on. Project phase is shown in project table as a flag indicating current phase. After project closure project phase is updated to inactive phase and project is moving to project archive – tab with all finished projects.

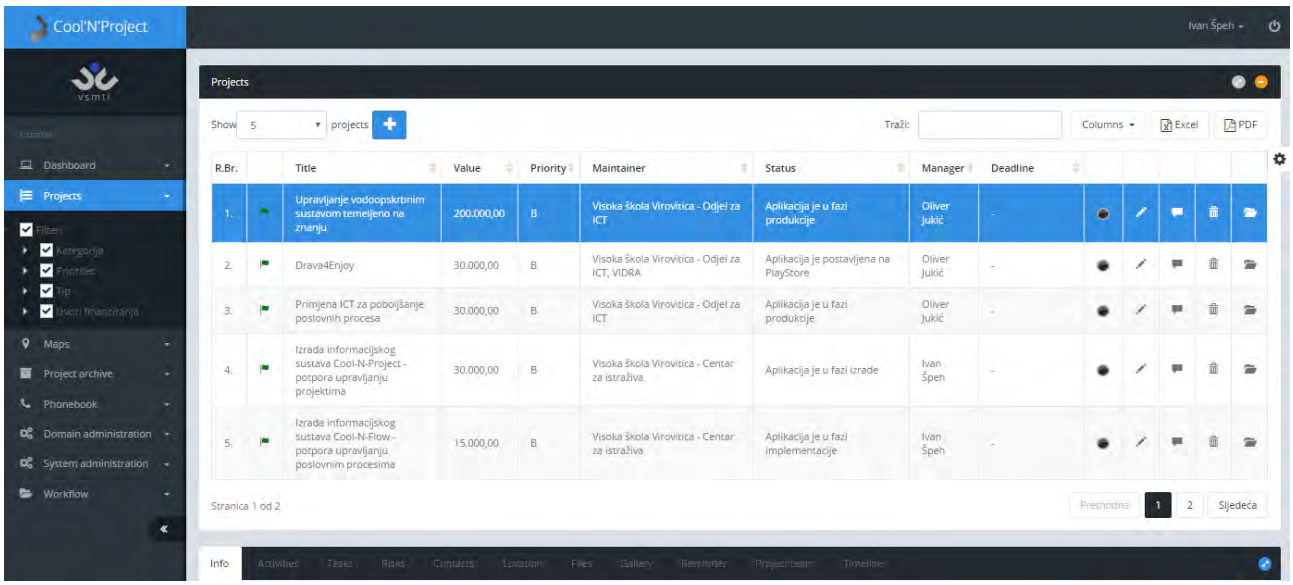


Image 3. Application GUI

Project visibility determines what kind of project team will be. There are three types of visibility – public, custom and private. If public visibility is selected, every member of the domain will be part of the project team. When creating a project user can choose specific users or entire department if custom visibility is chosen. And finally, the user can have his own project for a private purpose.

After choosing project visibility, project manager and project members can be selected from dropdowns. A specific user role is assigned to every project user, where the role can be generic or custom from different project management functions (administration, planning, design, data management, technical support, communication marketing) [14].

Project location – geographic latitude and longitude are used for showing project(s) on google maps as shown in image 4.

In the current version of this software a risk management is not included, but certainly it will be included in next generations.

B. Project tabs

Project portfolio is made of tabs containing a different type of content which is populated during project execution phase. Each container listed below provide different type of project information:

- Project Info – quick overview of basic project information such as current status, description, value, finance sources, maintainer, project team and deadline;
- Activities – every project has fragments (or phases) which need to be completed for successful project closing. These fragments are called project activities.
- Tasks – tab containing decision-making support where project manager and project team can assign tasks to each other
- Contacts – phonebook of particular project
- Location – geographical presentation of the project with google maps as background
- Files – document repository of project
- Gallery – images and photos organized by project activities
- Reminders – if some meeting, task or some other project obligation it could be useful to set a reminder to notify project team about it.

These containers together with project row in the main table make project portfolio. Project tabs can be expanded with more option, such as Gantt diagrams or finance analyze, which will be included in next version of the application.

During the project execution phase, decision-making support through Workflow Management System (WfMS) in project activity tab is available. Workflow system partially or completely automates of a business process, in whole, or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules [15]. Work procedures that have same execution flow can be

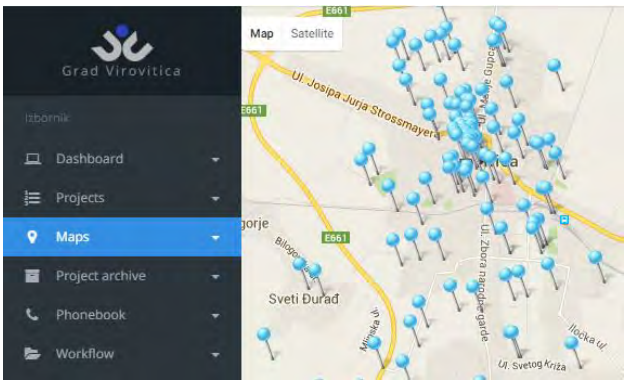


Image 4. Projects on google maps

described as unique task type in the system. Configuration of task types is stored in configuration database with following parameters:

A number of phases – task type can have more than one execution step – task phase. Every working procedure must have at least one phase.

Owner – after defining activities, the project manager can assign a task to project members after which he becomes task owner. Worker – every phase has assigned user who is responsible for executing it. The worker can be a specific member of project team, or alias name, for example, head of the department.

Deadline – a final time that a task must be completed by. Indicators of deadline are shown as progress bars. The closest the deadline is, deadline bar is turning his color into red. If deadline of the task has expired, email can be sent at specified time of a day.

Actions – task phase has predefined actions that worker can choose as the result. According to chosen action, the task is closing or moved to next phase, next worker. Typical types of worker actions are *approve* – task is moving to next phase (or finished if the last phase is active), *reject* – worker is rejecting task, *forward* – worker is forwarding current active phase to another worker, *cancel* – owner can cancel the task or *postpone* - owner can reschedule tasks.

So, in the project task tab project manager can assign a task to specific project activity and specific project member. Task creation and deadline is visible in project calendar and in worker/owner table of tasks.

C. Implementation

The process of implementation Cool-N-Project software in local government units can be divided into 5 basic phases: analyze, design, build, realization and support. After identification of processes in units, functional specification containing visual identity, hardware requirements and database design is published. According to functional specification, programming team developed an application which was published in cloud surround. Using cloud surround additional costs for hardware infrastructure and hardware maintenance at each local government unit are avoided. Once the application is published, difference tests are performed. According to test results, upgrades are made on the application. During realization phase, users (local government unit employees) are trained and additional functions are added in application functionality. Every local government unit inside the system is presented as a domain with a unique identifier and assigned administrator, users and default settings.

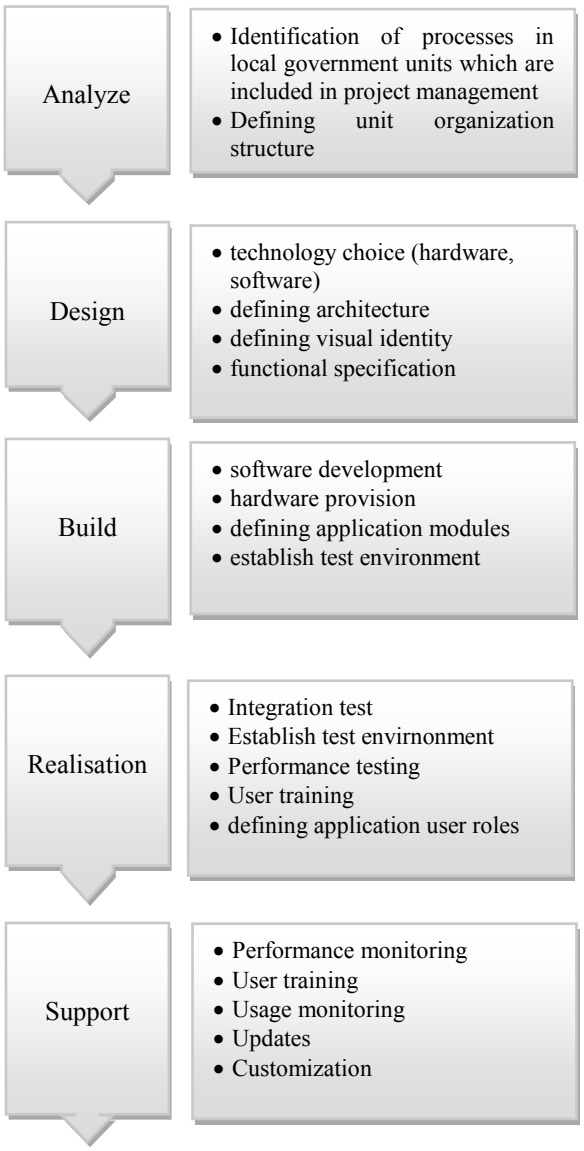
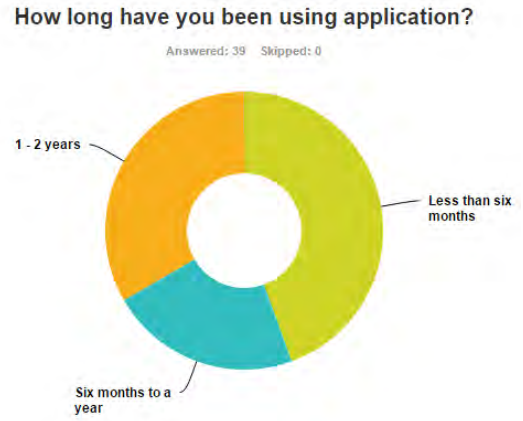


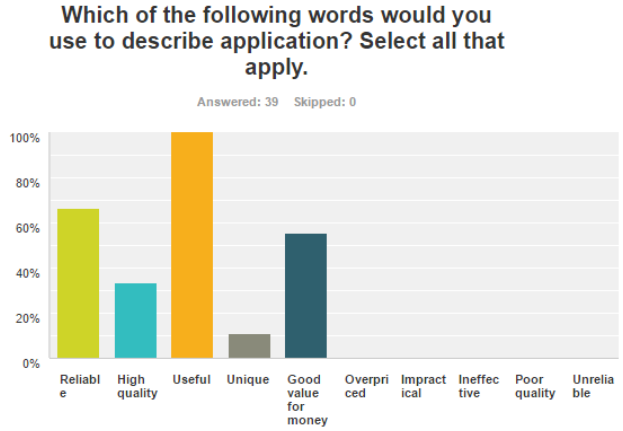
Image 5. Phases of implementation

D. User satisfaction and PMIS success

Results of the short questionnaire for the measurement of user satisfaction and project management improving are presented. Respondents are employees of local government units from Virovitica-Podravina county. The purpose of the questionnaire is to see to what extent does the application improve project management and how users can describe the application.



Development of described application started in 2014. The first version was published after 6 months in several local government units in Virovitica-Podravina county.



Only 10% of respondents said that this type of application is unique. As mentioned in the introduction there are many other web applications which can provide similar functionality, but in most cases they are not affordable to local government units, so users are satisfied with the price. All respondents described the application as useful which proves that project management is improved.

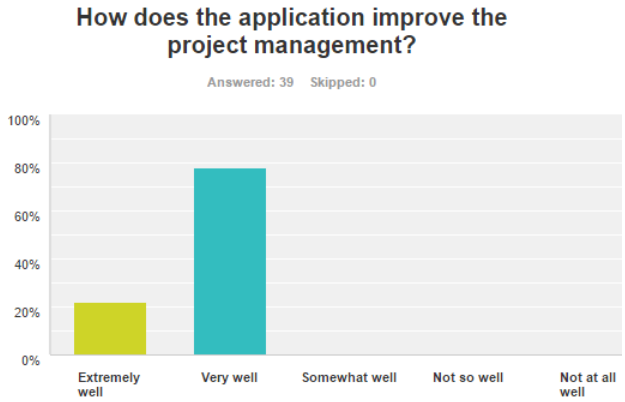
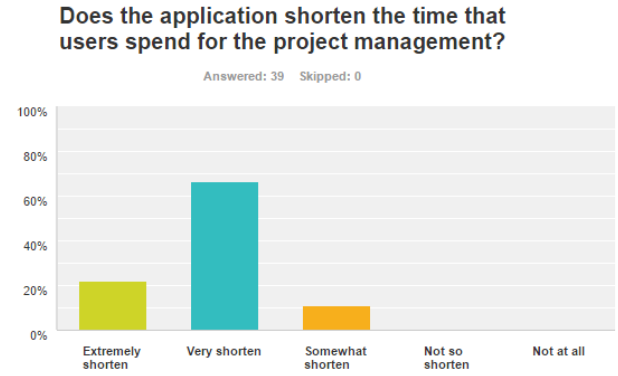


Image 8. Projects on google maps



In the literature review of project success, one of the critical success factors of the project is time [7]. To improve time efficiency of project management several features are developed or included as a plugin in the application. More than 60% of respondents confirmed that using this type of application time efficiency is improved.

Users are asked to write down additional functionality which cannot be found in current versions. These are submitted changes:

- Integrated SWOT analysis tools
- Timeline
- News feed from the projects
- Enrollment of users in multiple projects
- Improved project portfolio
- Improved project search
- Risk management

IV. CONCLUSION

In this paper, a web-based application aimed for project tracking and monitoring used in local government units is described. The need for this type of application is derived from the fact that local government unit has multi-project environment characteristics. In such organizations management is faced with multiple challenges and problems in project execution phase which can be solved or, at least, less complicated, using this kind of application. There are several other reasons listed in the paper which explain why multi-project environment should use PMIS tools. In the next version, more features, like Gantt diagrams and finance reporting will be included.

V. REFERENCES

[1] Elonen, S., Artto, K. A., "Problems in managing internal development projects in multi-project enviroments", Internal Journal of Project Management 21, vol 6, 2003., 395-402

[2] K. Rava, "Multi-Project Management in Organization", Information System Project Management, 14. Lecture, 2014.

[3] Raymond, L., Bergeron, F. "Project management information systems: an empirical study of their impact on project managers and project success", International Joournal Of Project Management 26, vol 2. 2008. 213-220

[4] Patanakul, P., Milosevic, D., "The effectiveness in managing a group of multiple projects: factors of influence and measurement criteria", International Journal Of Project Management 27, 2008. 216-233

[5] Marjolein C.J. Caniëls, Ralph J.J.M. Bakens, "The effects of Project Management Information System on decision making in a

multi project environment". International Journal Of Project Management, vol 30, 2012., 162-175

[6] Kahura, M., "the Role of Project Manangement Information Systems towards the Success of a Project:The Case of Construction Projects in Nairobi Kenya", vol 3, 2013.

[7] Lee, S. K., Yu, J.H., "Crtitical Success Factors for Project Management Information System in Construction", KICEM Journal of Construction Engineereing and Project Management, 2011

[8] http://www.pmi.org, accessed february 2016.

[9] Havelka D., Rajkumar, T.M., "Recovering troubled projects: prescriptions for sustained recovery", Issue in Information systems, vol 7, 2006

[10] Karim, A.J., "Project Management Information Systems (PMIS) factors: An empirical study of their impact on project management decision making (PMDM) performance", Research Journal of economics, business and ICT, vol 2, 2011 ,

[11] http://project-management.com/top-5-project-management-phases/, accessed February 2016.

[12] White, D., Fortune J., "Current practice in project management an empirical study", International Journal of Project Management,

[13] Project management: A System Approach to Planning, Scheduling, and Controlling John Wiley

[14] K.Burnett, The Project Management Paradigm, 1998.

[15] E. Zembia i I. Obłąk, "The Survey of Information System in Public Administration in Poland,", Interdisciplinary Journal of Information, Knowledge, and Management, vol 9, 2014., 31-56

[16] http://project-management.com/software/top15/,accessed February 2016.

ANALYSIS OF ICT USE IN PRIVATE ACCOMMODATION RENTALS IN CROATIA

Ljiljana Zekanović-Korona
Department of Tourism and Communication Studies
University of Zadar
Franje Tuđmana 24i
ljkorona@unizd.hr
Jurica Grzunov
Department of Tourism and Communication Studies
University of Zadar
Franje Tuđmana 24i
jgrzunov@unizd.hr

Abstract - The key factor for national growth and evolution is the development of economy, which in these modern times is causally linked both with the society of knowledge and to digital economy. Likewise, economy based on innovation factors is the core of the development of a society of knowledge. On the other hand, the key factors of ICT growth required for such development are hardware, software and networks.

In present times, tourism is the sector in which ICT has seen its outstanding use. Tourism and innovative digital technologies are in synergy to create new and interesting business opportunities for all stakeholders. Therefore, needs and requirements that emerge from the tourism industry encourage new ICT solutions, whilst some of the information and communication technologies are recognized as vital instrument in overcoming the crisis that has affected the tourism sector. The results of many studies showed a constant large-scale increase in the number of online accommodation bookings.

In this paper, the usage of advanced ICT for accommodation booking purposes was analyzed. The target group of this research were the owners of private accommodation in Croatian tourism. The analysis of the research results refers to an online survey conducted through social networks in which 25 closed-ended questions sought to determine which of the online services for accommodation booking and host rental support were most frequently used and to determine the way in which they are managed and to what extent the accommodation owners are (dis)satisfied with the individual components of these services. The results of this research showed that, regardless of age or the level of education, apartment owners in Croatia frequently use ICT for accommodation booking purposes.

Keywords: ICT, Tourism, Online booking, Accommodation, Digital economy, E-business.

I. INTRODUCTION

ICT is a driving force and has a historically important role in the fifth technological revolution. Today, the availability of communications is a right of every individual, and technology is a necessary tool that enables it. The technology is based on broadband Internet, mobility and cloud computing. Innovative solutions commonly called "Technology for good" are the key factors for networked society. These products, i.e. solutions and services, are covering a whole spectrum of needs, and the process of networking is in full swing. The well-being of people, society and business is induced with the help of technology that becomes the basis of smart social development. It is estimated that by the year 2020 we will have had more than 50 billion networked devices i.e. as many as six devices per every connected person.

Many new information and communication technologies have seen their first application in tourism, which was also the case 60 years ago when the development of global distribution systems (GDS) started. The relationship of tourism and ICT has always been considered mutually stimulating. The tourism industry constantly generates some innovative applications that demand finding new ICT solutions.

Today, during the global crisis that affected the tourism sector to some extent as well, some of the ICT technologies are considered to be crucial in overcoming the difficulties encountered by entrepreneurs in the tourism sector. IPK International predicts a large yearly increase in the number of online bookings. Also, it is expected that the demand for online virtual business meetings and videoconferences is to increase due to (as analysis indicated) the great decline in demand for business travel and the MICE sector (Meeting, Incentive, Conference, Event). [1]

In times of economic crisis most of the tourism companies are advised to control and reduce costs, reduce sales prices and use more aggressive informational campaigns regarding their new offers by using e-marketing. To increase e-marketing and e-sales initiatives one has to use innovative ICT solutions, i.e. CRM systems and RM systems already used by large hotel chains and international travel agencies. The primary strategic objective is to increase sales with the help of ICT, i.e. online distribution and marketing of the area of tourism in which everyone can see a chance for success.

Despite wars, terrorist attacks and a growing fear of pandemics as well as unstable economic conditions, people around the world are travelling more than ever. International tourist arrivals grew by 4.4% in 2015 to reach a total of 1,184 million in 2015 [3].

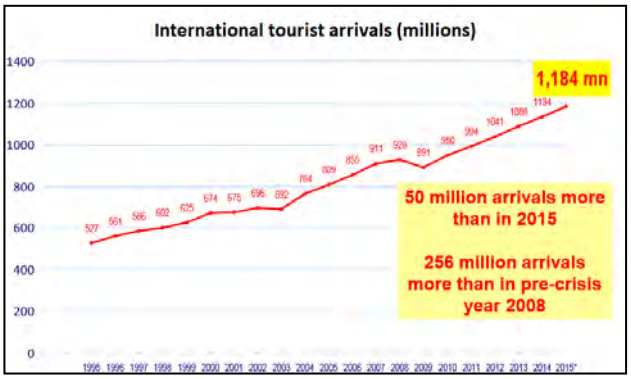


Figure 1. International tourist arrivals in millions from 1995 to 2015 (Source: UNWTO)

The global economy slowed in 2015 but a growth is expected in 2016, providing a solid basis for international travel and tourism [4]. Also, despite economic challenges, online bookings in Europe expanded by more than 7% yearly from 2013 to 2015 and will continue to climb at a similar pace through 2016 [5].

According to the research made by Croatian Institute of tourism in 2014, internet was ranked second on the list of 8 as a source of information which tourists had been using while preparing their vacation and also during their stay in Croatia in the summer of 2014. The Internet was used by almost 30% of respondents. 75% of guests made a reservation over one or more booking services before arriving to the destination in 2014. 72% of guests booked their accommodation in advance and 16% of them booked their transportation in advance. 73% of guests who booked accommodation in advance used the Internet and 71% of guests who booked transportation in advance also used the Internet. 43% of all respondents booked their accommodation by calling the accommodating facility directly, but they used the Internet as a source of information. [7]

II. Online Bookings and Travel Industry

In 2015, online bookings continued to boom with a further 7% rise to a 66% share in all travel bookings worldwide but there are signs that they could be reaching the saturation level of about 70% in mature markets.

The global market share of travel agency bookings is now more or less stabilised at 24%, in accordance with World Travel Monitor figures. There has as well been a sharp increase in bookings by smartphones this year. China (10% share) took first place in the share rankings, followed by the USA (7% share) and Japan (5% share). Social media channels such as social networks, review portals, blogs and forums have also become very popular with international travellers, and about 70% of them are active social media users [3].

It is believed that social media now influence nearly one fourth of all international trips, especially travellers' choice of destination and accommodation. The number of active Facebook users has now reached an estimated 1.5 billion people around the world while other social platforms such as Twitter, LinkedIn, Google+, WhatsApp and TripAdvisor continue to grow steadily in many countries. More and more international travellers are also actively posting their holiday experiences on social media, with the percentage now rising to nearly 50%.

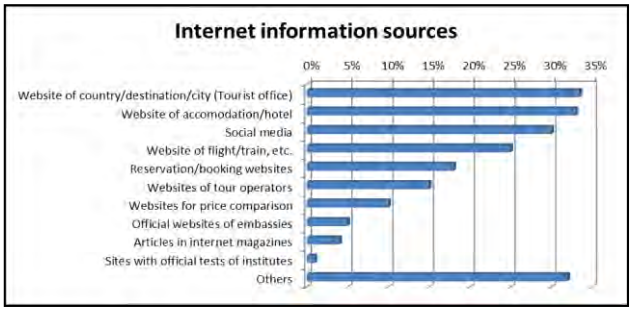


Figure 2. Internet information sources (multiple answers possible) (Source:World Travel Monitor 2014, IPK International)

Technology-based „sharing economy' firms are increasingly penetrating the travel industry by offering consumers far more individualised choice than traditional suppliers and are disrupting the established business model by using private resources for lower-cost commercial activities. Traditional travel suppliers need to respond quickly and innovatively if they are to stop losing market share. US-based accommodation provider Airbnb and the ride-sharing app Uber are just two of the newcomers that have made an ever bigger impact on the travel industry. The dramatic and controversial growth of so-called „sharing' firms such as Airbnb and competitors such as Homeaway, Wimdu and others has seen them grab a significant share of the accommodation market in recent years. Over the last five years, the so-called „para-hotellerie' segment, covering different forms of private accommodation ranging from holiday homes, bed & breakfast and hostels to staying with friends and relatives, has increased by 35% to reach a 40% share in the overall world outbound accommodation market.

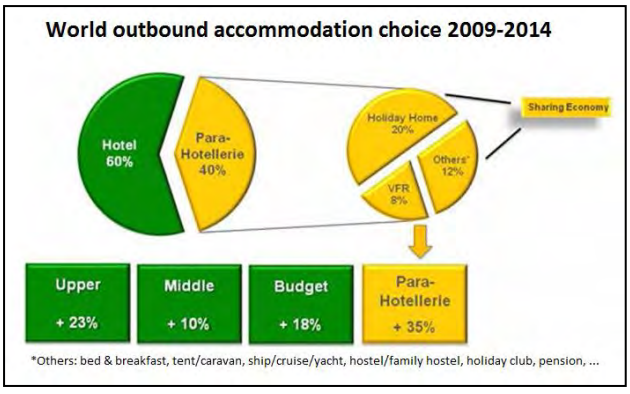


Figure 3. Rising market share of Para-Hotellerie (Source: World Travel Monitor 2014, IPK International)

In the last few years, it has become obvious that sharing economy poses a major challenge to the traditional travel business. The basic idea is to use available resources more efficiently, whether this is accommodation, car transportation or other activities. "The sharing economy is absolutely disruptive and is changing how we use things. The impact on travel is

huge,"[1]. In the accommodation sector, for example, Airbnb now offers about 1.2 million accommodations in 34,000 cities worldwide, making it one of the world's ten largest "hospitality brands" alongside long-established leading hotel companies. If compared to traditional travel companies, these sharing economy rental companies have completely different business models. Essentially, the models are technology platforms that market privately-owned assets or resources such as accommodation or cars on a commercial basis. For consumers, they are attractive because they offer lower prices, better accessibility and flexibility, ease of use, and are more user-focused, including transparency and interactive communications. From a business perspective, such firms are asset-light because they own the sales platform but not the resources. This means they can operate at low cost and also pass on legal liability to the providers who are private contractors. Meanwhile, one interesting new trend is that other start-ups are now being launched to fill in gaps in the sharing economy, such as handover of accommodation keys and cleaning services. However, hotel groups are starting to respond to the challenge of Airbnb by broadening their offers. For example, some chains are adding apartment options and putting more emphasis on their „local' status. In addition, they are becoming more transparent by incorporating customer feedback, using social media reviews and improving customer communications. The emergence of these „sharing economy„firms is also a response to the increasingly extreme expectations of consumers who are spending more time on social media. The rapid growth of the web and mobile-based companies in the travel market reflects the relatively slow reaction of the traditional travel industry to the internet and social media channels. Although most travel companies have responded in recent years to digital opportunities, some focus purely on generating bookings and do not have a full social business model that, in addition, can create customer loyalty.[2]

III. QUESTIONNAIRE METHODOLOGY

The aim of the research was to analyze the way the apartment owners in Croatia use advanced information and communication technologies for accommodation booking purposes.

The survey was conducted among a quasi random sample of 160 respondents from the group of 2000 registered apartment owners in wider Zadar region in Croatia (111 females, i.e. 69.38% of the sample and 49 male respondents, i.e. 30.62% of the sample). The questionnaire, which consisted of 25 questions (mostly closed type with multiple choice answers), was distributed via paper mail and online via social networks and e-mail to apartment owners whose contact information were publicly displayed online. The results obtained by this research are not as statistically relevant as much as they represent a good preliminary base for further researches.

[1] Jason Clampet, co-founder of New York-based travel industry intelligence company Skift.

[2] Albert Brenner, managing director of the Düsseldorf-based Digital Transformation Group

IV. RESULTS AND DISCUSSION

The analysis of the results shows that the majority of subjects who rent apartments in Croatia are aged 26-35 years (31.25%), followed by those of age 36-45 years (25.63%), then 46-55 years (16.87%), 18-25 years (13.75%). This suggests that the apartment owners in Croatia are mostly young and middle aged people. After examining the distribution of respondents by TRI index (below) it becomes obvious that the younger subjects are more willing to use new technologies that enable an increased activity in tourism sector. The educational structure of the respondents is as follows: completed only primary school - 2.5% of respondents, a high school - 35.63%, baccalaureate - 15.62%, college education - 25.63%, Master degrees and PhD - 9.38%. The apartment owners completed secondary education mostly and they are followed by university-educated respondents. Further analysis shows that most of the respondents (apartment owners) are employed in public sector (26.25%). For vast majority of such public employees, accommodation renting probably represents an additional activity which serves as a supplement to their primary income, while only 18.13% of owners are unemployed and renting is their primary activity that generates income. The sample also includes about 12.5%, of independent entrepreneurs who own apartments and about 21.88% of owners are employed in the private sector. And last, there are 13.75% of the retired and 7.5% of students in the sample who also own apartments. From the above data it can be seen that almost 50% of all respondents are engaged in renting out apartments as supplementary rather than primary activities. If you add pupils, students and the retired to that number then there is a total of 69.38% of owners who are renting apartments as additional (rather than primary) activity. The analysis of the monthly incomes of households, excluding the renting income, shows the following: 10% of the owners have more than 12000 kn income, 21.88% of them have between 8001 and 12000 kn income, 37.5% of them have 4001-8000 kn, 20% of the owners earn less than 4000 kn, and 10.63% of the owners have no other income. (Figure 4)

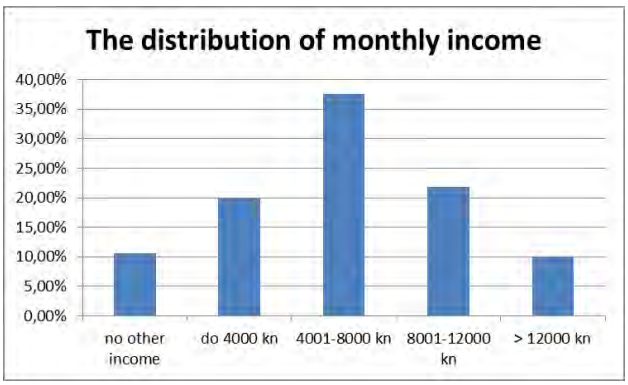


Figure 4. The distibution of the respondents monthly income (Source: processed by authors)

It is evident that most apartment renters have average monthly income somewhere around the national

average monthly salary in Croatia so they need this additional activity to achieve a better standard of living. In the questionnaire there was a set of questions from which the TRI-index (*Technology Readiness Index*) was calculated for every respondent. Technology Readiness Index (TRI) is used to calculate the willingness of respondents to adopt and use new technologies. This index is based on the examination of feelings that the respondent expresses for new technologies. Components, i.e. feelings covered by the TRI-index are: optimism, innovativeness, discomfort and insecurity. The above feelings are measured by the Likert scale as a series of statements in which the respondents can assess to what extent they agree or disagree with a given statement on a 1 to 5 scale. The obtained result is the assessment of how much a particular individual or respondent is willing to accept new technologies. Readiness assessment includes the evaluation range of -20 to +20. If the above assessment scale is divided into five equal parts, it is possible to distinguish between the following categories of respondents:

1) laggards (from -20 to -13)
2) paranoids (from -12 to -5)
3) skeptics (from -4 to +4)
4) pioneers (from +5 to +12)
5) researchers (from +13 to +20). [6]

Therefore, if the level of TRI-index is higher, the tendency of respondents to the acceptance and use of new technologies is greater. The sample did not include any of those who are reluctant to use new technologies (laggards), and only 1.25% of respondents have paranoid feelings when they use new technologies, but all of them aged between 56 and 65 years. Also, according to TRI index, among the respondents there are 24.38% of skeptics and 55.63% of pioneers when it comes to using new technologies. And, last but not least, almost 18.75% of all respondents are in the category of researchers, which means they are feeling very comfortable while using new technologies and are usually the first ones to buy and try them when they come to the market. The total number of pioneers and explorers in the sample sums up to 74.38%, which indicates a strong positive association between the usage of ICT and tourism for the purpose of renting accommodation. From the above data it is evident that the respondents are mainly employed people and that, in addition to their regular jobs, in order to perform this additional activity it is necessary for them to use the assistance of technology that saves them time and accelerates the exchange of information. Today's potential guests want immediate answers to their questions, and in present time that is almost impossible without the help of the latest information and communication technologies.

As the number of guests searching for accommodation information online increases on a daily rate, it becomes very important for apartment owners to increase the volume of their online presence. This can be done in many ways, e.g. by using online ads and banners, social networks, private websites, etc. Therefore, a few questions in the survey addressed these topics, and the results are as follows: 51.25% of the respondents already own a website containing information about their apartments, while 26.25% of them are in the process of creating one or are planning to do so hereafter. Only 21.25% of the respondents do not own a website and are not planning to get one. One of the main reasons for this could lie in the fact that the number of available online services for advertising and rental mediation is rapidly increasing and some of the apartment owners feel that it is easier and more profitable to invest in a good online advertising campaign via some of these existing services than to spend money on their own website. Also, an increasing number of apartment owners are actually just sharing their own apartment (a part of the house or perhaps just a bedroom), and to do so it doesn't make much sense to own a private website when they can use e.g. Airbnb. The null hypothesis associating the TRI index with owning a private accommodation website was tested, and the result of the chi-square test was 2.4659 (d.f. =4, p=0.05), which means that the null hypothesis can be accepted. Hence, there is no association between the level of acceptance of new technologies and owning a website which advertised apartments (or planning on creating one). Also, around 64.63% of renters who own a website stated that they have created it themselves, and almost all of them belong to the group of pioneers and researchers according to their TRI indices, i.e. they can be considered advanced users of new technologies. Furthermore, the advertising of apartments on social networks was also analyzed in the survey, and 30% of owners responded negatively, i.e. they do not advertise on any of the social networks, while 42.5% of them advertise their accommodation capacities on Facebook. Also, 39.38% of the respondents do not advertise their apartments on one of the domestic online classifieds, and among those who do, 25% of them use Apartmanija.hr and 5% use Njuskalo.hr. On the other hand, 36.875% of respondents do not use foreign online classifieds to advertise their apartments (mostly the ones who own their private website), while 10% of renters advertise their apartments on foreign online classified section. Despite high fees (around 30% of all rental income), almost 37% of the respondents entrusted the business of renting their apartments to one of the local tourist agencies that intermediate in rental accommodation. Because of the local tax legislation, all of the registered apartment owners must register their guests within 24 hours upon their arrival. In past times, this was done by filling out the guestbook in paper form manually and then taking it to the police station and tourist board. In recent times, many various online applications have been released thus making this task easier and quicker and thereby saving time and money for the apartment owners. These applications are used by 75.63% of the respondents, while 18.88% of them still fill in the guestbook in paper form. Despite the high fines and the available technology that makes it really easy, there are still apartment owners (about 5% of them) who do not register their guests in any way. Among those who do use online applications for registering guests, 95% of them consider these applications to be very simple and user-friendly. At the beginning of 2016, the Croatian national tourist board in association with Croatian Ministry of Tourism released a new online service for the registration of guests called eVisitor, which is now mandatory for all apartment owners at the state level. Although most of the

respondents do not advertise in foreign online classifieds, almost all of them use foreign online accommodation rental booking services such as Booking.com, Airbnb, TripAdvisor, etc. Majority of the respondents use more than one of these accommodation rental booking services at the same time, thereby improving their chances of renting their apartments to guests from many different countries and making the most of the pre-season and post-season rental periods. While online advertisements are paid in advance and are generally rather expensive, considering the fact they usually do not guarantee any bookings, aforementioned online accommodation rental booking services operate on commission basis and make it a safer and more attractive option for apartment owners. Most frequently used online accommodation rental booking services in the surveyed sample were Booking.com (83.75%), Airbnb (55.63%), TripAdvisor (50.63%) and HouseTrip (36.25%). Furthermore, 76.25% of the respondents stated that they have set up and kept maintaining their online listings themselves, and 17.5% of them with the help of their family and friends. The analysis of general customer satisfaction with the online accommodation rental booking services shows that the respondents have best assessed Airbnb (avg. score of 4.06, st.dev. 1.17), followed by HomeAway (avg. score of 3.68, st.dev. 1.22), TripAdvisor (avg. score of 3.62, st.dev. 1, 02), Expedia (avg. score of 3.62, st.dev. 1.16), and Booking.com (avg. score of 3.43, st.dev. 1.32). (Figure 5)



Figure 5. General satisfaction with online accommodation booking services (Source: processed by authors)

Table 1 shows host service fees ratings (1-high service fee,...,5-low service fee)

From the data given in table 1. it is easy to notice the difference in the respondents' opinions regarding the amount of service fees that each of these online booking services charges to apartment owners. Booking.com charges 15% of the total rental cost, therefore has the lowest rating average (2.53), while Airbnb (3.76), Tripadvisor (3.64), etc. charge only 3% commission fee to the apartment owners, and a certain amount to the guests for using their services.

Finally, for the two of the most frequently used online accommodation booking services, Airbnb and Booking.com, the respondents assessed some of the main components of these services, such as graphical interface, payout preferences, customer support, user-friendlines,

Online service	*AVERAGE*	*MODE*	*ST.DEV.*
Booking.com	2,53	1	1,40
Airbnb	3,76	5	1,27
TripAdvisor	3,64	5	1,29
Expedia	3,33	5	1,62
HomeAway	3,41	5	1,38
Housetrip	3,39	4	1,13
Wimdu	3,62	5	1,52
Atraveo	2,95	3	1,39
Casamundo	3,27	5	1,48

Table1. Host service fees ratings (Source: processed by authors)

mobile application, etc. From the chart (Figure 6.) that compares average ratings of individual components it is clearly visible that the average overall satisfaction with individual components is substantially better in the case of Airbnb users, especially when it comes to graphical interface, ways of communicating with guests, payout preferences, customer support and host service fees. Some of the other components of these services, such as user-friendliness, mobile application, calendar and pricing options and review system, received high ratings in both cases, which leads to a conclusion that both of these very popular online accommodation booking services pay a lot of attention to their users and their needs in terms of application usage itself.

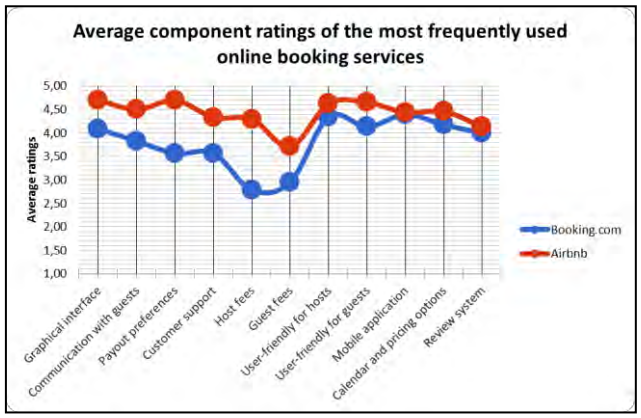


Figure 6. Average component ratings of Airbnb and Booking.com (Source: processed by authors)

V. Conclusion

As it was mentioned at the beginning of this paper, some of the ICT technologies have proved crucial in overcoming the difficulties encountered by entrepreneurs in the tourism sector, especially private accommodation owners many of whom do not have any other source of income apart from renting apartments. The ‚shared economy' model completely changed the way accommodation rental market operates, and many homeowners spotted the opportunity to generate additional income by using their surplus assets such as vacation homes, guest houses, spare rooms, etc. The easiest way to do so is by using ICT to access worldwide

tourist market. Modern technologies make it possible for practically everyone who owns some kind of accommodation to rent it without having to invest a lot of money into advertising, all that is needed to start renting (in addition to meeting local legislation) is a computer or a smartphone connected to the Internet, and a little time to set up the apartment listing details on one of the online accommodation booking websites. These online booking applications are usually very user-friendly and can be easily comprehended. Since the analysis of this research showed that the majority of the apartments owners are young or middle aged people who are comfortable with using new technologies, it is easy to conclude that this emerging new market has a bright future and that it is just a matter of time when traditional ways of rental mediation are going to become obsolete and eventually seize to exist if they do not adapt to the new technologies and ways of doing e-business.

It is visible from the data that online booking of accommodation is growing very quickly from year to year. This research shows that we should continuously work on instructions about the use of information technologies for all generations of accommodation hosts. One of the main goals of local tourist communities should be helping small accommodation hosts. ICT technologies are finding their largest change in tourism and if we want to increase our income in tourism, it is necessary that all participants in the chain of renting tourist capacities are mostly familiar with new technologies. Today it is necessary for all kinds of information to come instantly and very quickly to the users in all segments of society, especially in tourism, and it is possible only with the help of new technologies. In tourism as a growing branch in our country, there is a need for sinergy of ICT technology and the increase in the number of tourists. This research has showed that digital economy and information technologies are more and more present and inevitable in today's tourism.

VI. LITERATURE

[1] Vrdoljak-Šalamon, B.: infotrend.hr - onLine edition (2009): Turizam i ICT u doba krize.

[2] Panian, Ž. (2013.): Elektroničko poslovanje druge generacije. p.7, Faculty of Economics and Business, Zagreb.

[3] UNWTO World tourism barometer, Volume 14-2016. http://dtxtq4w60xqpw.cloudfront.net/sites/all/files/pdf/unwto_barom16_01_january_excerpt.pdf

[4] ITB world travel trends report 2015/2016 http://www.itb-berlin.de/media/itb/itb_dl_all/itb_presse_all/WTTR_ITB2016_8_Web.pdf

[5] Phocuswright's European Online Travel Overview Tenth Edition

[6] Zekanović-Korona, Lj. & Grzunov, J. (2014) Evaluation of shared digital economy adoption: case of *AirBnB*, MIPRO DE, Opatija, Croatia.

[7] Croatian Institut for Tourism: Croatian tourism in numbers,Vol.9,1/2015

Records management challenges and opportunities: An Australian perspective

A. Davies* and R. Bergami**
*Victoria University, College of Business, Melbourne, Australia
** Victoria University, College of Business, Melbourne, Australia and University of South Bohemia in Ceske Budejovice, Faculty of Economics, Department of Trade and Tourism, Czech Republic
adam.davies1@live.vu.edu.au;
roberto.bergami@vu.edu.au;

Records management practices face challenges in an ever increasing digital environment. As governments grapple with an evolving legislative landscape, the transition from paper to electronic records is critical in reducing future compliance costs and regulatory burden. Digitization and paper reduction are often cited as key initiatives to achieving this transition. Yet, governments in Australia continue to struggle in implementing these initiatives across different industries and jurisdictions. This paper examines the progress of government agencies in transitioning to electronic records management, identifying areas of success, failure, or changes that may be required in current practices. Although the focus is on the Australian environment, the findings may have international relevance, as records management is not limited to domestic transactions, especially where transnational firms are involved. The importance of this research is highlighted by the need to inform advancements in global communication options, the development of cloud-based solutions, and the adoption of industry-specific electronic records, such as in health. The conclusion is that as technology is forever changing the manner in which information is created and managed, governments must ensure appropriate reporting allows for clear determinations to be made on the cost and benefits of managing records electronically, whilst maintaining regulatory compliance.

I. INTRODUCTION

The evolution of the record, from papyrus to parchment, to bytes and bits, has fundamentally changed records management practices. Record-keepers, tasked with the management of vast archives of paper records, now face the challenge of transitioning to a digital world. In addressing the ever increasing cost of record retention, governments around the world, have directed their entities to both embrace digital records and digitize their paper information assets.

The Australian Government's Digital Transition Policy [1], outlines the digital directive for all government entities to embrace electronic records management and digitization. The focus of this policy is to improve record governance, while reducing the cost of storage and preservation of paper records. To achieve this transition, Australian government entities are authorized to use converted records authorities, allowing the destruction of many original records after they have been copied, converted, or migrated [2]

Digitization, or the conversions of paper archives, provides organizations with the capability to streamline operations, reduce financial outlays, and evolve information governance practices. Going digital, or focusing on reducing paper, can also be associated with notions of going paperless [3] or the paperless office. However, these notions have been challenged, with the claim the paperless office is a lie we have been telling ourselves since 1975 [4].

In examining the success of organisations to reduce paper and transition to digital environment, this paper, will focus examine compliance to the Digital Transition policy, to identify the readiness of Australian government organisations to manage further record evolutions, such as electronic health records [5], block chains [6], and onion routing [7].

II. BACKGROUND

The Australian records management landscape is mature, progressive, and visionary. Records management globally, has followed Australia, through the acceptance of the International Standards Organisation (ISO) 15489-1:2001 Information and documentation -- Records management -- Part 1: General. (ISO) 15489-1:2001. This standard was launched by ISO, on October 3rd, 2001 and is considered to be the first international standard on record management [8]. The genesis of ISO 15489-1:2001, however, can be found in Australia Standard AS4390:1-6, which is noted as being the first ever national standard devoted solely to records management, the foundation on which ISO 15489-1:2001 was developed [8].

The evolution of the Australian records management landscape is not, however, without its challenges. The federated nature of the Commonwealth of Australia renders certain areas of records management discernibly more difficult. One of these areas is the transition from paper records to electronic records management. The process available to Australian government entities in converting paper records to electronic format is shown at Figure 1.

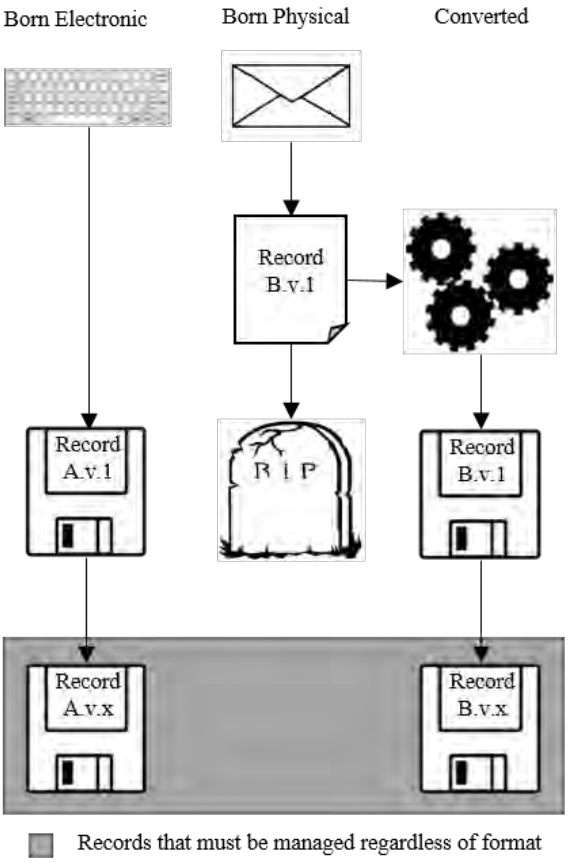


Figure 1 - Process for applying the converted records disposal authorities

The process of applying the converted records disposal authority beings with the determination of the records preservation status [2, 9, 10, 11, 12, 13, 14, 15, 16]. Once determined, the physical record is then converted, with the digital rendition becoming the true record, and the physical record being destroyed. The process, in its notable lack of complexity, does provide challenges to organisations as they consider their risk exposure in destroying any record that may be used as evidence of decision making.

Another challenge, introduced through the federated nature of the Commonwealth of Australia is that, while the Australian Government's Digital Transition Policy applies to all governmental entities, those entities within Australia can either be local, state, national or federal. This structure of government entities differs slightly from the formal governmental structure of Australia, which has three levels: federal, state and local [17]. The introduction of a national entity is a recent evolution, whereby two or more state governments may combine their sovereign responsibilities into a single intergovernmental agency (IGA), often referred to as a 'national body'.

Research into the records management implications of Australia IGAs is in its infancy; however this has not stopped researchers in determining the introduction of IGAs has forever changed the perception on how records are to be managed within multi-jurisdictional organisations [18]. The challenges facing IGAs are reflective of those facing multi-national corporations, cross government organizations and academic institutions with transnational programmes. The relevance of understanding the Australian perspective on records management is, therefore, paramount to improving the success of future digital adoptions, wherever they may occur.

Australia's unique records landscape provides researchers opportunities to investigate the implications of intra-jurisdictional and multi-jurisdictional organisations, as they continue on the journey of digital adoption. This paper, being limited in scope of its analysis, will focus on the effectives of Australian government agencies to implement conversion practices as a basis for discussing challenges these organisations may face in adapting to manage more evolved forms of digital records.

III. METHODOLOGY

Due to size limitations, the paper will focus on the digital adoption of Australian government entities, through their application of converted records authorities. The selection of Australia, as the geographical location to concentrate this research reflects the mature records management landscape, and also the vast array of governmental structures reflecting the challenges of a broad range of organisational entities.

The selection of Australian government entities, over non-government entities, reflects the focus of records management literature within Australia, and more broadly. By following the literature this paper seeks to improve the depth of analysis that can be undertaken in determining the current status of records management compliance. The selection of converted records authorities as the mechanism in which to determine how Australian government entities are adopting electronic records management is to ensure a consistent baseline, approved within the regulatory framework available to government entities.

To support the integrity of this research, this paper examines and selectively reports on the findings of Australia's audit regulators, available within the public domain, between 2000 and 2015. This will be followed by an examination of future electronic records management trends, then using the collective audit findings the paper aims to determine the readiness of these agencies to evolve their records management practices.

In concluding, this paper provides future researchers with an understanding of the potential of Australian government records management practices to transition from one form of to another. The knowledge derived from the collective findings of Australian audit regulators may be used to determine trends, compare against international counterparts and assess milestones in Australia's digital adoption.

IV. DISCUSSION

Digital adoption by Australian government entities has been a fragmented array of success and failure. In investigating the adoption of electronic records, the Victorian Auditor General's Office (VAGO) found despite the growing use of electronic records, the vast majority of agencies (84 per cent) were still maintaining records in hard copy [19]. This reliance on the use of paper over digital, may not demonstrate the most appropriate allocation of government resources. Indeed, the Western Australia audit office found it may be more cost effective to move towards digitising records sooner rather than later [20].

A detailed analysis of VAGO reports, between 2008 and 2013, found the overall status of Victoria government entities to be non-compliant against their records management obligations, as summarised in Figure 2. [21]

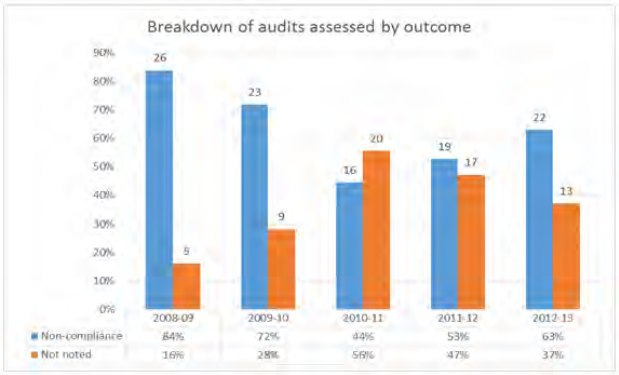


Figure 2 – VAGO audits containing noted records management breaches, 2008-2013

In assessing the overall compliance of the Victoria State Government, VAGO identified an inversion point in the compliance to records management practices within 2010-11. This pinnacle of records compliance coincides with research into observation that records management in 2010 was at a crossroads. The confidence with which the profession entered the twenty-first century has ebbed away [22].

The challenging question facing both government and non-government organisations is: do we have paper minds trying to cope with electronic realities? [23] The habitual nature of paper use was identified in audits by the ACT Auditor-General's Office where it was found "all audited agencies captured and maintained records in paper form, even those that had a formal electronic system that could capture documents in a more efficient manner." [24]

The proliferation of records, can in part, be attributed to organisations reluctance to destroy converted paper records along with the perception that paper offers protection. Organisations seeking to apply the converted records authorities must be willing to destroy their paper records or risk duplicating large quantities of records and record storage, as shown in Figure 3.

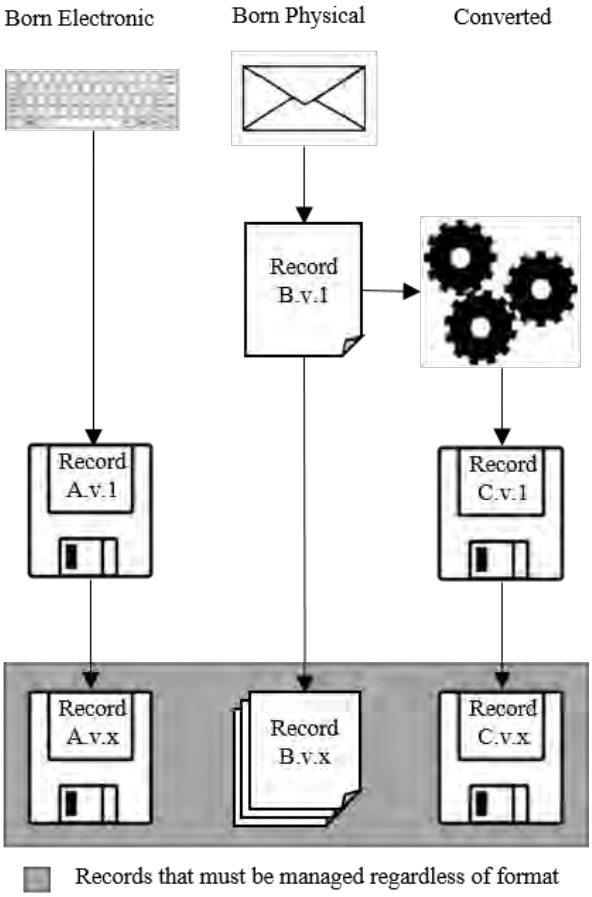


Figure 3 - Reality of applying the converted records disposal authorities

The culture of Australian government entities to capture both paper and electronic formats of records can also occur when the primary record is born digital. The Australian National Audit Office found that organisations generally advocate a simplistic 'print-to-paper' approach to staff capturing records from the electronic environment [25]. This degradation in evolutionary process has in part reversed, as cost constraints, post the global financial crisis, are requiring organisations to improve their allocation of resources.

In providing case studies of government organisations embracing the digital adoption, the National Archives of Australia (NAA) highlighted a scenario whereby $150,000 was saved in one year by transitioning to digital records management in one business area [26]. The theme of cost reduction is common in the case studies provided by the NAA. Indeed, the Australian National Audit Office itself adopted a digital approach from which physical storage costs of $80,000 have so far been avoided since almost no new paper files have been created [27].

In adopting a transition from paper, Australian government entities have demonstrated a slow and serpentine progress. Failure to embrace digital adoption may result in substantial financial outlays, such as the Victoria Department of Human Service failure adhere to the records management requirements of the Public

Record Office Victoria, resulting in an annual $1.75 million burden, on taxpayers' funds [21]. In exploring the progress of Australia government agencies, it was found that, like their international counterparts, Australian entities needed external influence, be it regulatory policy, governance oversight, or corporate failure, to promote changes in culture and operations.

Records managers, regardless of their geographical location, need to be mindful of future trends in record evolution. The history of such evolution has been one of singularities. Records have evolved in form or function, at specific times in history, as technology advanced down a single repetitive path. The rise of electronic records however, allows for a multi-facet record evolution, with different fields pulling records, and the subsequent management of records, in diverging directions.

One prominent influence, driving governments to adopt electronic records is the introduction electronic health records (EHR). The focus of EHR's, as a cost reduction mechanism, is prevalent within the literature, given the inflationary trend in the US$1.9 trillion spent on US health care in 2005, such savings could be quite significant [28]. The continued digital adoption within the health industry has many researchers noting electronic records are shaping up to be the future of health care [29].

The health industry is not, however, alone. The financial industry, through the potential of block chains records, is now faced with the realisation that electronic crypto-records may be better than currency, with research citing its promise to lower transaction costs, transform developing economies and generally reshape the financial system [30]. As communications technologies bring the world closer together, concerns surrounding data sovereignty, information flows and privacy are being discussed with greater frequency.

The evolution of the internet, cloud services, and mobile connectivity, has allowed organisations to store records in geographically diverse locations. Transmission and the privacy of records held in geographically dispersed locations, has raised its own set of challenges. Privacy and anonymity now feature in main stream conversations regarding the transfer and collection of records. Onion routing, a means of communicating anonymously, is one potential technology that could become main stream considering over half a million people use the onion-routing network Tor, every day" [7].

These advancements equally have the potential to influence the evolutionary direction of records. In embracing future record evolutions, the records profession may need to acknowledge the time of Electronic Record and Document Management (EDRM) solutions, is perhaps over. Technologies continue to come and go. If the records management profession desires to retain its rightful place in organisations' information governance structures, it will need to evolve alongside their primary responsibility: the record.

As organisations embrace digital adoption, the role of international standards such as ISO 15489 will continue to fade. Research into the appropriateness of ISO 15489; conclude the concepts contained in that standard were written for Records and Information Management (RIM) professionals in their role as principal record-keepers [31]. In the new world of digital records, however, RIM professionals will no longer be the principle record-keepers and as such new and innovative standards will need to be developed and endorsed.

The potential for record evolution is vast. The records profession, as observed through the outcomes of Australian government entities, is progressing towards a digital world. Digital technologies will continue to advance the way in which records are created, captured and managed. Records professional, wherever they might be, will need to adapt to a more technical environment, embrace change and be willing to assess and discard past traditions and old practices. The future is bright, though as it is embraced, vigilance is required to ensure a continuum of digital adoption.

V. CONLUSION

Digital adoption within Australian government entities is alive and well. Investigation into the findings of Australia's audit regulators demonstrated a progression away from paper. The challenges identified in this progression focus on cost as the main driver behind reducing paper, and the rate in which progress is being adopted. Transitioning away from paper is not just about cost. Electronic records provide organisations the freedom to unshackle themselves from physical locations, to federate their information assets, and also to compartmentalise security regimes.

In retaining some of their physical attributes, electronic records, continue to separate content from context. As records evolve through the use of block chain, or other technologies, records professionals will need to adapt to records without content. The success of future transitions will rely on organisations understanding the move from paper to electronic records is not just about conversion and deletion. Organisations, such as those in Australia, need to manage the transition by enhancing their operations to remove paper based processes, or enhance systems to no longer require paper input.

Electronic health records, block chains and onion routing are just some of the areas in which records professionals should become and remain aware. The literature on these topics outlines their advantages, but these will require different and unique forms of record management. The future of records is unknown. In a world where the tradition of records professionals to migrate vast quantities of content to solutions such as an EDRMS may no longer be required. Records professionals may need to accept the future of federation records management practices.

The records managers of tomorrow will be different to those of today. Technologies continue ascension over our daily lives has forever changed where, when, and how records are created. I the world of tomorrow it will be for

records managers to find a path through the chaos, to the sunlit uplands of democratic transparency [32].

REFERENCES

[1] National Archives of Australia, Digital Transition Policy, Australian Federal Government, viewed 04 February 2016, http://www.naa.gov.au/records-management/digital-transition-and-digital-continuity/digital-transition-policy/index.aspx

[2] National Archives of Australia, General Records Authority 31: *Destruction of source or original records after digitisation, conversion or migration*, viewed 04 February 2016, http://www.naa.gov.au/records-management/agency/keep-destroy-transfer/general-records-authorities/GRA31/index.aspx

[3] Z. Liu and D.G. Stork, 2000. Is paperless really more?. Communications of the ACM, 43(11), pp.94-97, 2000

[4] D. Casey Flaherty and C. Lovato Digital signatures and the paperless office. Journal Of Internet Law, 17(7), pp.3-12, 2014

[5] N.G. Weiskopf and C,.Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20(1), pp.144-151, 2013

[6] G.W. Peters and E. Panayi. Understanding Modern Banking Ledgers through Blockchain Technologies: Future of Transaction Processing and Smart Contracts on the Internet of Money. Available at SSRN 2692487, 2015

[7] J. Feigenbaum, A. Johnson and P. Syverson. Probabilistic analysis of onion routing in a black-box model. ACM Transactions on Information and System Security (TISSEC), 15(3), pp.1-28, 2012

[8] J. McLeod. Assessing the impact of ISO 15489-a preliminary investigation. Records Management Journal, 13(2), pp.70-82, 2003

[9] WA State Records Office, Disposal Schedule for Source Records, viewed 04 February 2016, http://www.sro.wa.gov.au/sites/default/files/gda_sourcerecords.pdf

[10] Public Record Office Victoria, DA 2159: PROS 10/01 G1 Guide to the GDA for Converted Source Records, viewed 04 February 2016 http://prov.vic.gov.au/wp-content/uploads/2015/01/1001G1v1.3.pdf

[11] ACT Territory Records Office, Standards for Records Management – No 9 Records Digitisation and Conversion, viewed 04 February 2016. http://www.territoryrecords.act.gov.au/__data/assets/pdf_file/0003/434631/Guideline-No-9-Records-Digitisation-and-Conversion-February-2011.pdf

[12] Tasmanian Archive & Heritage Office, DA 2159: Disposal Schedule for Source Records, viewed 04 February 2016, http://www.utas.edu.au/__data/assets/pdf_file/0006/23298/DA-2159.pdf

[13] NSW State Record Office, General retention and disposal authority 45: original or source records that have been copied, viewed 04 February 2016, http://www.records.nsw.gov.au/recordkeeping/rules/retention-and-disposal-authorities/general-retention-and-disposal-authorities/files/ga45-original-or-source-records-that-have-been-copied/view

[14] Queensland State Archives, General Retention and Disposal Schedule for Original Paper Records that have been Digitised, viewed 04 February 2016, http://www.archives.qld.gov.au/Recordkeeping/GRKDownloads/Documents/General_Retention_and_Disposal_Schedule_for_Original_Paper_Records_that_have_been_Digitised_QDAN656.pdf

[15] State Records of South Australia, General Disposal Schedule No 21 For management and disposal of source documents and digitised versions after digitisation, viewed 04 February 2016, http://government.archives.sa.gov.au/sites/default/files/20140327%20General%20Disposal%20Schedule%20No%2021%20Final%20V3_Copy.pdf

[16] Northern Territoy Government, Advice 5: Scanning of Records, viewed 04 February 2016 http://www.nt.gov.au/dcis/info_tech/records_policy_standards/records_management_advice/scanning_records.shtml

[17] Commonwealth of Australia, How government works, viewed 04 February 2016. http://www.australia.gov.au/about-government/how-government-works

[18] A. Davies, R. Bergami and S.J. Miah. Implications of Managing Health related Records and Relevant Information Systems within Intergovernmental Agencies. 26th Australasian Conference on Information Systems (ACIS 2015) pp.1-10

[19] Victorian Auditor General's Office, 2007-08:15, Records Management in the Victorian Public Sector

[20] Auditor General Western Australia, Records Management in the Public Sector, pp.1-22

[21] A. Davies.Records management and compliance in the Victorian State Government, Honours thesis, Victoria University, 2014

[22] J. Lappin. 'What will be the next records management orthodoxy?', *Records Management Journal*, vol. 20, no. 3, pp. 252-64, 2010

[23] T. Cook.. Electronic records, paper minds: the revolution in information management and archives in the post/custodial and post/modernist era.[Based on a presentation delivered by the author during his November 1993 Australian tour.]. Archives and Manuscripts, 22(2), pp.300-328., 2004

[24] ACT Auditor-General's Office, Records management in ACT government agencies, pp.1-55

[25] Australian National Audit Office, Recordkeeping, pp.1-76, 2002

[26] National Archives of Australia, How $150,000 was saved in one year by transitioning to digital records management in one business area: Department of Education, Employment and Workplace Relations, viewed 04 February 2016, http://www.naa.gov.au/records-management/digital-transition-and-digital-continuity/digital-transition-policy/digital-transition/case-study-three.aspx

[27] National Archives of Australia, An improved digital information and records system results in savings, major business efficiencies and good user acceptance: Australian National Audit Office, viewed 04 February 2016, http://www.naa.gov.au/records-management/digital-transition-and-digital-continuity/digital-transition-policy/digital-transition/case-study-two.aspx

[28] S. Kumar and K. Bauer. The business case for implementing electronic health records in primary care settings in the United States. *Journal Of Revenue & Pricing Management*, *10(*2), pp.119-131, 2011

[29] T.L. Kiviat. Beyond bitcoin: Issues in regulating blockchain transactions. *Duke Law Journal*, *65(*3), pp.569-608, 2015

[30] T.A. O'Sullivan, N. A., Billing, and D. Stokes. Just what the doctor ordered: Moving forward with electronic health records. Nutrition & Dietetics, vol.68, pp.179-184, 2011

[31] P. Joseph, S. Debowski and P. Goldschmidt. Paradigm shifts in recordkeeping responsibilities: implications for ISO 15489's implementation. *Records Management Journal*, *22(1)*, pp.57-75, 2012

[32] D. Ryan. What is the essence of records management?. Records Management Journal, 16(3), pp.124-130, 2006.

Effectiveness analysis of using Solid State Disk technology

A. Skendžić*, B. Kovačić ** and E. Tijan***
*Polytechnic "Nikola Tesla", Gospic, Croatia
** University of Rijeka, Department of Informatics, Rijeka, Croatia
*** University of Rijeka, Faculty of Maritime Studies, Rijeka, Croatia
*askendzic@velegs-nikolatesla.hr, **bkovacic@uniri.hr, ***etijan@pfri.hr

Abstract - In the daily use of computer applications, large quantities of data are being written and read. Central Processing Unit, Random Access Memory and computer Hard Disk Drive are involved in that process. In 2008 / 2009 the massive commercial use of hard drives based on Solid State technology began, which are now rapidly replacing mechanical hard drives, made up of mechanical and electronic components (actuator, actuator arm, the motor that allows the rotation speed from 5400 rpm to 15000 rpm, controller, cache, etc.). Solid State Disk technology uses the same electronic interface as hard disk drives to communicate with the rest of the computer, which facilitates Solid State Disk installation and replacement of mechanical drives. Hard drives constructed in Solid State technology have the memory controller and flash memory on which the device performance primarily depends. Speed of writing and reading of data from the hard drive is extremely affecting the operation of the entire computer system, especially database search, the playback / recording of audio-video and daily data read / copy. The research described in this paper will compare the speed of writing / reading of the mechanical hard drive with the speed of writing / reading of the hard drive in SSD technology. The comparison will be conducted through synthetic tests with ATTO Disk Benchmark tool. Solid State Disk technology can also be viewed according to the „Green computing" concept, meaning reduced expenses (less energy consumption, less cooling costs for data centers, servers, PC-s) and higher performance). Comparison results between these two technologies will provide a cost-effective argumentation in favor of switching to SSD technology.

I. GREEN COMPUTING

Green computing is a term relating to environmental sustainability in the context of computing or ICT. This includes design, production and use of computers, servers and other sub-systems (monitors, printers, storage devices, and networking/ communications systems) in an effective way, with minimal or no adverse impact on the environment [1].

The goals of green computing are similar to green chemistry: reduction of hazardous materials, maximizing energy efficiency throughout the entire product life cycle and the promotion of recycling or biodegradability of used products. Green computing is very important to all types of devices and systems, ranging from hand-held devices to large data centers [2].

Features of green computing are [3]:

- carbon-free computers,
- computers that consume less power,
- the use of solar energy,
- silent computers,
- lead-free computers.

The carbon which is used in computer manufacturing has become a serious threat to the environment. This means that computer industry needs to take certain steps (in this case, one of the steps should be the production of computer elements with no carbon content) in order to contribute to the prevention of devastation vegetation areas.

Computers can consume less power by installing low-voltage processors which can save energy. Renewable energy sources (such as solar energy) can be used to power computer circuits. Computers are great sources of artificial noise in work environments. Noise pollution is becoming one of the major problems of environmental degradation. Personal computers can produce noise levels of 50 decibels, which can create a stressful atmosphere for the users and the environment (the people nearby). The noise level, as recommended by the World Health Organization should not exceed 35 decibels. The use of lead in modern computer manufacturing is an exception. Microprocessors are manufactured without the use of lead.

The technological approach of making Solid State Disks minimizes the use of harmful compounds during the production (no moving parts and smaller number of electronic components), and takes into account the minimal power consumption with increased disk performance.

II. SOLID STATE DISK TECHNOLOGY AND DRIVES

The development of solid state disk (SSD) technology can be traced back to 1970's. Mass popularization of SSD technology began in 1995 when the first SSD drives entered the commercial use. SSD drive technology is based on NAND flash non-volatile memory [4]. There are two different types of flash memory architecture, as follows:

- NAND Flash - transistors are connected in series, about a million erase cycles,
- NOR-Flash - transistors are connected in parallel, 10000 - 100000 erase cycles.

Unlike the NAND technology whose main purpose is storage, the NOR technology is usually used for operating system storage in special purpose devices (mobile devices such as smartphones). Figure 1 provides a graphical representation of NAND and NOR memory architecture [5].

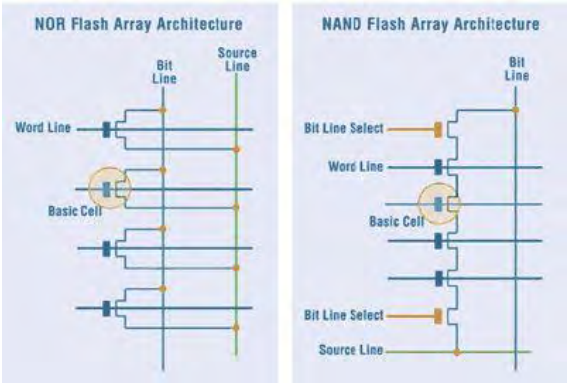


Figure 1. Graphical representation of NAND and NOR memory architecture

Two major advantages of flash memory are storing information without a constant power source (Non-volatile flash memory) and a lower price. Faster reading/writing speed is achieved because of parallelized NAND chips [6]. To increase the performance further, segmenting logically sequential data is used (in RAID 0 disk arrays) and multiplexing is used. NAND comes with an integrated controller, which manages bad blocks and wear levelling algorithms [7]. NAND memory can be divided as shown in Figure 2 [8]: Single Level Cell (SLC, storing 1 bit of data), Multi Level Cell (MLC, storing 2 bits of data) and Triple Level Cell (TLC, storing 3 bits of data.

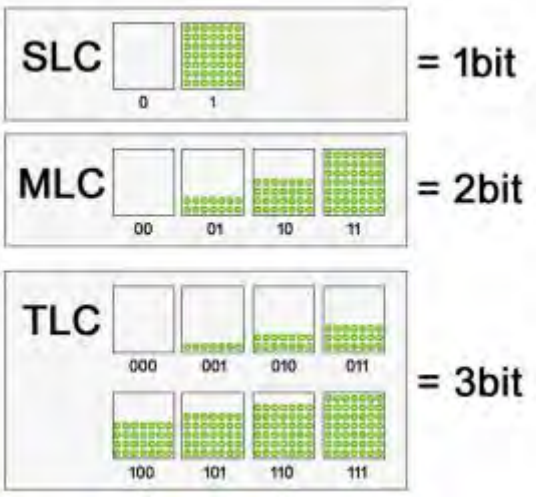


Figure 2. NAND technology: SLC, MLC, TLC

Magnetic hard disk drives (HDD) can simply "overwrite" the old data with the new data. In flash memory (SSD) the old data needs to be deleted before new data could be written. This process prolongs the "write" operation. Attempted solutions to this problem led to the introduction of the TRIM command. With the TRIM command, the operating system informs the controller which memory blocks are no longer in use, so they can be erased by the controller when the device is in the idle mode. This command is implemented in all operating systems from year 2008. Furthermore, flash memory has a limited operating life. Write / delete cycles cause physical degradation of the oxide layer (insulator) of the SSD elements. To slow down this process, the Wear levelling method [8] is used. Algorithms built into the SSD controller uniformly distribute cycles of writing / erasing of the entire solid-state drive, prolonging its lifespan. Such algorithms can be implemented in two variants or methods: dynamic wear leveling method and static wear levelling method [9].

In dynamic wear levelling method, each time a block of data is re-written to the flash memory, it is written to a new location. Static wear levelling selects the target block with the lowest number of total deletions, clears the contents of the block if needed and writes the new data in the block. In other words, static blocks that do not change are periodically moved so that low usage cells are used by other data. This effect enables an SSD to continue operating until the majority of the blocks are near the end of their lifespan.

The essential characteristics of magnetic hard disks are: seek time, rotational latency, and spin-up [10]. Seek time is the time necessary for the head mechanism to position itself above the track in which the requested data is located. Rotational latency occurs if the required disk sector is not positioned exactly under the head, so it takes time before (with the turning of the disk) head positions itself directly above the data. Cumulatively, rotational latency and seek time last from 5 to 20 milliseconds, and are dependent on the speed of the disc. Spin-up is the time period from a stopped disc state to an operational speed (15,000 or 7,200 rpm), and it can last several seconds (only during system boot). Such delays do not exist on the SSDs, so the access to information is much faster, almost instantaneous from the moment the SSD device is switched on.

The most important part of the SSD is its controller. The interface type and, more importantly, the performance of the unit are dependent on the SSD controller. Some of the controller functions are error correction, bad block mapping, read/write caching and encryption.

III. DATA WRITE / READ SPEED MEASUREMENT

Through the ATTO Disk Benchmark testing tools [9], the speeds of writing/reading will be compared. The selected ATTO performance measuring tool is used by Hitachi Corporation [10] in the production and testing of hard drives, and it is considered to be a relevant tools for performance testing. Specifications of the test computer are shown in Table 1.

TABLE I. TEST COMPUTER CONFIGURATION

Main board	GIGABYTE GA-M61SME-S2, NFORCE 405 chipset
Processor (CPU)	AMD 7750 Dual core 2,7 GHz
RAM	4 GB DDR II (800CMHz) Dual channel
Graphic card	ATI RADEON 4600
HDD (SSD)	KINGSTON 64 GB SATA II (system disk)
HDD (magnetic)	WESTERN DIGITAL BLACK 500 GB 7200 rpm SATA II
Operating system	MS Windows 7 + ATTO Disk benchmark

When performing the test, the TRIM option on the SSD drive was enabled.

Factory data for the test model SSD were as follows:

• Maximum speed of sequential data read: 260 MB / s,

• Maximum speed of sequential data write: up to 100 MB / s.

The read / write speed test of the SSD (Kingston V300) and the mechanical hard drive (WD DIGITAL BLACK 500) was performed using ATTO disk benchmark tool. It should be noted that mechanical hard drives are in the active mode during reading and writing, and then consume the most electricity, and produce the most noise and heat. Data from the tests (values of random data read and write in selected blocks) is shown in Table 2.

TABLE II. THE MEASURED VALUES OF THE READ/WRITE DATA OPERATION IN THE DEFAULT BLOCK SIZE

The size of the test blocks	**4 KB**		**64 KB**		**512 KB**		**8192 KB**	
Read operation (**MB/s**)	**SSD**	**HDD**	**SSD**	**HDD**	**SSD**	**HDD**	**SSD**	**HDD**
	56636	56039	127543	107648	132960	111615	133328	109565
Write operation (**MB/s**)	**SSD**	**HDD**	**SSD**	**HDD**	**SSD**	**HDD**	**SSD**	**HDD**
	47300	66941	94894	105213	97084	104044	98 689	105683

The smallest measured values shown in Table 2 relate to read / write of data in blocks of 4 KB, while the largest values are at the test block size of 512 KB and 8192 KB. The differences in the read / write speed of blocks larger than 512 KB are negligible. SSD drives have lower performance when working with large amounts of small size files, and with compressed data. The decline in SSD performance is also notable when the disk is nearing the full capacity (the performance starts to degrade at around half capacity).

Compared with factory declared read / write speeds of the test Kingston V200 64GB SSD, the results of sequential data write speeds are almost identical. Sequential data read performance was only half of the factory declared. The reason for this can be the SATA 2.0 interface (theoretical limit of a maximum of 300 MB / s). The results of the sequential data read were at the level of SATA 1 (150 MB / s).

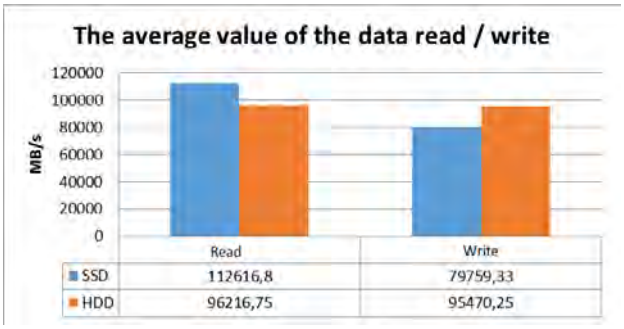


Figure 3. The average value of the data read/write

After the read / write test, the power consumption was compared. The results are shown in Table 3 [11], [12].

TABLE III. POWER CONSUMPTION IN DIFFERENT MODES OF OPERATION

Disk	Power Consumption	Idle	Read	Write
Kingston SSD V300 60 GB SATA III	0.640 W	1.423 W	2.052 W	2.052 W
Western digital WD 500 CAVIAR BLACK SATA III		6.1 W	6.8 W	6.8W

It is obvious that SSD technology power consumption is around 3 times lower in relation to the rotating disks.

IV. SSD DRIVES AND GREEN COMPUTING

When measuring the performance of mechanical disk drives and SSD drives, significant differences can be observed. The advantage of SSD drives is reflected in the speed of read / write and the average access time (less than one millisecond, compared to 4 milliseconds or more for magnetic disk drives). Although the listed facts are in favor of SSD, it is important to consider the following:

a) It is important to use a faster version of the SATA 3.0 (6 Gb / s, 600 MB/s) in relation to the SATA 2.0 (3 Gb / s, 300 MB / s) interface. Connecting the SSDs to a slower SATA 2.0 interface will cause the decline of SSD performance, such as in the test case. The specified sequential data read / write of the test drive is 260 MB / s (read) and 100 MB / s (write). In the test itself, due to limited SATA 2.0 interface, Kingston V300 SSD achieved lower results than factory declared (maximum of 133 MB / s).

b) An extremely large number of disk controllers exist on the market today. Faster and more advanced disk controllers offer higher performance of SSDs.

c) The capacity of SSD. Performance is reduced when the disk is nearing full capacity.

TABLE IV. ADVANTAGES AND DISADVANTAGES OF SSD / HDD TECHNOLOGY [13]

Attribute	SSD (Solid State Disk)	HDD (Hard Disk Drive)
Power consumption	2-3 w	6-7 w
Performance (write)	250 – 550 MB /s	50 – 150 MB / s
Number of components	A small number of components	A larger number of components and moving parts (mechanical)
Noise	No moving parts	Moving Parts / vibration / sound
Capacity	About 1 GB	About 6 GB
MTBF	2 million working hours	1,5 million working hours
Price per gigabyte	0,10 $ (price per 1 TB model)	0,06 (Price per 4 TB model)

The main drawback of SSDs is a limited number of writing / deletion cycles of NAND memory:

•3000 for commercial Multi-level cell,

•1000 for commercial Triple-level cell,

•100000 for Single level cell (industrial use, expensive technology).

The difference of SSDs compared to conventional hard drives is also visible in a way of price calculation. Instead of expressing prices per GB or TB, the price per IOPS (Input / Output operations Per Second - the number of operations per second that the unit can perform) is used. For comparison, Table 5 shows the average values of all the important characteristics of hard drives and flash / DRAM SSD drives.

TABLE V. SUMMARY OF AVERAGE VALUES OF ALL THE IMPORTANT CHARACTERISTICS OF HARD DRIVES AND FLASH (DRAM) SSD DRIVES [13]

TECHNOLOGY	CAPACITY (GB)	LATENCY
"large" HDD	2 500	120 000
"fast" HDD	700	70 000
Flash SSD	250	2 000
Flash SSD (read only)	250	450
DRAM SSD	250	20

SSD technology provides significant improvements compared to mechanical drives which justifies their use in green computing. However the current price of SSD technology significantly affects the application of SSD in green computing.

V. CONCLUSION

In this paper, in addition to the concept of NAND and NOR memory architecture that makes an essential element of SSD, the comparison of read / write drive speed (SSD technology and magnetic hard disk) is presented. SSD technology is faster, but there is a large difference in the price per megabyte. Although the price of SSDs is expressed in the number of IOPS operations, the price per megabyte is still often used and compared. Thus, SSDs are in the highest price range when it comes to permanent storage of data. The operating speed of SSD, low power consumption, a small number of physical components and absence of any noise are all in favour of the application of SSD in green computing. Furthermore, large data centres in the future could, by using the SSD technology, significantly save on electricity consumption for the operation of the data servers, increase performance and use less energy in cooling systems. SSD technology is also more durable. The drawbacks of SSD technology are currently the higher price in comparison to magnetic hard drives ($ 0.10 to 1 TB compared to $ 0.06 for the 4 TB model), which significantly affects the use of SSD. With the development of technology and lowering the costs, SSD can expect significantly larger representation in green computing.

REFERENCES

[1] http://www.kosacrta.hr/info-tab/najvaznije-stvari/green-computing/ (31.1.2016.)

[2] http://www.academia.edu/6982393/Power_Management_Techniques_for_Data_Centers_A_Survey (06.01.2016.)

[3] San Murugesan, "Harnessing Green IT: Principles and Practices," IEEE IT Professional, January–February 2008, p 24-33

[4] http://sistemac.carnet.hr/node/1211 (14.9.2015.)

[5] http://www.racunalo.com/memorije-u-mobilnim-telefonima-rom-interna-flash-memorija- stock-rom-castom-rom-sve-sto-trebate-znati/ (14.9.2015.)

[6] http://web.archive.org/web/20071013150729/http:/www.corsair.com/_faq/FAQ_flash_drive_wear_leveling.pdf. June 2007 (14.9.2015)

[7] Perdue, Ken. "Wear Leveling Application Note" . Url: http://www.eettaiwan.com/STATIC/PDF/200808/EETOL_2008IIC_Spansion_AN_13.pdf (14.9.2015.)

[8] http://infoteh.etf.unssa.rs.ba/zbornik/2015/radovi/RSS-5/RSS-5-12.pdf (14.10.2015.)

[9] https://www.attotech.com/disk-benchmark/ (05.01.2016.)

[10] http://www.hitachi.com (05.01.2016.)

[11] http://www.wdc.com (31.1.2016.)

[12] http://www.kingston.com (31.1.2016.)

[13] http://www.storagereview.com/ssd_vs_hdd (31.1.2016.)

[14] http://sistemac.carnet.hr/node/1219 (14.9.2015.)

Information and communication technologies and the new forms of organized crime in network society

Marija Boban, PhD*
*Assistant Professor at University of Split Faculty of Law/ Split, Croatia
E-mail: marija.boban@pravst.hr

Abstract – The network society and the complexity of the changes that have affected digital economy is legally accompanied by the remains and traces of the past, including the addition of positive encountered numerous negative companions. This common, and now generally accepted phrase, encompasses the totality of criminal acts unauthorized affects the use, integrity and availability of technical, software or database computer system or the confidentiality of digital data. Increasingly, criminal groups are today using the Internet in the commission of crimes and offenses that are considered "traditional". The most dangerous, among them, are new forms of organized crime. The starting point of the paper presents the concept of the role of Internet as infrastructure for new forms of organized crimes with an emphasis on regulative and legal framework for prevention of new internet crimes as well as global sharing of information threw databases in order to fight against criminality of new era. The author is also giving an overview of legal practice of sanctioning of new forms of organized crime in criminal law of Republic of Croatia threw legal regulation and conventions, what actually presents the main request for the efficiency of government repression as the body and at the same time and their placement in the legal framework which is acceptable from the standpoint of the rights of citizens, particularly as well as from the standpoint of current criminal procedure law in today's information society.
Key words: information and communication technologies, organized crime, legal framework, databases, criminal law

I. INTRODUCTION

The virtual human world is becoming increasingly the natural an environment in which one lives and works on a daily basis. In this world, the technologiy meets the needs who arise historically by developing its human forces versatile communication and equal use of heritage culture and civilization. In only the beginning of a new epoch-making world turns, shows the reality of everyday life in which half of humanity has not yet reached that stage of development, and the other half live in the different developed communities where the old and new "meet" in the variety of shapes and combinations.

Protection against computer crime becomes an important function of society, because with each new technological innovation occurs abuse committed on computers or with their help. They meet at all local, national and regional levels, and interweaving of computer networks and the expansion of the Internet and the world wide open spaces. It contributes greatly to the digital format available data and information, as well as the possibility that the atrocities committed by remote access, using, managing and manipulating data. Favors them, also, the fact that the same phenomena and processes is also intertwined permitted and illegal, and the effects are achieved by the same means and equal methods and procedures that are used in everyday work and the use of information technology.

The new era and network society is well described by Castells who determines comprehends three sociological dimensions — production, power, and experience — stressing that the organization of the economy, of the state and its institutions, and the ways that people create meaning in their lives through collective action, are irreducible sources of social dynamics — that must be understood as both discrete and inter-related entities. Moreover, as an established cybernetic culture theoreticians with their Internet development analysis, stress the roles of the state (military and academic), social movements (computer hackers and social activists), and business in shaping the economic infrastructure according to their (conflicting) interests. [1] Discussions are conducted also about the educational and cultural potentials of the Internet, but mostly without a precise definition of the content of the term. Undisputed, information and media literacy is an essential part of literacy and culture of modern times and precondition of internet (user) security.

The contemporary approach to business, mostly eBusiness as the form of doing business in digital economy, and communication through global information networks has led to the increasing integration not only the people but also of information systems. With all its benefits that such an operating mode carries, there are a number of risks to the safety of all participants, as well as information systems themselves. Data and information, which are within such systems processed, stored and distributed, as well as the technical basis that it provides, are increasingly the target of such abuse, hence the set as the main task of demand for their efficient and complete protection. [2]

II. Internet as infrastructure for committing criminal acts of organized crime

Procreation and exponential growth and development of the global information and communication network - the Internet [3] – was accompanied by a rapid increase in the amount of information and new, interactive way of communicating which knows no territorial boundaries, radically changing the traditional way of work and life. [4] Also, the emergence of eBusiness as well as electronic banking and virtual money, electronic mail and telephone, and ever faster introduction and adoption of e-business, the issue of effective and comprehensive protection of personal data (privacy) and information systems, is gaining in importance and becomes the primary interests of the individual and the community in general as well as networks developed within the defense industry, and the need for cooperation between the academic community today is available to ordinary citizens.

The openness of the architecture of the Internet and its continued development in which users were also the creators and contributed to its further development, were it's main strength of development. In addition, the cost of connecting to the Internet weren't high, and the software was available, and often free. [5] The situation in which users are also the main protagonists of the development of technology shortens the process between the stages of development, testing, use and further modifications to the system, which is one of the reasons for the rapid development of the Internet, not only in terms of expansion of network nodes, but also in terms of different services that continuously evolving (as shown on Figure 1.)

Increase in the number of Internet users is connected with change and growth in the number of available

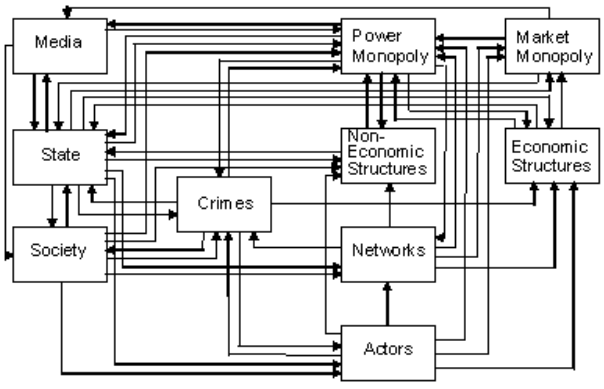


Figure 1. The internet as (infra)structure of organized crime [6]

information. The Internet has become broadly available, less structured and somewhat chaotic. A huge amount of information available and the large number of available contacts provide new opportunities and at the same time overwhelm users. The standard protocols and formats allow the exchange of data available, but there is still plenty of room for experimentation and creativity to create new artistic, cultural and organizational forms. Discussions, about the Internet as the "information superhighway" and "cyberspace", generally focuses more on technology development and commercial benefits, which is resulting with new opportunities.

The main preconditions for this development of the Internet was an open, decentralized, interactive network architecture, network protocols, which then must also be open and that can be easily modified, and institutions / governance structure and development of the Internet, which must be in accordance with the principles of openness and cooperation that it does not would be hampered. Unfortunately, today, precisely such kind of development also brought some negative consequences - especially in the field of organized crimes.

III. Criminal justice model in the fight against organized crime

A. Definition of organized crime

Definitions of what constitutes the organized crime vary widely from country to country. Most often it is defined as any group having some manner of a formalized structure and whose primary objective is to obtain money through illegal activities. Such groups maintain their position through the use of actual or threatened violence, corrupt public officials, graft, or extortion, and generally have a significant impact on the people in their locales, region, or the country as a whole. Organized networks are typically involved in many different types of criminal activity spanning several countries. These activities may include trafficking in humans, illicit goods, weapons and drugs, armed robbery, counterfeiting and money laundering. [6]

B. Conceptual elements of the definition of organized crime

Elements of content conceptual planning are related to conditions required for the participation of more than two persons who together constitute serious offenses over a long period of time or indefinitely, in advance are divided criminal tasks and roles, the existence of secret internal structures and internal control and discipline of members, methods. Classical methods of organized crime are violence, intimidation, corruption, the so-called. money laundering, connection with criminal associations across national boundaries (trans-national), and the influence on politics, media, executive and judiciary) and goals are always acquisition of profit or power.

Conceptual elements of the definition of organized crime can be classified into two groups:

a/ Characteristics of associations of organized crime

a) the acts were committed by several persons who are organized as a group;

b) members of the group are involved in the hierarchical structure in which there are leaders and subordinate members, the association is often permanent - members can be changed, and that the organization does not fall apart;

c) associations work on the principle of confidentiality;

d) preserve the solidarity of members of the organization based on internal control, which often contain severe measures including discipline and violence, which can be applied for external protection association and its activities; e) a part of the association organized so that is involved in legitimate economic activity and is used for example. money laundering. [7]

b/ Characteristics of criminal activity

a) Organized Crime primarily maintained market of illegal trade;

b) certain criminal activities contain violence, intimidation to achieve the objective of the association. Also, the purpose of the use of violence can be and to prevent disclosure and gaining a monopoly on certain illegal market;

c) some criminal activities constitute money laundering in terms of involvement of illegal income (crime-generated profits) in the legal market;

d) criminal activities are performed by a professional, sophisticated methods similar to conceal the crime;

e) Organized crime is often transnational and international nature [8]

C. Legal definition of the concept of organized crime

Criminal Law (CL) in the concept incriminates association in a criminal association (CL Art.328.) and second in committing the criminal offence of taking part in a criminal association (CL Art.329.). [9]

Legal definition of the concept of organized crime is that criminal Association is made of at least 3 people who have joined together with the common purpose of committing one or more crimes for which is adjudicated prison sentence of 3 years or more and that does not include an association that make people accidentally connected directly to committing one criminal offence (CL Art. 328, paragraph 4). [9] Criminal Association is the foundation of the concept organized crime.

D. The concept of organized crime of legal entities in eBusiness

Legal persons are not only fiction in the field of organized crime, they do exist, and occupy an important role in the social organization and are capable of their activities cause damage as well as individual. Severe forms of economic crime in digital economy today are mainly a consequence of not an individual but a series of people in companies with the ambition of making profit at all costs – especially on digital markets threw eBusiness. For example, in Germany about 80% of severe economic crime committed within companies. [8]

Market competition is forcing operators to gain profit by any means necessary threw various crimes (fraud, money laundering, tax evasion, abuse of the bankruptcy, works to the detriment of consumers and the like.) These criminal offences can be attributed to the legal person in whose name and for whose account committed.

We distinguish between legal entities there are formed for committing criminal offenses or the situation when they, after the establishment, mainly engaged in criminal activity by acquiring illegal profit that becomes their main driving force. They can also just be a facade for the real criminal organization. There are numerous and diverse examples of the involvement of legal entities in the field of organized criminal activities in eBusiness such as industrial and financial companies in the US which are practicing bribing high-ranking officials of other countries for the purpose of accepting profitable contracts and various financial institutions may also be participants in the crime activities: banks, savings banks, exchange offices, brokers, import, export and other commercial companies which are doing so called enterprise - transaction model of crime in regions of high technological development, which is manifested through the activities of legal institutions (using ICT technologies and mostly eBusiness model).

What makes it difficult is the fact that conceptual definition of crime of legal entities is peculiar socio-psychological situation in which criminal activity takes place as a rational and professional activity of a legal person. Conducted criminal act, in the interest of legal person, is not only an expression of the will of individuals, but also the result of the spiritual atmosphere that rules in the legal a person who is in the German literature as "group spirit" or "criminal attitude within the legal person" and the American "culture of the corporation".[8] In that case, the legal entity creates the belief that good is everything that serves their interests, which among members of the legal person leads to the conclusion that the violation of legal norms is lesser evil if it serves our common interests (especially if the legal person is willing to compensate the convicted individual, for example: they will pay a fine for him). Such a group or team spirit weakens, or neutralizes, the resistance of members of a legal person under the violation of legal norms, and increases their willingness to accept risks that outside legal person would not accept

The proposals for common European solutions include criminality for participation in a criminal

organization, in addition to citizens, legal persons and their activities become the means or cover criminal organizations.

There is the possibility of the existence of two types of criminal organizations:

- The first organization with the sole or dominant criminal goals (eg. those for the sale of drugs) and
- Second legal business organizations in which at some levels perform criminal activity where appear difficulties in establishing individual criminal responsibility for acts committed in its favor. ".[7]

The important role in organized crime in eBusiness also plays the coherence and impact of crime on the of legal entities expansion of organized crime, especially transnational. These is protected under UN Convention against Transnational Organized Crime: "Each State Party shall adopt the necessary measures, in accordance with its legal principles, to establish the responsibility of legal entities for participation in serious crimes involving organized criminal group ".[10] The international community envisages preventive measures that would prevent abuse of legal entities by organized crime groups by establishing a public record of legal and natural persons engaged in the establishment, management and funding of legal entities; introducing the possibility of dissolution based on a court order or otherwise, persons convicted of offenses covered by the Convention, to a certain period could not work at the site managers of legal entities; establishment of national records of persons who are prohibited from management of legal persons and others.

The Council of Europe has also adopted the 2000 Criminal Law Convention on Corruption which prescribes liability of legal persons for the offenses of active bribery, trading in influence and money laundering, "which is to their benefit by any natural person, acting alone or as part of a legal entity, which has the leading role within the legal person, and on the basis of: the right to represent the legal person or authority to take decisions on behalf of the legal person or authority to exercise control within the legal person and its inclusion as an accomplice or instigator in the above-mentioned acts. "→ legal the persons responsible for those crimes in the case of "when, due to lack of supervision which it had to carry out a natural person, has been a criminal offense for the benefit of that legal person, committed by an individual under the jurisdiction of that legal entity." ".[11]

The next step in fighting the new forms of organized crime was the Convention on Cybercrime VE (2001) who stipulates the liability of legal persons and provides a definition of the crime of the legal person as an criminal offense which is in favor of legal persons committed by a a person acting either individually or as part of the body of the legal person who has a leading position within legal person, based on: a) the authorization to represent the legal person; b) authority to take decisions on behalf of the legal person; c) managerial powers within the legal person. Also, criminality of legal entities includes the case where due to lack of supervision or control, physical person acting under the authority of and on behalf of the legal person is allowed committing the offense established by the Convention. ".[12]

IV. INFORMATION AND COMMUNICATION TEHCNOLOGIES AND NEW FORMS OF ORGANIZED CRIME

Today, when we are talking about Internet security, or rather insecurity, we often forget the motives and reasons that led to its creation, and the purpose of which is supposed to serve. Since the original purpose of this network was not a commercial nor focused on what the network today represents, and what it's actually used for today, nor expected such a large extent and the number of users, security is not given sufficient amount of attention. So with the protocols and rules, according to which computers now communicate within the network, which is primarily produced to ensure the efficiency, flexibility and openness of the system, it has produced a huge amount of difficulties in the security of communication and information exchange.

A. Global sharing of information as the way of fighting internatinal crime

The success of international police investigations is dependent upon the availability of up-to-date, global data. Police investigations function threw INTERPOL. INTERPOL is the world's largest international police organization, with 190 member countries. Their role is to enable police around the world to work together to make the world a safer place threw high-tech infrastructure of technical and operational support helps meeting the growing challenges of fighting crime in the 21st century. [14] Threw that high-tech infrastructure they facilitate global information sharing by managing a range of criminal information databases which enable the global law enforcement community to connect seemingly unrelated pieces of data, thereby facilitating investigations and enhancing international police cooperation. It provides its member countries with instant, direct access to a wide range of criminal databases, containing millions of records on fingerprints, DNA, stolen motor vehicles, firearms, stolen and lost travel documents and more. The databases share information contributed by member countries on a daily basis and with the following features:

Information is shared through the I-24/7 secure police communications system;

- In compliance with international standards;
- Legally founded;
- Technologically advanced;
- Embedded with security features;
- Flexible and can be customized.

All databases, except the one of child sexual exploitation images, are accessible through the I-24/7 Dashboard, a restricted-access Internet portal.

B. The main INTERPOL databases for fighting organized crime using eBusiness model

The police investigation continues to extend access to its databases beyond its National Central Bureaus (NCBs) to frontline law enforcement officers, such as border guards in member countries, allowing them to search the databases on wanted persons, stolen and lost travel documents and stolen motor vehicles. These solutions allow an officer to submit a query simultaneously to a national database and an INTERPOL database and receive responses from both within seconds.

The main INTERPOL databases are:

Nominal Data – contains more than 158,000 records on known international criminals, missing persons and dead bodies, with their criminal histories, photographs, fingerprints, etc.

DNA Profiles – contains around 150,000 DNA profiles from 73 countries. DNA profiles are numerically coded sets of genetic markers unique to every individual and can be used to help solve crimes and identify missing persons and unidentified bodies.

Fingerprints – INTERPOL manages an Automated Fingerprint Identification System which contains more than 216,000 sets of fingerprints and more than 8,400 crime scene marks submitted by member countries either electronically or by mail.

Child sexual exploitation images – 5,785 victims and 2,965 offenders have been identified by investigators using the INTERPOL International Child Sexual Exploitation

Stolen and Lost Travel Documents – holds information on more than 45 million travel documents reported lost or stolen by 169 countries. This database enables NCBs and other authorized entities, such as immigration and border control officers, to ascertain the validity of a suspect travel document in seconds.

Stolen Administrative Documents – contains information on almost 700,000 official documents which serve to identify objects; for example, vehicle registration documents and clearance certificates for import/export.

Stolen Motor Vehicles – provides extensive identification details on approximately 6.8 million vehicles reported stolen around the world. In 2014, more than 132,000 stolen vehicles were identified using the database.

*Stolen Works of Ar*t – allows member countries to research records on almost 45,000 pieces of artwork and cultural heritage reported stolen by 129 participating countries.

Fusion Task Force – a database of more than 10,000 persons suspected of being linked to terrorist activities. Some 160 member countries currently contribute information.

Firearms – The INTERPOL Firearms Reference Table allows investigators to properly identify a firearm used in a crime. It contains more than 250,000 firearms references and 57,000 high-quality images. The INTERPOL Ballistic Information Network is a platform for the international sharing and comparison of ballistics data, holding more than 270,000 records from 17 participating countries. [15]

V. THE SANCTIONING OF NEW FORMS OF ORGANIZED CRIME

Firstly, we must state that sanctioning of new forms of organized crime are the request for the efficiency of government repression of the body and at the same time and their placement in the legal framework is acceptable from the standpoint of the rights of citizens, particularly from the standpoint of current criminal procedure law

The sanctions are very important because of its repressive and preventive function. The response to organized crime are those sanctions more appropriate to the nature of this group of criminal activity and what is important that they are focused just on profit as the driving force of organized crime and acquisition with more expressed motive of acquiring political power and influence on the overall economic, social and political relations in an area. Because of that, the most important sanctions are confiscation and forfeiture get.

Confiscation is generally accepted as an additional sanction of forfeiture by the perpetration offense or used for its commission. The most of the legal systems of confiscation of all assets of the perpetrator canceled due to abuse in totalitarian regimes while this sanction somewhere reserved for certain crimes eg. It proceeds to a person who has conducted the crime in any form – including confiscation of income (cash, bank accounts, etc.).

In Croatia, it is applied to individuals even when the acts were committed within a criminal organization and sanction the confiscation of the offender's personal property Croatian Criminal Law does not provide. [9]

Organized crime also presents a threat to the security and stability of every country including the Republic of Croatia Organized crime favors state-political integration processes, liberalized border regime or the movement and

residence of foreigners, changed and liberalized the possibilities of transnational economic and financial operations because the consequences of globalization open organized crime wide field of action. [13]

VI. CONLUSION

The development of information and communication technology in network society has brought new dangers in the field of criminal law and organized crimes, which arises awareness of the need for new actions in legal as well as in regulation and security measures as shown in the paper.

The security and new treaties in the legislation and detailed elaboration of all the possibilities of abuse of and new forms of criminal offenses of organized crimes s seems the only way to prevent the effective sanctioning of criminal acts of violation of human rights as well as criminal law in general. Criminal groups now use the Internet as an "infrastructure" for facilitating illegal immigration, trafficking in human organs, selling of rare animal species, cybercrime and cyber terrorism as the new forms of organizes crimes.

Also, developing personal awareness of the need to protect by any software, cryptographic protection or limiting access to personal data by third parties, especially in the eBusiness model with criminal acts against the person eg. identity theft and bank account theft) increases the level of security and protection operations in the focus on eBusiness.

REFERENCES

[1] M. Castells, ”The Rise of the Network Society, The Information Age: Economy, Society and Culture Vol. I. Cambridge, MA; Oxford, UK: Blackwell, 1996.

[2] D. Dragičević, „ Computer Crime and Information Systems", Informator, 1999.

[3] B. Leiner, et. al.,: „A Brief History of the Internet“, - available at http://www.isoc.org/internet/history/brief.shtml (05. 03. 2016.)

[4] V. Grbavac, M. Plenković, „Communication systems at the turn of the 21st century", Informatology, no. 27., 1995.

[5] V. G. CERF, R. E. KAHN, "A protocol for packet network interconnection", IEEE Trans. Comm. Tech., vol. COM-22, V 5, May 1974., pp 627-641

[6] M. Boban, “National Security Systems”, Authorized teaching materials, University of Split, University Department of Forensic Science, Split, 2016.

[7] A. Kurtović, A., “Legal arrangements in the light of the measures warning, security measures and confiscation of the proceeds, Croatian Annual of Criminal Law and Practice, Zagreb, vol. 7, br. 2/2000., pp 376-377.

[8] D. Derenčinović: “Legal nature of forfeiture of property and its importance in the prevention of organized crime”, the paper is presented at the conference "Withdrawal of the proceeds of crime - the role and obligations of state authorities of the Republic Croatian", Police and security, 1999., pp 3-4

[9] A. Kurtović, G. Tomašević: Organized crime: forfeiture of the gains originating from crime”, the paper is presented at the conference "Withdrawal of the proceeds of crime - the role and obligations of state authorities of the Republic Croatian", Police and security, 1999., pp 3-4

[10] Criminal Law, Official Gazette of Republic of Croatia, No 125/11, 144/12, 56/15, 61/15

[11] The United Nations Convention against Transnational Organized Crime, adopted by General Assembly resolution 55/25 of 15, November 2000.

[12] Criminal Law Convention on Corruption, European Treaty Series - No. 173, Strasbourg, 27.I.1999.

[13] Convention on Cybercrime, CETS No.185, Treaty, Budapest, 23/11/2001

[14] SOA, Official pages of Croatia SOA, National Security, https://www.soa.hr/hr/soa/podrucja/kriminalitet/

[15] INTERPOL official webpage - http://www.interpol.int/About-INTERPOL/Overview (05. 03. 2016.)

[16] INTERPOL official publications - Databases (COM/FS/2015-03/GI-04) – available at http://www.interpol.int/News-and-media/Publications2/Pages/Publications (08. 03. 2016.)

Digitalizacija lokalne uprave na primjeru Istarske županije

Lea Ordanić i Nataša Šarić-Kekić
IN2 d.o.o., Zagreb, Hrvatska
Lea.Ordanic@in2.eu
Natasa.SaricKekic@in2.eu

Sažetak – Suvremeno upravljanje poslovanjem županije zahtijeva poslovni sustav koji omogućuje učinkovito planiranje proračuna i praćenje njegove realizacije, adekvatnu evidenciju računovodstvenih događaja, pripremu plana nabave i evidenciju njegove realizacije te ažurnu izradu financijskih i proračunskih izvještaja. Za realizaciju navedenih zahtjeva korišteno je Microsoft Dynamics AX 2012 poslovno informacijsko rješenje koje zbog svoje prilagodljivosti, mogućnosti proširenja odnosno nadograđivanja postojećih funkcionalnosti omogućuje najviše fleksibilnosti u realizaciji inicijalnih zahtjeva korisnika i kasnijih dopuna.

I. UVOD

Poslovanje županije ne uključuje samo računovodstvenu i financijsku evidenciju poslovnih događaja i pripadno izvještavanje već se temelji na adekvatnom proračunskom planiranju i praćenju realizacije proračuna. Ovi podaci su od iznimne važnosti za lokalnu zajednicu koja očekuje transparentno poslovanje lokalne samouprave te brzo i odgovarajuće izvještavanje. Isto tako, podloga proračunskom planiranju je strateško planiranje na razini države koje prikazuje strateške prioritete i ciljeve vladine politike.

Postojeće proračunske, računovodstvene i porezne aplikacije koje je koristila Istarska županija, obzirom da su bile realizirane na različitim platformama, uzrokovale su povećanje troškova međusobne integracije te oduzimale više vremena za planiranje proračuna i praćenje njegove realizacije. Takav način praćenja poslovanja nije mogao osigurati odgovarajuće kontrolne mehanizme realizacije proračuna te promptno plasiranje financijskih i proračunskih izvještaja što su bili osnovni zahtjevi korisnika u novom programskom rješenju.

II. CILJEVI I ODABIR POSLOVNOG RJEŠENJA

A. *Ciljevi projekta*

Korisnici su uvođenjem novog poslovnog sustava očekivali stabilnu, upravljivu i prilagodljivu platformu koja će osigurati integrirano poslovanje županije uz obuhvaćanje svih procesa proračuna (od planiranja do online kontrole realizacije proračuna) te istovremeno omogućio vanjskim (proračunskim) korisnicima pristupanje sustavu i aktivno sudjelovanje u procesima planiranja proračuna i praćenja realizacije.

Uvođenjem novog sustava očekivano je, u prvoj fazi, osigurati automatizaciju poslovnih procesa iz područja proračunskog računovodstva, planiranja novog proračuna, izrade izmjena i dopuna proračuna tijekom poslovne godine te procesa javne nabave, a u narednim fazama svih ostalih procesa Istarske županije što uključuje evidenciju i donošenje novih poreznih rješenja, praćenje redovne i prisilne naplate poreza te donošenje i praćenje naplate koncesija.

Ovim načinom postigla bi se digitalizacija, automatizacija i jedinstvenost svih procesa kako bi svi sudionici i interesne skupine imale mogućnost promptnog uvida u željeni set informacija.

B. *Odabir poslovnog rješenja*

Poslovno informatički sustav Microsoft Dynamics AX 2012 je odabran iz razloga što je riječ o svjetskom rješenju čiji je rad provjeren kroz poslovanje velikog broja tvrtki, a uz to ima podršku najvećeg svjetskog proizvođača softvera – Microsofta. Sustav podržava poslovanje velikog broja tvrtki gospodarskog i javnog sektora čiji složeni procesi mogu biti jednostavno postavljeni ugrađenim funkcionalnostima sustava.

S korisničkog aspekta, ovaj sustav je korisnicima blizak zbog prepoznatljivog Microsoft korisničkog sučelja te zbog integracije sa MS Office alatima.

No, najvažnija prednost ovog poslovno informatičkog sustava je jednostavno proširenje odnosno nadograđivanje postojećih funkcionalnosti sustava ako za to postoji potreba. [1]

C. *Shema novog poslovnog sustava*

Središte novog poslovnog sustava je MS Dynamics AX 2012 koji je povezan sa vanjskim aplikacijama na istoj Microsoft platformi.

AX je prvenstveno namijenjen računovodstvenom i financijskom poslovanju županije, evidenciji dugotrajne imovine, procesima nabave te održavanju svih matičnih šifrarnika. Planiranje proračuna i praćenje realizacije proračunskih korisnika kao i pregledi stanja su izdvojeni u zasebno web rješenje kako bi omogućili jednostavniji pristup velikom broju udaljenih korisnika. Shema novog poslovnog sustava prikazan je slikom 1.

III. PLANIRANJE I PRAĆENJE REALIZACIJE PRORAČUNA

A. *Planiranje proračuna*

U procesu planiranja proračuna Istarske županije, aktivnu ulogu imaju i proračunski korisnici. Oni koriste spomenuto web rješenje za planiranje svojih troškova unutar proračuna županije, a također i za praćenje realizacije tih planova. Osim toga, njihovi podaci su podloga za planiranje i realizaciju proračuna županije.

Jedinstvenost matičnih podataka u MS Dynamics AX 2012 rješenju osigurava jednostavnu integraciju prema web rješenju za unos planiranog proračuna. Matični podaci ključni za planiranje proračuna su proračunske klasifikacije prikazane tablicom 1. [2]

Pojednostavljeni tijek procesa planiranja uključuje tri uloge: financije, pročelnike i proračunske korisnike kako je prikazano slikom 2. Uloga financija je definiranje osnovnog plana rashoda i izdataka koji se potom šalje pročelnicima. Pročelnici svakog od razdjela unose limite na željenoj razini klasifikacija prema proračunskim korisnicima, ali također unose i planirane vrijednosti proračuna za dio proračuna koji se odnosi na njihov razdjel. Proračunski korisnici unose planirane vrijednosti na razini računa (konta) unutar dozvoljenih limita te nakon toga vraćaju plan na odobrenje pročelnicima. Pročelnici odobravaju (ili vraćaju na doradu) proračun svakog pojedinog korisnika, a financije objedinjuju plan dok isti ne postane plan proračuna županije za narednu godinu.

Tijek rada omogućuje vraćanje proračuna na doradu svakoj od nižih uloga, a konačna potvrda proračuna učitava se u AX te služi kao podloga za izvještavanje i kontrolu realizacije proračuna.

Sličan tijek rada kao i za donošenje izvornog plana proračuna koristi se i za donošenje rebalansa, preraspodjelu sredstava te izmjenu plana trošenja. Nakon potvrde od strane financija sve navedene izmjene prenose se u AX u obliku razlika u odnosu na prethodnu verziju proračuna određene godine.

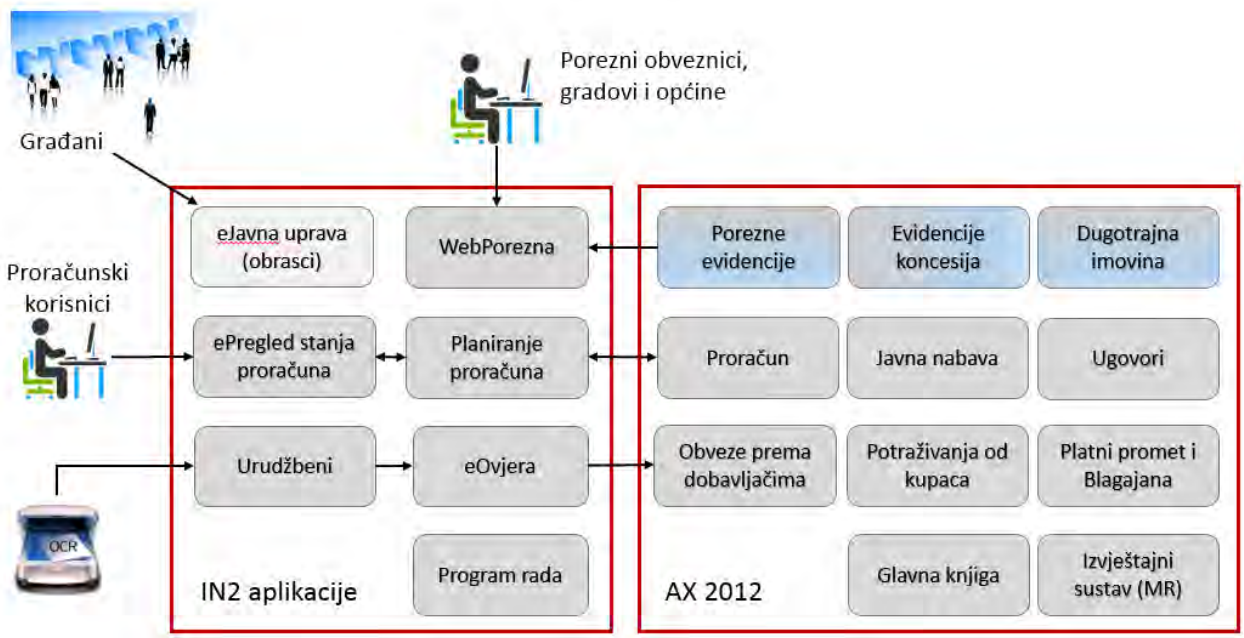
Bitno je napomenuti da unosom proračuna u AX svaka stavka proračuna dobiva jedinstveni broj koji se naziva pozicija. Obzirom da je pozicija jedinstvena za svaku stavku proračuna, poznavanjem pozicije poznate su i sve ostale proračunske klasifikacije.

Web rješenje za unos proračuna osigurava jednostavan dohvat novih vrijednosti klasifikacija ako su iste prethodno otvorene u AX-u, a potrebne su za planiranje određenog rashoda ili izdatka u novoj godini.

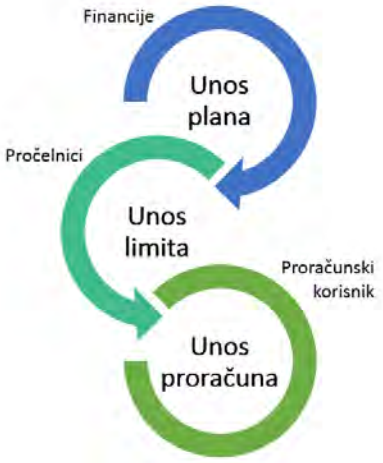
U sustav su implementirane kontrole kako proračunski korisnici ne bi mogli raspodijeliti iznose koji premašuju dobivene limite, a također i prava pristupa korisnika na određeni set informacija koje mu pripadaju.

TABLICA I. PRORAČUNSKE KLASIFIKACIJE

KLASIFIKACIJA	PODJELA KLASIFIKACIJE
Organizacijska klasifikacija	Razdjel Glava Proračunski korisnik
Programska klasifikacija	Program Aktivnost / Projekt
Funkcijska klasifikacija	Funkcija
Ekonomska klasifikacija	Računski plan
Izvori financiranja	Izvor financiranja
Lokacijska klasifikacija	Lokacija



Slika 1 Prikaz sheme poslovnog rješenja



Slika 2 Prikaz tijeka rada donošenja proračuna

Preglednost i jednostavnost korištenja web rješenja, glavne su pogodnosti za krajnje korisnike koji tijekom procesa planiranja podatke mogu odmah pratiti u obliku pripremljenih dinamičkih izvještaja na svim razinama i kombinacijama proračunskih klasifikacija. Prikaz dijela web rješenja prikazan je slikom 3.

Obzirom da su korisnici planiranje proračuna do sada radili u Excel datotekama, ovo web rješenje nudi ogromno pojednostavljenje procesa planiranja zbog automatskog pregleda podataka, tijek rada u kojem istovremeno sudjeluje velik broj proračunskih korisnika, pročelnika i korisnika financija te ugrađenih kontrola unosa zbog čega konačni proračun i njegove izmjene mogu biti donesene u najkraćem mogućem roku.

B. Realizacija proračuna

Proces realizacije proračuna sastoji se od realizacije dijela proračuna Istarske županije i realizacije dijela proračuna proračunskih korisnika.

Realizacija proračuna Istarske županije prvo se evidentira u aplikaciji urudžbenog zapisnika gdje se prilaže i skenirani primjerak dokumenta. Ako je riječ o financijskom dokumentu (računu, predračunu i sl.) isti se prenosi u aplikaciju eOvjere koja je povezana sa AX-om preko zajedničkih šifrarnika. Prijenosom u eOvjeru pokreće se automatsko prepoznavanje podataka nad skeniranim primjerkom računa OCR (Optical character recognition) metodom pri čemu se automatski prepoznaju svi važni elementi računa kako ih korisnici ne bi morali ručno unositi.

Uloga eOvjere je odobriti financijskih dokumenata kroz definiran tijek rada sustava. Svaki korisnik eOvjere ima pravo uvida u dokumente u svojoj ingerenciji. Uz skenirani dokument, korisnici imaju uvid u prepoznate financijske podatke te unose razradu po pozicijama i iznosima koje će ovaj financijski dokument teretiti.

Podaci poput imena dobavljača, OIB-a, bankovnog računa, iznosa, datuma dospijeća te drugi automatski se generiraju u eOvjeri i koriste za kasniji prijenos u AX.

Tijekom procesa odobrenja, svaki korisnik može provjeriti trenutno stanje pozicije koju je odabrao da tereti financijski dokument. Pritom ima uvid u plan pozicije, njenu trenutnu izvršenost, rezerviranost zbog prethodno unesenih ugovora ili narudžbenica te preostali iznos nad tom pozicijom.

Potpuno spreman financijski račun nakon konačnog odobrenja dokumenta od strane župana/zamjenika župana, prenosi se u AX-u gdje se automatski knjiži.

Izvorni proračun 2016 - Plan donesen

Prihodi
Rashodi projekcija
Prihodi projekcija

Pozicija	Šifra	Naziv	Funkcija	Izvršenje 2014	Tekući plan	Izvršenje 2015	Limit	Razlika	Novi plan
0	001	STRUČNA SLUŽBA SKUPŠTINE							
	00101	STRUČNA SLUŽBA ZA POSLOVE SKUPŠTINE (Plan donesen)							
	1000	Javna uprava i administracija							
	A100001	Plaće zaposlenika							
	11001	Nenamjenski prihodi i primici							
160001	3111	PLAĆE ZA REDOVAN RAD	01310						
160002	3132	DOPRINOSI ZA ZDRAVSTVENO OSIGURANJE	01310						
160003	3133	DOPRINOSI ZA ZAPOŠLJAVANJE	01310						
	A100002	Ostala materijalna prava zaposlenika							
	11001	Nenamjenski prihodi i primici							
160004	3121	OSTALI RASHODI ZA ZAPOSLENE	01310						
160005	3132	DOPRINOSI ZA ZDRAVSTV.OSIGUR.NA DODATNA PRAVA ZAPOSLENIKA	01310						
160006	3133	DOPRINOSI ZA ZAPOŠLJAVANJE NA DODATNA PRAVA ZAPOSLENIKA	01310						
160007	3212	NAKNADE ZA PRIJEVOZ, ZA RAD NA TERENU I ODVOJENI ŽIVOT	01310						
	A100003	Troškovi upravnog tijela							
	11001	Nenamjenski prihodi i primici							
160008	3211	SLUŽBENA PUTOVANJA	01330						
160009	3213	STRUČNO USAVRŠAVANJE ZAPOSLENIH	01330						
160010	3293	REPREZENTACIJA	01330						
	1004	Predstavničko tijelo							

Slika 2 Prikaz dijela web rješenja za unos proračuna

Financijski dokument je tada spreman za plaćanje generiranjem potrebne datoteke za prijenos na Internet bankarstvo.

Realizacija dijela proračuna proračunskih korisnika unosi se u web rješenje na istom mjestu gdje se planira proračun. Proračunski korisnici na mjesečnoj bazi imaju informaciju koja su im sredstva i za koju namjena doznačena (ova informacija automatski dolazi isplatom sredstava iz AX-a proračunskim korisnicima), a ista moraju rasporediti unutar doznačenog iznosa. Podaci se nakon potvrde također automatski učitavaju u AX i knjiže.

Proces realizacije proračuna potpuno je upravljiv jer svaka odgovorna osoba za svoj dio realizacije proračuna ima pravo uvida u dokumente koji terete pozicije i aktivno sudjeluje u njihovom odobrenju i povezivanju na poziciju proračunu koju će iskoristiti. Integracijom poslovnih rješenja urudžbenog zapisnika i ovjeravanja dokumenata osigurana je ispravnost unesenih podataka u svakom trenutku, a posljedično automatiziran proces knjiženja prethodno unesenih podataka.

C. Praćenje realizacije

Praćenje realizacije treba svakom sudioniku planiranja i izvršenja proračuna u svakom trenutku dati ispravnu informaciju o izvršenosti i rezerviranosti pojedine pozicije proračuna. Za navedenu potrebu je u dijelu web rješenja omogućen pregled stanja pozicije, razdjela i ugovora.

Odabirom pregleda stanja pozicije korisnici vide njen plan, realizaciju, rezerviranost i preostalu vrijednost pozicije. Također su vidljivi i dokumenti koji čine realizirani i rezervirani iznos.

Svi korisnici koji imaju pravo na navedeni pregled mogu vidjeti podatke koji su u njihovoj ingerenciji.

Kod realiziranih stavaka svaki dokument koji je prošao sustav ovjeravanja moguće je direktno pogledati u eOvjeri, a za preostalu realizaciju pozicije (npr. trošak plaća zaposlenika) prikazane su transakcije glavne knjige AX.

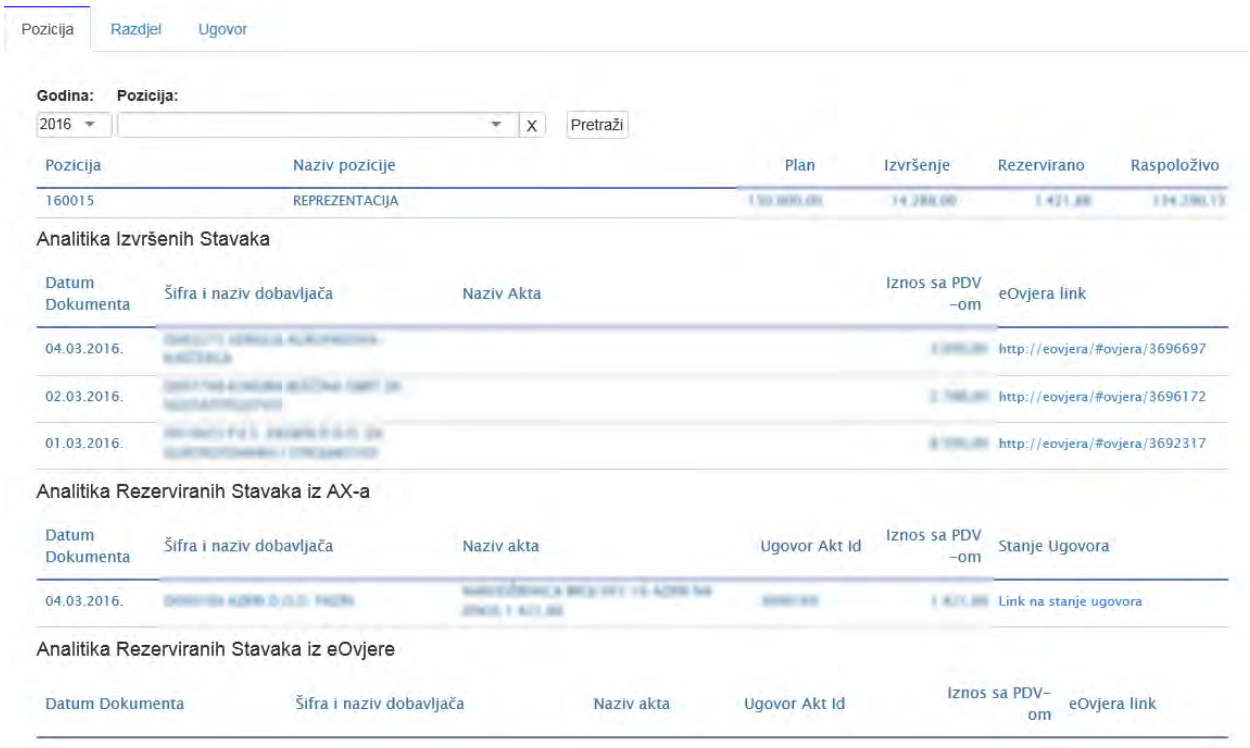
Kod rezerviranih stavaka vidljivi su podaci o dokumentima koji nisu završili proces ovjeravanja te ugovori i narudžbenice koji rezerviraju iznos koji će dolaskom ulaznog računa biti realiziran. Naravno, unos ulaznog računa u sustav eOvjere koji je realizacija određenog ugovora odnosno narudžbenice neće duplicirano rezervirati stanje pozicije, a ista je logika implementirana i za račune na temelju prethodno plaćenih predujmova.

Pregled realizacije razdjela samo donosi informaciju pregleda više pozicija odjednom, a svaka prikazana pozicija može se direktno detaljnije analizirati.

Na sličan način implementiran je pregled stanja ugovora kako bi se na jednostavan način mogla pratiti realizacija ugovora javne nabave. Prikaz dijela web rješenja vidljiv je na slici 3.

D. Izvještavanje

Izvještavanje županije ne temelji se samo na financijskim izvještajima već i izvještajima izvršenja proračuna.



Slika 3 Prikaz dijela web rješenja za pregled stanja pozicije, razdjela i ugovora

Opći dio proračun čine Račun prihoda i rashoda, Račun financiranja i Sažetak. Ova skupina izvještaja prikazuje ukupne prihode i primitke te rashode i izdatke na razini ekonomske klasifikacije

Posebni dio proračuna čini prikaz izvršenja po organizacijskoj, ekonomskoj i programskoj klasifikaciji za rashode i izdatke.

Financijski izvještaji sadrže informaciju o financijskom položaju te uspješnosti ostvarivanju postavljenih ciljeva i novčanim tokovima.

Oba seta izvještaja automatski su dostupni u AX-u jer podatke dohvaćaju iz plana proračuna i pripadne realizacije.

IV. NABAVA

Za potrebe javne nabave, nužno je osigurati praćenje podataka od plana nabave do realizacije ugovora.

Proces javne nabave započinje planom nabave koji se kreira u AX-u. Plan nabave moguće je kopirati iz prethodne godine te napraviti odgovarajuće prilagodbe, a iz plana nabave se nakon potvrde automatski kreira predmet nabave.

Nakon navedenog u AX-u je moguće kreirati nove ugovore. Ovaj proces započinje iz urudžbenog zapisnika gdje se prvo evidentira novi ugovor, a potom se njegovi podaci automatski prenose u AX. Novi ugovor definira koje pozicije očekuje teretiti obzirom da su to nužni podaci za pregled stanja pozicije, razdjela i ugovora u dijelu rezervacije pozicije.

Realizacija ugovora se prati povezivanjem financijskog dokumenta u eOvjeri sa šifrarnikom ugovora iz AX-a. Prijenosom računa u AX, podatak o ugovoru se također prenosi zbog čega je račun vidljiv u listi realizacije ugovora.

Sve navedene informacije koriste se za izvještavanje, a izvještaji poput plana nabave i realizacije ugovora su dostupni u AX-u.

V. RAČUNOVODSTVO I FINANCIJE

Standardne funkcionalnosti AX poslovnog rješenja omogućuju provođenja procesa računovodstva i financija Upravnog odjela za proračun i financije u Istarskoj županiji.

Obzirom da su ove funkcionalnosti standardno podržane važno je samo istaknuti da AX podržava načelo modificiranog nastanka događaja što znači da su prihodi priznati u trenutku kad su naplaćeni, a do tada se evidentiraju kao obračunati nenaplaćeni prihodi, dok se rashodi evidentiraju u izvještajnom razdoblju na koji se odnose. [2]

Bitan prilagođen proces ovog dijela realizacije projekta je automatiziranost učitavanja datoteke bankovnog izvatka. Učitavanjem datoteke banke sustav prepoznaje sve klasifikacije koje se prate na prihodima i primicima prema podatku iz poziva na broj odobrenja, a također knjiženjem izvoda automatski se generiraju e-mail poruke prema odgovornim osobama kojima je bitna informacija je li neki priljev pristigao za njihov projekt ili proračunskog korisnika.

VI. OTVORENOST INFORMACIJA LOKALNOJ ZAJEDNICI

Istarska županija je svjesna povećanja informatičke pismenosti lokalne zajednice i potrebe da sve više servisa i informacija bude dostupno na Internetu. [3] Iz navedenog razloga u prvoj su fazi javnosti otvoreni servisi apliciranja za javne potrebe u zdravstvu i kulturu te pregled podataka poreznih obveznika gradovima i općinama (a u narednim fazama i samim proračunskim korisnicima).

A. Javne potrebe za kulturu i zdravstvo - eObrasci

Važna funkcionalnost za lokalnu zajednicu je web rješenje za podnošenja zahtjeva za javne potrebe u kulturi i zdravstvu.

Korisnici koji imaju afiniteta prema organizaciji kulturnih događaja ili žele omogućiti realizaciju javnih potreba vezano za segment zdravstva građana mogu putem web rješenja upisati sve potrebne podatke za prijavu natječaja za dobivanje sredstava za kulturu i zdravstvo koje potom Upravni odjeli putem definiranog tijeka rada odobravaju ili osporavaju. Osobe koje su poslale zahtjeve imaju u svakom trenu mogućnost uvida u stanje svog zahtjeva.

Na temelju navedenih odluka o financiranju generiraju se ugovori koji se automatski urudžbiraju obzirom da je rješenje povezano sa aplikacijom urudžbenog zapisnika te se također takvi ugovori prenose u AX kako bi se mogli pratiti u pregledu stanja ugovora, razdjela i pozicije obzirom da i oni rezerviraju određene pozicije proračuna.

B. Web Porezna

Web Porezna je zamišljena kao web rješenje za online pregled matičnih podataka i poreznih kartica poreznih obveznika gradovima i općinama te u kasnijoj fazi samim poreznim obveznicima.

Gradovi i općine koje su zatražile da im poreznu evidenciju za gradske i općinske poreze provodi Istarska županija putem ovog web rješenja mogu vidjeti i statističke podatke o generiranim zaduženjima i pripadnoj naplati, stanju i dospijeću poreza.

U narednim fazama projekta porezna kartica bi bila dostupna i poreznim obveznicima koji bi time dobili uvid u sva zaduženja i naplatu županijskih te općinskih i gradskih poreza koji obračunava Istarska županija.

VII. PLAN ZA NAREDNI PERIOD

U tijeku je implementacija procesa poreznih evidencija i evidencije koncesija, te procesa evidencije i praćenja realizacije svih ostalih vrsta ugovora, kao i implementacija procesa eNarudžbenica u programskom sustavu AX. Za potrebu navedenih procesa kreirat će se posebni moduli koji nadograđuju standardne funkcionalnosti.

Modul Porezne omogućuje unos i praćenje matičnih podataka poreznih obveznika, zaduženje županijskih te općinskih i gradskih poreza, procesa praćenja naplate (redovne i prisilne) te pripadne statističke izvještaje.

Modul Koncesija omogućiti će evidentiranje podataka o koncesionarima, praćenje ugovora za različite vrste koncesija, generiranje zaduženja te pripadno praćenje naplate.

Modul Ugovori omogućiti će evidentiranje svih vrsta prihodovnih ugovora koje u svojim poslovnim procesima koristi Istarska županija te praćenje realizacije istih. Ovaj modul će se integrirati sa postojećim on line pregledom stanja ugovora.

Modul eNarudžbenice omogućiti će online kreiranje i odobravanje narudžbenica dobavljača te njihov prijenos u AX kako bi poslužili za realizaciju ulaznih računa i praćenje realizacije proračuna.

VIII. ZAKLJUČAK

Digitalizacija javne uprave prvi je korak u ostvarenju dostupnosti informacija lokalnoj zajednici. Sve veće potrebe za transparentnošću poslovanja javne uprave zahtijevaju adekvatan sustav praćenja poslovanja, a implementacijom AX-a i povezanih web rješenja korisnici Istarske županije, proračunski korisnici iste i lokalna zajednica dobili su izuzetnu podlogu jednostavnijem i transparentnom poslovanju.

LITERATURA

[1] Microsoft, „80300A- Introduction to Microsoft Dynamics AX 2012“

[2] Andreja Milić, Nikolina Bičanić, Dušanka Marinković Drača, Josip Šaban, Krešimir Vranar, Jasminka Rakijašić, Ivana Jakir – Bajo, “Proračunsko računovodstvo i financijsko poslovanje”, neobjavljen

[3] Božidar Rački, „IŽ Case Study Implementacija AX“, izjava korisnika, neobjavljen

Digitalna ekonomija – rezultanta disruptivnih tehnologija

Mladen Mauher
Tehničko veleučilište u Zagrebu, Zagreb, Hrvatska

mladen.mauher@tvz.hr

Sažetak - Informacijske i komunikacijske tehnologije, uz rađanje novih znanosti (*znanost o uslugama, znanost o podacima*), u destrukcijama modela tradicionalnih ekonomskih sustava, ostvaruje njenu transformaciju u digitalnu ekonomiju: područje društvenih znanosti u kojem znanstvenici proučavaju i objašnjavaju: Proizvodnju (*Industrija 4.0*) i ponudu (*sekvencioniranje i transformacija linearnih u temporalne nelinearne prostorno stohastičke instance ponude*); Raspodjelu (dio ekonomskog procesa u kojoj se rezultat *globalno kooperativne* proizvodnje raspodjeljuje na sve sudionike proizvodnje); Razmjenu (*u multidimenzionalnim interakcijama globalizacije, regionalizacije i nacionalne ekonomije*); Potražnju i potrošnju materijalnih dobara (roba) i usluga u određenom vremenu (*minimizacija vremena uz ubrzavanje društvenih procesa*), prostoru (*globalizacija*), društvu (*rastuća mobilnost*), uz kontinuirani rast kompleksnosti ekonomskih sustava i ekonomskih transakcija.

I. UVOD

Ekonomija je područje društvenih znanosti u kojem znanstvenici proučavaju i objašnjavaju: Proizvodnju i ponudu; Raspodjelu (dio ekonomskog procesa u kojoj se rezultat društvene proizvodnje raspodjeljuje na sve sudionike proizvodnje); Razmjenu (); Potražnju i potrošnju materijalnih dobara (roba) i usluga u određenom vremenu, prostoru, društvu [1].

Razvoj i interakcije tehnologije i društva u proteklih 240 godina dovodile su do tehnoloških revolucija – od „Industrijske revolucije“ (strojevi, tvornice, plovni putevi); revolucije pare, ugljena, željeza i željeznice; revolucije čelika, teške industrije (električne, kemijske, građevinske, pomorske); revolucije automobila, nafte, petrokemije i masovne proizvodnje; revolucije informacijske tehnologije i telekomunikacija. Ulaskom u 21. stoljeće društva ulaze u revolucije temeljene na biotehnologijama, bioelektronici, nanotehnologijama i tehnologijama novih materijala [2].

Nova računarska revolucija uvodi nas u prostore kibernetičko-fizičkih sustava (Cyber-Physical Systems[1]) uz vidljive trendove ubrzanja primjene „pametnih“ stvari (*Internet of Things-IoT*), njihove integracije u kompleksne sustave na globalnom nivou, te posljedične biološke evolucije[2]. Sve nas to dovodi u Četvrtu industrijsku revoluciju (*4th Industrial Revolution*).

Tehnološko-revolucijski ciklusi transformiraju cijelo društvo i ekonomiju svojim eksponencijalnim rastom i strukturnim promjenama, te kvantnim skokom u inovacijama, reverznim inovacijama i oblikovanju novih proizvoda/usluga i korespodentnih glokalnih proizvodnih potencijala.

Interakcija društva i tehnologije otvara prostore rješenja novih društvenih izazova: Zdravlje, demografske promjene i kvaliteta života; Sigurnost harane, održiva poljoprivreda i šumarstvo, istraživanje mora, pomorja i unutarnjih voda i bioekonomija; Sigurna, čista i učinkovita energija; Pametni, zeleni i integrirani promet; Klimatska aktivnost, okoliš, učinkovitost resursa i sirovine; Uključiva inovativna i promišljena društva; Sigurna društva.

II. POJAVNOSTI NOVIH EKONOMSKIH MODELA

A. *Ekonomija dijeljenja*

Ekonomija dijeljenja (Shareconomy) ili 'zajednička, suradnička potrošnja' je ideja da stvari treba koristiti umjesto posjedovati, razmjenjivati ih umjesto na njih trošiti novac, posuditi ih umjesto kupiti, poklanjati umjesto bacati. Za to su najzaslužnije društvene mreže, a njih poglavito koriste mladi, koji su naučili u virtualnom svijetu s drugima dijeliti svoje fotografije, informacije, svjetonazor i intimu, pa im ni dijeljenje stvari u stvarnom životu ne pada teško [3].

Identificira trendove u suvremenoj ekonomiji u kojoj strojevi sve više zamjenjuju ljude. Ti strojevi više nisu željezne, mehaničke naprave iz doba industrijske revolucije, već umreženi, tehnološki sofisticirani uređaji koji međusobno komuniciraju. Rifkin smatra kako dolazi doba u kojemu će granični trošak (odnosno trošak proizvodnje dodatne jedinice proizvoda) pasti na nulu uslijed pritiska konkurencije i povećanja produktivnosti i učinkovitosti. Opisuje kako Internet povezuje sve i svakoga, i kako se milijarde senzora ugrađuju u domove, urede, uređaje, vozila, proizvodne linije, logističke mreže, pa i ljude. Opadajući troškovi proizvodnje imaju dalekosežne posljedice na društvo, te milijuni ljudi već sada dijele svoja saznanja i surađuju u stvaranju novih proizvoda i usluga po relativno niskim troškovima [4].

[1] • Cyber – computation, communication, and control that are discrete, logical, and switched • Physical – natural and human-made systems governed by the laws of physics and operating in continuous time • Cyber-Physical Systems – systems in which the cyber and physical systems are tightly integrated at all scales and levels.

[2] The exponential proliferation of embedded devices (afforded by Moore's Law) is not matched by a corresponding increase in human ability to consume information! Increasing autonomy (human out of the loop), direct world access.

B. Suradnička (kolaborativna) ekonomija

Ekonomski model u kojemu zajednički dostupne tehnologije omogućuju ljudima da međusobno dijele i dobiju ono što žele. U mnogim se oblicima podudara sa ekonomijom dijeljenja.

Rastuća prihvatljivost i razvoj suradničke ekonomije vidljiva je na portalu www.meshing.it

C. Zelena ekonomija

Koncepti zelene ekonomije, odnosno kreiranja zelenih poslova i niskougljičnog razvoja, s dostizanjem gotovo 100 postotne proizvodnje energije iz obnovljivih izvora [3]. Temelji se na principima:

- Integriranog pristupa kontroli emisija – svi okolišni pokazatelji postrojenja, emisije u sve sastavnice okoliša, korištenje resursa, buka, energetska učinkovitost sprečavanje nesreća, zatvaranje postrojenja.
- Prevencije – sustavna primjena najboljih raspoloživih tehnika.
- Pregovaranja o uvjetima dozvole između nadležnih tijela, operatera i drugih dionika.
- Transparentnosti – sudjelovanje javnosti, objava dozvola i rezultata praćenja emisija.

D. Cirkularna ekonomija

Cirkularna ekonomija odnosi se na industrijsko gospodarstvo koje je svjesno obnovljivo, nastoji iskorištavati obnovljive izvore energije (OIE), smanjuje i nastoji eliminirati upotrebu kemikalija i stvaranje otpada.

Temelji se na slijedećim načelima:

- Dizajn otpada: Otpad ne postoji ako su materijali od kojih se sastoji neki proizvod namijenjeni razgradnji ili ponovnoj upotrebi, bilo u tehničkom ili biološkom smislu.
- Izgradnja otpornosti uz raznolikost: Raznolikost i prilagodljivost odlike su koje treba promovirati.
- Promicanje iskorištavanja obnovljivih izvore energije: Svi sustavi trebali bi težiti upotrebi obnovljivih i održivih izvora energije.
- Ljudska aktivnost podupire: ekosustave i izgradnju prirodnog kapitala, ljudsko zdravlje i sreću, zdravo i povezano društvo i kulturu.
- Resursi se iskorištavaju kako bi se ostvarivala vrijednost (financijska i druge vrste).
- Razmišljanje o „sustavima“: Ključno je razumjeti kako dijelovi utječu jedni na druge unutar cjeline i kako cjelina utječe na dijelove. Elementi se promatraju u kontekstu okoliša i društva.
- Razmišljanje u slojevima: Kod biorazgradivih materijala vrijednost je u načinu na koji se njihovi dijelovi mogu razgraditi ili ponovo upotrijebiti.

E. Bihevioralna ekonomija

„Bihevioralnu ekonomiju možemo definirati kao kombinaciju ekonomije i drugih društvenih znanosti koje više opisuju ponašanje. Javlja se kad ekonomisti kombiniraju istraživanja i metode iz ekonomije i drugih društvenih znanosti s ciljem poboljšavanja deskriptivne vrijednosti ekonomske teorije“ [5].

- Bihevioralna ekonomija bavi se pitanjem utjecaja psihologije pojedinca na ekonomsko odlučivanje. U stranoj literaturi se koristi i izraz „psihološka ekonomija (Brain Game ekonomija) “ ili „ekonomija ponašanja.“

F. Ekonomija utemeljena na znanju

Stupovi ekonomije utemeljene na znanju (trokut znanja) obuhvaćaju [6]:

- Obrazovni sustav
- Informacijsku infrastrukturu
- Ekonomske poticaje i institucionalni režim
- Inovacijski sustav

G. Druge pojavnosti ekonomskih modela

- Algoritamska ekonomija
- Aditivna ekonomija (*3D Printing*)
- Kriptoekonomija (*Block-Chain Economy*)

III. STRATEŠKE ODREDNICE RJEŠAVANJA DRUŠTVENIH IZAZOVA

A. Strateški pokretači Europske unije (EU)

Društveni su izazovi u programu Obzor 2020 [7] fokusirani na područja na koja će se koncentrirati sredstva i istraživačke aktivnosti za potporu ključnim ciljevima programa:

- Personalizirana zdravstvena skrb
- Održiva sigurnost hrane
- Plavi rast: realizacija potencijala oceana
- Pametni gradovi i zajednice
- Konkurentna energija s niskom emisijom CO2
- Energetska efikasnost
- Mobilnost za rast
- Otpad: izvor za recikliranje i ponovnu upotrebu sirovina
- Inovacije vezane za vodene resurse: jačanje vrijednosti vodenih resursa za Europu
- Prevladavanje krize: nove ideje, strategije i upravljačke strukture za Europu
- Otpornost na katastrofe: sigurna društva, uključujući prilagođavanje klimatskim promjenama
- Digitalna sigurnost

Obzorom 2020 nastoji se ostvariti vodstvo i industrijsko vodstvo u razvojnim tehnologijama (LEIT) koja podržava kolaborativne istraživačke i inovacijske projekte, s jakim fokusom na primijenjena istraživanja strateških tehnologija Europe, odnosno ključnih razvojnih tehnologija (KET) u područjima informacijske i komunikacijske tehnologije, nanotehnologije, naprednih materijala, biotehnologija, napredne proizvodnje i prerade, te istraživanja svemira.

[3] Obnovljivi izvori energije obuhvaćaju energiju mora, energiju vjetra, energiju iz biomase, energiju plina iz deponija otpada, geotermalnu energiju, aerotermalnu i sunčevu energiju te plin iz postrojenja za obradu otpadnih voda i bioplin.

1) Digitalna znanost

Promjene u dinamici znanosti i istraživanja omogućene su eksponencijalnim rastom podataka, dostupnošću digitalnih tehnologija te su vođene globalizacijom znanstvene zajednice kao i sve većim društvenim zahtjevima za rješavanjima velikih izazova našeg vremena. Promjene imaju učinak na cjelokupni istraživački ciklus, od početka istraživanja do objavljivanja kao i na način na koji je ciklus organiziran [8].

Digitalna znanost predstavlja radikalnu transformaciju prirode znanosti i inovacija radi integracije IKT u procese istraživanja i Internet kulture otvorenosti i dijeljenja. Ona je daleko otvorenija, više globalna i kolaborativna, kreativnija i bliža društvu. Temelji se na primjeni e-Infrastrukture za podatkovno i računalno intenzivna istraživanja u virtualnom i kolaborativnom okruženju [9].

2) Digitalni plan za Europu

Od 1995. godine informacijske i komunikacijske tehnologije (IKT) pokreću porast produktivnosti i rast u Europskoj uniji. Pojam informacijskih i komunikacijskih tehnologija obuhvaća širok spektar tehnologija, od informacijske tehnologije (IT), preko telekomunikacija, televizije i radija te svih vrsta audio i video obrade i prijenosa, sve do mrežnih kontrolnih i nadzornih funkcija. Tijekom posljednja tri desetljeća tehnološka „konvergencija" brisala je jasne granice između telekomunikacija, radiodifuzije i informacijskih tehnologija. Pametni telefoni, tableti i hibridna televizija najočitiji su primjeri te pojave. Iako je linearna radiodifuzija i dalje glavno sredstvo distribucije informacija i zabavnih sadržaja u Europi, sve je više audiovizualnih sadržaja dostupno na zahtjev, a zbog eksponencijalnog porasta 4G internetske povezivosti i „interneta stvari" (Internet of things – IoT), što uključuje povezane automobile, nosive uređaje i senzore, internet postaje sve prisutniji [10].

Digitalni plan za Europu utvrđuje 101 mjeru u 7 prioritetnih područja djelovanja [11]:

- stvaranje jedinstvenoga digitalnog tržišta,
- poboljšanje interoperativnosti informacijskih i komunikacijskih proizvoda i usluga,
- poticanje povjerenja i sigurnosti na internetu,
- osiguranje pružanja znatno bržeg pristupa internetu,
- poticanje ulaganja u istraživanje i razvoj,
- poboljšanje digitalne pismenosti, znanja i e-uključivosti,
- primjena informacijskih i komunikacijskih tehnologija u rješavanju ključnih izazova društva, kao što su klimatske promjene, povećanje troškova zdravstvene skrbi i starenje stanovništva.

3) Europske tehnološke platforme

Europske tehnološke platforme (ETPs) [12] su industrijski vođeni forumi usmjereni za ustanovljavanje kratkoročnih i dugoročnih planova istraživanja i inovacija na europskom i nacionalnim nivoima potpomognute privatnim i javnim financiranjem. Do sada su oblikovane slijedeće platforme:

- Advisory Council for Aviation Research and Innovation in Europe
- Association for R&D actors in Embedded Systems
- Aquaculture Technology and Innovation Platform
- Biofuels Technology Platform
- Construction Technology Platform
- Photovoltaic Technology Platform
- Rail Research Advisory Council
- Road Transport Research Advisory Council
- Robotics Technology Platform
- Steel Technology Platform
- Advanced Engineering Materials
- Technology Platform for Global Animal Health
- Technology Platform for High Performance Computing
- European Technology Platform for Nanoelectronics
- Technology Platform for the Future of Textiles and Clothing
- Technology Platform for Sustainable Chemistry
- Technology Platform for Wind Energy
- Technology Platform on Logistics
- technology Platform on Nanomedicine
- Technology Platform on Renewable Heating & Cooling
- Technology Platform on Smart Systems Integration
- Technology Platform on Sustainable Mineral Resources
- Food for Life
- Forest Based Sector Technology Platform
- Integral Satcom Initiative
- Manufuture
- Net!Works
- Networked and Electronic Media
- Networked European Software and Services Initiative
- Photonics 21
- Plants for the Future
- Smart Grids European Technology Platform
- Sustainable Farm Animal Breeding and Reproduction Technology Platform
- Sustainable Nuclear Energy Technology Platform
- Technology Research Platform for organic food and farming
- Water supply and sanitation Technology Platform
- Waterborne
- Zero Emissions Platform

4) Strategija pametne specijalizacije, klasteri i megaklasteri

Europska unija pokrenula je inicijativu izrade Strategija pametne specijalizacije kao novi pristup gospodarskom razvoju koji je baziran na ciljanoj podršci istraživačko razvojnim aktivnostima i inovacijama [13]. Države članice usredotočile su se na stvaranje novog modela gospodarskog rasta koji će povećati ukupnu konkurentnost EU i smanjiti razlike u razvoju između gospodarstava svojih 28 članica. Nova Kohezijska politika EU za programsko razdoblje 2014. – 2020. 4 kao „ex-ante" uvjet zahtjeva od zemalja članica identificiranje područja specijalizacije koja najbolje odgovaraju njihovom inovacijskom potencijalu, a koja su temeljena na sredstvima i sposobnostima za korištenje EU sredstava u području istraživanja, tehnološkog razvoja i inovacija. Cilj je omogućiti učinkovitije korištenje ESI

fondova i povećati sinergiju između EU, nacionalnih i regionalnih politika [14].

IV. Strateške odrednice Republike Hrvatske

A. Strategija znanosti i obrazovanja

Doprinos visokog obrazovanja zapošljavanju i rastu, kao i njegovu međunarodnu privlačnost, moguće je poboljšati uspostavom uskih i učinkovitih veza između obrazovanja, istraživanja i inovacija – tri strane „trokuta znanja“. Nedavni pomak ka otvorenim inovacijama uzrokovao je povećani protok znanja i nove vrste suradnje među obrazovnim institucijama, istraživačkim organizacijama i poduzećima [16].

Hrvatska treba razvijati nacionalni trokut znanja povezan s europskim i svjetskim znanjem, a ostvaren suradnjom obrazovnog i istraživačkog s poslovnim sektorom, posebice industrijom, i tako stvarati Hrvatski visokoobrazovni i istraživački prostor u kojem djeluju javne i privatne visokoobrazovne i istraživačke institucije te surađuju s gospodarstvom i društvenim djelatnostima [17].

B. Strategija pametne specijalizacije RH

Prioritetna područja Strategije pametne specijalizacije u Republici Hrvatskoj, s relevantnim tehnološkim i proizvodnim poljima, su: (1) Zdravlje i kvaliteta života, (2) Energija i održivi okoliš, (3) Promet i mobilnost, (4) Sigurnost i (5) Hrana i bio-ekonomija. Hrvatska je također definirala dvije horizontalne teme koje mogu pridonijeti većoj dodanoj vrijednosti hrvatske proizvodnje i potaknuti nastajanje novih gospodarskih aktivnosti, porast produktivnosti hrvatskog gospodarstva i nastanak novih i održivih prilika za zapošljavanje. Horizontalne teme su: (1) KET (ključne razvojne tehnologije) i (2) ICT (informacijske i komunikacijske tehnologije).

Strategija pametne specijalizacije Republike Hrvatske (S3) je strateški dokument koji predstavlja preduvjet za dobivanje sredstava iz ESI fondova i određuje prioritetna tematska područja ulaganja u području istraživanja, razvoja i inovacija za razdoblje 2014 – 2020 [18].

C. Strategija inovacije RH

Strategijom se dugoročno usmjerava razvoj i sustavno poticanje inovacija kao temeljne vrijednosti uspješnosti gospodarstva, ali i društva u cjelini [19].

Elementi inovacijskog sustava prilagođavaju se potrebi brzog transfera znanja i tehnologija u gospodarstvo te se u svim EU strateškim dokumentima poseban naglasak stavlja na investicije u tehnologiju i inovacije u 6 prioritetnih područja:

- KET tehnologije (eng. Key enabling technologies),
- napredne proizvodne tehnologije,
- tržište bio-proizvoda,
- čisti transport,
- pametne mreže i
- održivu gradnju i učinkovito korištenje sirovine.

D. Znanstveni centri izvrsnosti u RH

Cilj osnivanja ZCI-a je prepoznati i vrednovati istraživače i znanstvena istraživanja koja nose inovativnost, potencijal otkrića, odnosno moguću prekretnicu u znanstvenom istraživanju, a uz to su međunarodno relevantna u smislu kvalitete i vizije te usklađena sa strateškim potrebama i prioritetima Republike Hrvatske, odnosno ciljevima strategije Europe 2020 i Obzora 2020.

Svrha rada ZCI-a je unaprijediti hrvatsku znanost i ojačati njeno uključivanje u Europski istraživački prostor, odnosno potaknuti sudjelovanje u istraživačkim programima EU i drugim međunarodnim programima. ZCI treba ponuditi dodanu ukupnu znanstvenu kvalitetu koja je iznad pojedinačnih individualnih znanstvenih rezultata. ZCI na jednom mjestu okuplja interdisciplinarnu mrežu inovativnih znanstvenika i znanstvenih timova te gospodarskih i javnih subjekata koji sustavno provode istraživanje na temama od izuzetne važnosti za znanost i društvo.

Misija ZCI-a je da znanstvenim istraživanjima i njihovom mogućom primjenom pomiče granice istraživanja, znanja i društva općenito, a koje će povećati i/ili unaprijediti međunarodnu vidljivost i prepoznatljivost hrvatske znanstvene zajednice te koristiti razvoju gospodarstva i društva u cjelini [20].

U RH su ustanovljeni znanstveni centri izvrsnosti za:

- napredne materijale i senzore – CEMS
- znanost i tehnologiju (STIM)
- reproduktivnu i regenerativnu medicinu
- virusnu imunologiju i cjepiva
- integrativnu bioetiku
- hrvatsko glagoljaštvo
- školsku efektivnost i menadžment
- bioraznolikost i molekularno oplemenjivanje bilja
- bioprospecting mora
- kvantne i kompleksne sustave te reprezentacije Liejevih algebri,
- personaliziranu brigu o zdravlju
- temeljnu, kliničku i translacijsku neuroznanost
- znanost o podacima i kooperativne sustave

E. Centri kompetencija

Centri kompetencija (CEKOM-i) su individualni (umreženi) subjekti vođeni potrebama industrije, osmišljeni u svrhu pružanja podrške podizanju kapaciteta poslovnog sektora (uglavnom malih i srednjih poduzeća kojima nedostaju vlastiti kapaciteti za istraživanje i razvoj), kako bi provodili projekte istraživanja i razvoja (osobito one usmjerene na razvoj i primijenjena istraživanja i komercijalizaciju rezultata) u skladu s tematskim područjima navedenima u Strategiji pametne specijalizacije Republike Hrvatske [21]. Njihov glavni cilj je povećati konkurentnost poslovnog sektora kroz ulaganja u istraživanje i razvoj i povećanje ulaganja poslovnog sektora u istraživanje i razvoj.

CEKOM može biti konzorcij između poduzetnika i jedne ili više organizacije za istraživanje i širenje znanja uspostavljen u cilju učinkovite suradnje na projektima istraživanja i razvoja; inovacijski klaster, čiji su članovi uključeni u aktivnosti istraživanja i razvoja ili pravni

subjekt koji upravlja istraživačkom infrastrukturom koja obavlja ekonomske djelatnosti.

V. DISRUPTIVNE TEHNOLOGIJE

A. Kibernetičko-fizički sustavi

Pojmovne sastavnice: Kibernetički – računarstvo, komunikacije i umjetnom inteligencijom potpomognuti kontrolni sustavi koji su diskretni, logički i povezani; Fizički – prirodni i ljudskim radom stvoreni sustavi pokretani zakonima fizike u kontinuiranom vremenu kojima je pridružen „Internet stvari"; Kibernetičko-fizički sustavi – sustavi međusobne povezanosti kibernetičkih i fizičkih sustava na svim nivoima i obujmima.

Kibernetičko-fizički sustav je sustav međusobno povezanih suradničkih računalnih elemenata (*Cloud / Fog/ Due Communication Networks* [22]) koji komuniciraju i kontroliraju fizičke entitete korištenjem Interneta stvari i kolaborativne inteligencije (*Swarm Intelligence, i.e.).*

B. Industrija 4.0

Četvrta industrijska revolucija (često imenovana kao Industrija 4.0) sumarni je pojam temeljen na konvergenciji: Informacijskih tehnologija (Automatizacije poslovnih procesa, Automatizacija ureda), višeslojnih računalno-komunikacijskih tehnologija (Računarstvo u oblaku, Magla računarstvo, Rosa računarstvo), uvezan sa Operativnim tehnologijama (Aditivna proizvodnja, Automatizacija industrijskih procesa, Automatizacija tvornice, Umjetna Inteligencija, Robotika) i Internetom Stvari (Kibernetičko-fizički sustavi), koji obuhvaća brojne kontekstualne oblike virtualizirane automatizacije, kontekstualne interakcije i razmjene podataka u realnom vremenu između strojeva (*Machine-to-Machine* [23]) i njihovog okruženja, te uspostavu inovativnih inteligentnih proizvodnih sustava i produkata (Tvornica budućnosti, Pametna tvornica, Pametni produkt, Prediktivna prilagodba i održavanje).

Industrija 4.0 odnosi se na kontekstualnu, temporalno umreženu ekonomiju, temeljenu na primjeni IKT, osnaženu jednostavnom povezanošću pametnih jedinica proizvodnih resursa, tehnologija i procesa. Industrija 4.0 omogućava ostvarenje „Inteligentne tvornice".

Industriju 4.0 čine slijedeći arhitekturni slojevi:

Tehnološki pokretači: Auto-ID/Inteligentni senzori, Ugradbeni sustavi, Umrežavanje/Oblak, magla, rosa, Big Data analitika, Distribuirana umjetna inteligencija, Fleksibilne proizvodne metode (*3D Printing*).

Koncepti: Kibernetičko-fizički sustavi, M2M komunikacije, Decentralizirana ad hoc organizacija, Internet stvari i usluga, Interakcija čovjeka i stroja, Modularizacija/Umetni i proizvedi.

Aplikacije: Horizontalno umrežavanje (Opskrbne mreže), Vertikalno umrežavanje, Agilna proizvodnja (Upravljanje performansama), Samoregulirajuća intra logistika, Mapiranje digitalnih procesa/Virtualna tvornica, Inteligentna interakcija čovjeka i stroja, Prediktivna/preskriptivna analitika, Integrirano planiranje radnih procesa, Produktu usmjerene usluge, Nadzor stanja, Samokontrola kvalitete, Samooptimizacija.

Upravljanje performansama: Strateško upravljanje performansama, Operativno upravljanje performansama u funkcijama stvaranja vrijednosti (istraživanje i razvoj, priprema rada, proizvodnja, proizvodna logistika, održavanje, logistika), Vještine, procesi, IT sustavi [24].

VI. NOVA INDUSTRIJSKA REVOLUCIJA

A. Europske tvornice budućnosti

Tvornicu budućnosti opisuju slijedeće domene i karakteristike:

- Napredni proizvodni procesi – inovativni način proizvodnje sa postojećim i novim materijalima za postojeće i nove proizvode; kompleksne strukture, geometrije i volumeni (nanotehnologije, kontekstualno-integrirani proizvodni procesi, individualizirana proizvodnja visoko-performansnih fleksibilnih struktura; novi poslovni modeli proizašli iz disrupcije tradicionalnih poslovnih sustava),
- Adaptivni i pametni proizvodni sustavi – inovativna proizvodna oprema na komponentnom i sustavnom nivou, uključujući mehatroniku, kontrolne i nadzorne sustave (fleksibilni i rekonfigurabilni strojevi i roboti, ugradbene kognitivne funkcije, simbiotske sigurne i produktivne interakcije čovjeka i robota, pametna robotika, mehatronika i nove strojne arhitekture za adaptivne i revolutivno-razvojne tvornice, nano/mikro preciznost i makro-proizvodna oprema, multidisciplinarni inženjerski alati za mehatronički inženjering,
- Digitalne virtualne i resursno učinkovite tvornice – integrirani modeli tvornica za evolutivne proizvodne sustave, inteligentni sustavi prediktivnog održavanja, integrirano visoko performansno računarstvo u upravljanju životnim ciklusom tvornice, energetski monitoring, višeslojne simulacije i analitike za unapređenje kvalitete proizvodnih procesa, kontinuirano vrednovanje i razdioba proizvodnih rizika, tvornice reciklaže,
- Kolaborativne i mobilne korporacije – umrežene tvornice i dinamički opskrbni lanci/mreže, poslovno-proizvodna kolaboracija u oblaku, mobilna skladišta i aplikacije za agilne i otvorene opskrbne mreže, povezani objekti imovine i poduzeća u opskrbnim mrežama, kolaborativna prognostika i planiranje ponude i potražnje, upravljanje digitalnim pravima (DRM), multikorporativna kontrola pristupa temeljem uloga u proizvodnim poduzećima (MRBAC)
- Humanocentrična proizvodnja – jačanje uloge ljudi u tvornicama, korporacijski kolaborativni sustavi edukacije i učenja na daljinu, napredni informacijski modeli za kreiranje znanja i učenje, novi oblici interakcija kolaborativnog rada radnika i drugih resursa u proizvodnim sustavima, dinamička interaktivna sučelja tvorničkih radnika u dinamičkom radnom okruženju, napredna vizualizacija kompleksnih proizvodnih i podataka, povezivanje organizacijskog znanja u povezanim poduzećima,
- Kupcu/korisniku fokusirana proizvodnja – uključivanje kupca/korisnika u proizvodni vrijednosni lanac, proizvodna inteligencija za informativan dizajn

produkta, crowd sourcing za visokopersonalizirani i inovativni dizajn produkta, sakupljanje, analiza i anonimizacija korištenja produkta, mobilni uslužni kokpit za proširenu poslovnu ponudu, proizvodna rješenja za modularni, nadogradivi, rekonfigurabilan i razgradiv produkt, implementacija kreativnosti i korisničko vođene inovacije putem fleksibilnog dizajna produkta i proizvodnih procesa.

B. *Holonički poslovni sustavi*

Holon (grčki: ὅλον imenični oblik ὅλος cjeline) je nešto što je istovremeno cjelina i dio. Holon je evolucijski sustav (fenomen) samoorganizacijske raspršene strukture, sačinjen od drugih holona, čija je struktura izbalansirana između uređenosti i kaosa. Održava se razmjenom materije/energije i informacije/entropije s drugim holonima, a istodobno je cjelina koja interaktira u višem holonu. Protežu se od najmanjih subatomskih čestica i nizova do multiverznih razina sačinjavajući mnogo svjetova. Pojedinci, društva i njihove kulture su srednje razne holona sačinjenih od mnogostrukih interakcija.

Holonička mreža, u poslovnim sustavima, je skup kompanija koje sustavno i integracijski posluju i koje se konstantno rekonfiguriraju, sukladno izazovima i oportunitetima koje pred njih postavljaju kupci. Svaka kompanija u mreži ostvaruje različite procesne mogućnosti i naziva se holon. Svako konfiguriranje procesnih mogućnosti unutar holoničke mreže naziva se virtualna kompanija. Izvrsnost (trokut znanja) je temeljni kriterij za kvalifikaciju u holoničku poslovnu mrežu.

Usmjerenost ka trajnom zadržavanju i povećanju izvrsnosti svakoga pojedinog holona oblikuje novu istraživačku holoničku mrežu. Svaki holon u istraživačkoj holoničkoj mreži interaktira s korespodentnim fundamentalnim, razvojnim i primijenjenim istraživačkim holoničkim mrežama. Na takav se način ostvaruju društva/gospodarstva utemeljena na znanju.

VII. ZAKLJUČAK

Informacijske i komunikacijske tehnologije, uz rađanje novih znanosti (znanost o uslugama, znanost o podacima), u destrukcijama modela tradicionalnih ekonomskih sustava, ostvaruju njenu transformaciju u digitalnu ekonomiju: područje društvenih znanosti u kojem znanstvenici proučavaju i objašnjavaju: proizvodnju (Industrija 4.0) i ponudu (sekvencioniranje i transformacija linearnih u temporalne nelinearne prostorno stohastičke instance ponude); raspodjelu (dio ekonomskog procesa u kojoj se rezultat globalno kooperativne proizvodnje raspodjeljuje na sve sudionike proizvodnje); razmjenu (u multidimenzionalnim interakcijama globalizacije, regionalizacije i nacionalne ekonomije); potražnju i potrošnju materijalnih dobara (roba) i usluga u određenom vremenu (minimizacija vremena uz ubrzavanje društvenih procesa), prostoru (globalizacija), društvu (rastuća mobilnost), uz kontinuirani rast kompleksnosti i interakcija ekonomskih sustava i ekonomskih transakcija.

LITERATURA

[1] Polovina S. Đ. Medić: Osnove ekonomije, Medinek, 2002.

[2] Peres C.: Technological revolutions and techno-economic paradigms, Working Papers in Technology Governance and Economic Dynamics no. 20, The other canon foundation, Norway, Tallinn University of Technology, Tallinn 2009.

[3] Fellander,A., Ingram, C., Teigland, R.: Sharing Economy – Embracing Change with Caution, Entreprenörskapsforum, 2015.

[4] Jeremy Rifkin: „Društvo u kojem je granični trošak jednak nuli“ , New York Times, 2014.

[5] Weber, R.: Behavioral economics, The handbook of economic sociology, str. 91-108, 2005.

[6] Svjetska banka, http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/ECAEXT/EXTECAREGTOPKNOECO/0,,contentMDK:20422383~menuPK:921081~pagePK:34004173~piPK:34003707~theSitePK:677607,00.html, preuzeto 2016.

[7] http://www.obzor2020.hr/, preuzeto 2016.

[8] http://data.consilium.europa.eu/doc/document/ST-8784-2015-INIT/hr/pdf , preuzeto 2016.

[9] https://ec.europa.eu/digital-single-market/en/news/digital-science-horizon-2020, preuzeto 2016.

[10] http://www.europarl.europa.eu/atyourservice/hr/displayFtu.html?ftuId=FTU_5.9.3.html, preuzeto 2016.

[11] https://uprava.gov.hr/o-ministarstvu/ustrojstvo/uprava-za-e-hrvatsku/aktualni-projekti/digitalni-plan-za-europu-engl-digital-agenda-for-europe/911, preuzeto 2016.

[12] European Technology Platforms: http://ec.europa.eu/research/innovation-union/index_en.cfm?pg=etp, preuzeto 2016.

[13] EU Smart Specialization Strategy: http://s3platform.jrc.ec.europa.eu/home

[14] Europski observatorij klastera: http://ec.europa.eu/growth/smes/cluster/observatory/index_en.htm, preuzeto 2016.

[15] http://ec.europa.eu/education/policy/higher-education/knowledge-innovation-triangle_hr.htm Preuzeto 17.5.2016., preuzeto 2016.

[16] Strategija obrazovanja, znanosti i tehnologije, Hrvatski sabor, 2014.

[17] HAZU: Smjernice za strategiju odgoja, obrazovanja i tehnologije, 2012

[18] Strategija pametne specijalizacije RH: http://europski-fondovi.eu/sites/default/files/dokumenti/Nacrt%20Strategije%20pametne%20specijalizacije.pdf, preuzeto 2016.

[19] Nacrt prijedloga inovacijske strategije RH: http://www.europski-fondovi.eu/sites/default/files/dokumenti/Nacrt%20Inovacijske%20strategije%202014-2020.pdf, preuzeto 2016.

[20] Znanstveni centri izvrsnosti RH: https://www.zci.hr/hr/

[21] Operativni program Konkurentnost i kohezija 2014.-2020.: http://www.strukturnifondovi.hr/op-konkurentnost-i-kohezija-2014-2020-779, preuzeto 2016.

[22] Skala,K., Davidović,D., Afgan,E., Sović, I.,Šojat,Z: Scalable Distributed Computing Hierarchy: Cloud, Fog and Dew Computing, OJCC, V.2,I.1, 2015, u https://www.ronpub.com/publications/OJCC_2015v2i1n03_Skala.pdf, preuzeto 2016.

[23] Chen, K-C., Lien S-Y.: Machine-to machine communications: Thechnologies and challanger, Rlsevier, 2013, http://santos.ee.ntu.edu.tw/papers/2013/Machine-to-machine%20communications%20Technologies%20and%20challenges.pdf, preuzeto 2016.

[24] Sauter, R., Bode, M., Kittelberger, D.: How Industry 4.0 Is Changing How We Manage Value Creation, 2015, http://www.horvath-partners.com/fileadmin/horvath-partners.com/assets/05_Media_Center/PDFs/englisch/Industry_4.0_EN_web-g.pdf, preuzeto 2016.

S P

I n t e r n a t i o n a l C o n f e r e n c e o n

MIPRO Junior - STUDENT PAPERS

Steering Committee

Chairs: Dino Lukman, University of Zagreb, Croatia
Filip Hormot, University of Zagreb, Croatia

PAPERS

Technical Diagnosis of Basic Logic Gates

Z. Tucaković
Automatic Control and Electronics Department, Faculty of Electrical Engineering Sarajevo
Sarajevo, Bosnia and Herzegovina
ztucakovic1@etf.unsa.ba

Abstract - In this paper, fault diagnosis of elementary logic gates is presented and most common causes of gate failure with a special focus on the diagnostic tests are discussed. The practical part of this paper includes the solution implementation, and more specifically, building the testing prototype. To support this, electrical schematic of the circuit is included, as well as the final design of the device.

I. INTRODUCTION

During laboratory assignments in electronics, different problems might occur. Resistors, capacitors, transistors, as well as other components in complex devices such as logic gates might be dysfunctional or damaged. Many integrated chips have several integrated logic gates on them (i.e. gate 7400 has four NAND gates, each with two inputs). If one gate on a chip is damaged, usually another gate is used. After that, the chip is left in the laboratory and the person coming next will assume it is fully functional.

That is the reason behind checking the functionality of a chip. It can be done by setting logic zeros and logic ones on inputs of logic gates. This can be a cumbersome task, and that is the reason why automation of the process is useful, easier and faster. Automation of logic gates functionality check is the subject of this paper. The prototype of a device which checks the functionality of basic logic gates has been made and the user would be provided with the information which gates inside the chip are functional. References [1] and [2] deal with this problem.

In the second chapter, theoretical basis for basic logic gates is given. In the third and fourth chapter, causes of malfunction and generation of diagnostic tests are analyzed. Fifth chapter contains description of the prototype structure and its parts (specifically power supply, microcontroller [3], display [4], keyboard, zero insertion force socket, as well as integrated chip, whose functionality will be tested). In the sixth chapter the algorithm of testing is presented, and in the seventh is given programming code that implements the algorithm. Chapter eight contains appendices, i.e. photos of the prototype and its electronic scheme.

II. THEORETICAL BASIS FOR BASIC LOGIC GATES

The logic gate is an electronic circuit or a device, implementing a Boolean function [5]. Basic logic gates are usually NOT, AND, OR, NAND and NOR gates, and they represent logical negation, conjunction, disjunction,

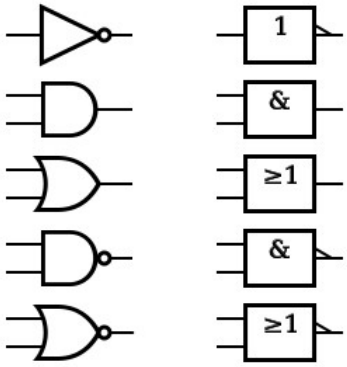


Figure 1: Common symbols for logic gates (NOT, AND, OR, NAND, NOR) and symbols according to IEC

TABLE I. TRUTH TABLE FOR ABOVE MENTIONED OPERATIONS

A	B	A'	B'	A∧B	A∨B	A|B	A↓B
0	0	1	1	0	0	1	1
0	1	1	0	0	1	1	0
1	0	0	1	0	1	1	0
1	1	0	0	1	1	0	0

Sheffer's and Peirce's operation, respectively [6].

Symbols of all these gates are given in Figure 1.

In Table I the truth tables for aforementioned operations are given.

III. CAUSES OF LOGIC GATES MALFUNCTION

A malfunction of a logic gate is a physical defect of one or more of its components, which causes malfunction of the entire system. Sometimes, a logic gate is permanently damaged. This class of malfunctions is caused by errors in element production or by the components aging [7].

The class of temporary malfunctions is caused by badly synchronized structures or asynchronous structures of digital systems, different disturbances caused by internal noise or external influences. In this class there are categories of random and categories of occasional malfunctions. Occasional malfunctions can become permanent with time, and these malfunctions are then called constant malfunctions. They usually occur in digital systems, discrete systems or in integrated circuits. They occur in the lowest levels of integration, as well as in Large-scale integration (LSI) and Very-large-scale-integration (VLSI).

However, in digital systems which use LSI and VLSI components, temporary malfunctions often occur. They are a special problem regarding malfunction detection, especially in attempts to identify what causes them. Many methods, which are formed for detection and identification of constant malfunctions can be used for diagnostics of temporary malfunctions, if diagnostics programs are processed in an on-line mode of the digital system. This applies only to digital systems, which include at least one digital computer.

In Figure 2, an implementation of a NOT gate is shown. The input is marked as *x*, and the output as *y*. If this circuit works correctly, then *y* should be negated *x*.

Possible physical defect of the NOT gate (Figure 2) and the value of output in that case are shown in Table II. In these cases, output holds the same value for both inputs (0 and 1). That value is permanent 0 or permanent 1 (p-0 or p-1), and represents a class of constant malfunctions.

Multiple malfunctions are also possible, but they are rare and can cause other types and classes of malfunctions. In essence, malfunction in a circuit also appears when a logic gate *does not* have output. That is defined by the truth table of a given operation. In this paper, the accent is not on finding the reason of malfunction, but on the identification of malfunction (if one exists).

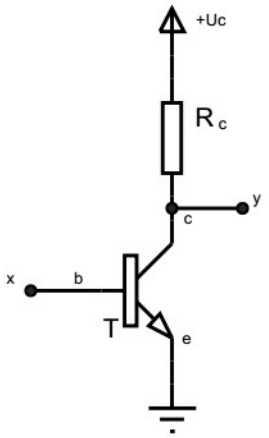


Figure 2: Possible implementation of NOT gate [7]

TABLE II. POSSIBLE PHYSICAL DEFECTS OF NOT GATE (FIGURE 2) [7]

Physical defects	Value of output *y*
Base (*b*) of transistor *T* is open-circuited	1
Collector (*c*) is open-circuited	0
Emitter (*e*) is open-circuited	1
Points *b* and *c* are short-circuited	0
Points *b* and *e* are short-circuited	1
Points *c* and *e* are short-circuited	0
Resistor R_c is open-circuited	0
Resistor R_c is short-circuited	1

IV. DIAGNOSTICS TEST GENERATION

Inspecting the functionality of logic gate varies from one gate to another. For example, if the integrated chip 7400 is inspected (it consists of four NAND gates with two inputs, and its pin arrangement is shown in Figure 3), then each of four NAND gates will be inspected as described in the following text, and results will be compared with the results of ideal, correct NAND gate.

Every method of checking logic gate functionality must set gate inputs on logic zeros or ones and compare output to the ideal, expected output (output of the functional gate). If they are not identical, inspected logic gate has a malfunction. If they are identical, next combination of logic zeros and ones should be set on inputs. This process continues until there are no combinations left untested.

There are many methods for checking logic gate functionality. In this paper, two will be considered:

- Checking logic gate functionality using sequential logic,
- Using a microcontroller.

Basic properties of each method are:

1. *Size of prototype*

Use of both sequential logic and microcontroller results in small final product, especially if a printed circuit board (PCB) is used.

2. *Connections between components*

Sequential logic demands much more components than a microcontroller, which makes the final product more complicated and increases the probability that one of used components has a malfunction. In sequential logic, a large number of flip-flops is necessary, because the final product is quite complicated automaton. On the other hand, there are less components if a microcontroller is used, but it requires a source code for it.

3. *Necessity for additional elements and the increase of the number of gates that are able to inspect*

Sequential logic cannot provide simple addition of external components (without making the prototype extremely complicated), unlike microcontroller. Keyboard, display and other elements that make the prototype more *user friendly*, can be easily added to microcontroller. The user utilizes a keyboard to specify which logic gate will be tested. Every additional component functionality must be implemented in the code, but it doesn't require much effort.

Subject to the above mentioned advantages, a *microcontroller* is chosen to be used for testing the functionality of a logic gate.

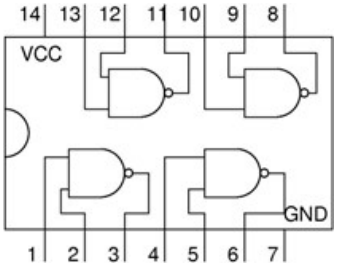


Figure 3: Pin arrangement of 7400 chip

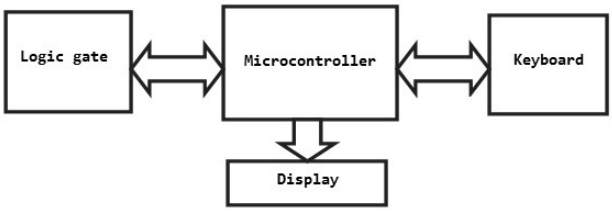


Figure 4: Block structure of the system

V. System Structure

The system consists of a microcontroller, a keyboard, a display and some additional elements, such as zero insertion force socket and an integrated chip whose functionality is tested [1], [2]. Figure 4 shows a block structure of the prototype. The central element is a microcontroller. It communicates with a logic gate, keyboard and display. Communications Microcontroller-gate and microcontroller-keyboard are bidirectional, unlike microcontroller-display which uses unidirectional communication (a display is used only for displaying data sent by microcontroller).

1. Power supply

Very important element of the developed system is power supply. For this matter, 230V/50Hz AC voltage is used. Microcontroller and logic gates demand 5V DC for their power supply, therefore AC-DC conversion is needed and the following elements are used: filter, switch, fuse, transformer, rectifier and regulator, resistors and capacitors (shown in Figure 8, lower part).

2. Microcontroller

Microchip's microcontroller PIC16F1939 is used for this purpose [3].

3. Display

Liquid Crystal Display (LCD) AV1624 of ANAG VISION is used for displaying results of logic gate testing, as well as user-machine interaction. Two rows, each with 16 characters can be displayed simultaneously [4].

4. Keyboard

Using a keyboard, a user can enter a number of a logic gate to be tested. The keyboard has 16 buttons (4 rows and 4 columns) and eight pins. First four pins represent rows, the rest represent columns. When a button is pushed, pins corresponding the column and the row of that button are on the same potential.

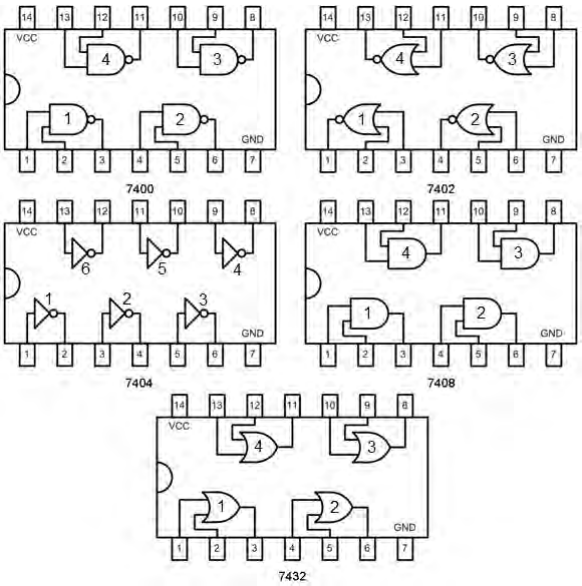


Figure 5: Pin rearrangements of gates 7400, 7402, 7404, 7408 and 7432 and gate numeration inside the chip

5. Zero insertion force socket

The zero insertion force socket (ZIF) is used to simplify usage for users. In this case, ZIF with 20 pins is used. An integrated chip, which is to be tested, is put in ZIF. Afterwards, the chip cannot be pulled out without moving the handle, there is a firm connection between ZIF and the integrated chip.

6. Integrated circuit

Testing of aforementioned gates is done for this matter. The following gates are used:
- 7400 - consists of four NAND gates with two inputs,
- 7402 - consists of four NOR gates with two inputs,
- 7404 - consists of six NOT gates,
- 7408 - consists of four AND gates with two inputs,
- 7432 - consists of four OR gates with two inputs.

Figure 5 shows pin rearrangements of listed gates.

VI. Testing Algorithm

Testing algorithm is shown in Figure 6 via flow charts [1].

VII. Source code

The algorithm is developed in programming language C, and written into the microcontroller. Code must include functions for initialization of microcontrollers' ports and initialization and testing the functionality of aforementioned gates, a function for identifying which button on keyboard is pressed (if any) and the main function in which, depending on the button pressed, certain procedures are given.

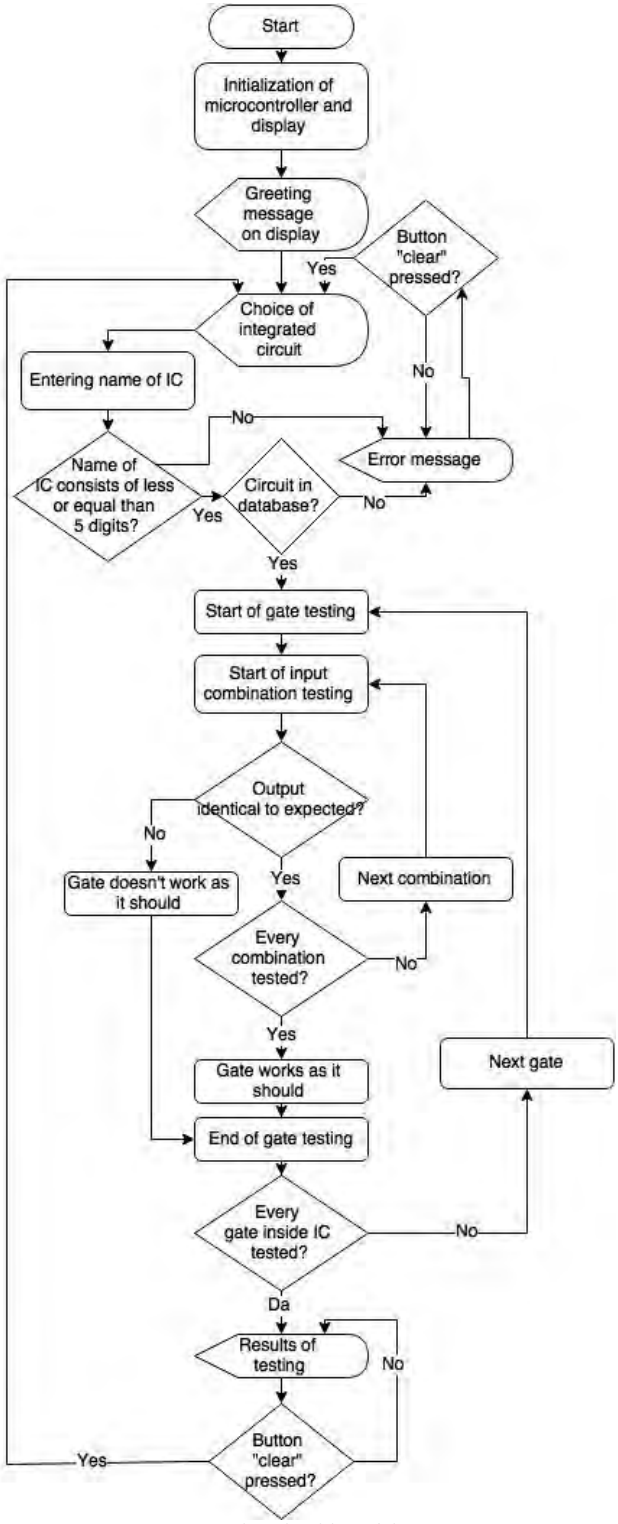


Figure 6: Testing algorithm of the prototype

A port (eight pins) connected with the keyboard is initialized as half input and half output port. Integrated chip to be tested is connected with two ports. The port connected to LCD is the output port. Pins connected to the integrated chip (via ZIF) are initialized as output pins if corresponding pin of the chip is input, and vice versa. A pin connected to power supply of the chip is declared as output. All mentioned pins are digital outputs or inputs.



Figure 7: Final design of the prototype

The procedure of functionality testing is described earlier. Every gate is presumed to be without malfunction until proven contrary. If for every combination of inputs, output of gate is identical to the expected output, then the gate is without malfunctions. If for at least one combination of inputs, the output is different, then the gate has a malfunction.

Checking whether any button is pressed is done by checking whether two pins (one from first half and the other from second) are on same potential, and that is done by setting one by one, four output pins on logic one, and testing if there is a logic one on any of input pins.

In the main function, depending on pressed button, certain procedure is given. If a user enters more than six digits that symbolize the gate's name, an error message will be written on LCD. If the gate name entered is not in database (7400, -02, -04, -08 and -32), there will be an error message.

VIII. Final Product

Figure 7 shows the final design of the prototype. In Figure 8 complete electronic scheme of the prototype is shown.

IX. Conclusion

In this paper, entire process of planning and production of electronic device is given. First, theoretical basis of logic gates, causes of malfunction and malfunction diagnosis are explained. After that, thorough analysis and planning software (programming code), as well as hardware (electronic components) is made. Functionality of code was checked, first through simulation, and later by a number of experiments on a real circuit. After numerous verifications, the final version of code is obtained. Only after that, the device could be built, and packed in a box. In that way, the prototype is simpler and safer for usage. Final product can be commercialized, and be very useful during laboratory assignments.

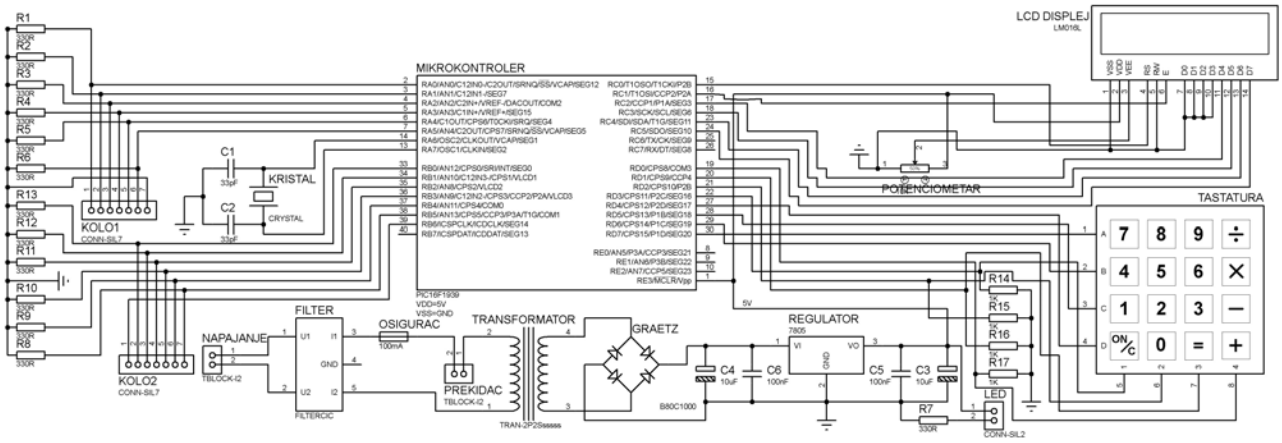


Figure 8: Complete electronic scheme of the prototype

The prototype, besides for *checking functionality of logic gates,* can be used for *identification of unknown 14-pinned logic gate from 74-series*. It is advised to implement functionalities for a larger number of logic gates through the programming code.

REFERENCES

[1] A. Bhattacharya, "Digital Integrated Circuit Tester (Using AT89s51 Microcontroller)", International Journal of Emerging Technology and Advanced Engineering, vol. 3, issue 6, pp.175-178, June 2013.

[2] A.A. Shinde, M.A. Tarkunde, "IC Tester Using 89s52 Microcontroller", International Journal of Computational Engineering Research, vol. 2, issue 7, pp.24-27, November 2012.

[3] Microchip's PIC16F1939 Microcontroller Datasheet, http://ww1.microchip.com/downloads/en/DeviceDoc/40001574C.pdf

[4] ANAG VISION's AV1624 LCD Datasheet, http://www.iq-tm.de/astro/EQ6/184594-da-01-ml-LCD_Modul_16x2_Zeichen_LED_de_en.pdf

[5] M. Ahić-Đokić, "Logički dizajn", Faculty of Electrical Engineering Sarajevo, Sarajevo, 2006.

[6] Ž. Jurić, "Diskretna matematika za studente tehničkih nauka", Faculty of Electrical Engineering Sarajevo, Sarajevo, 2011.

[7] Dž. Hasanbegović, "Osnove tehničke dijagnostike digitalnih sistema", Svjetlost, Sarajevo, 1985.

Developing a Parking Monitoring System Based on the Analysis of Images from an Outdoor Surveillance Camera

I.V. Sukhinskiy, E.A.Nepovinnykh and G.I. Radchenko
*South Ural State University, Chelyabinsk, Russia
si1en7ium@gmail.com, kwadraterry@gmail.com, gleb.radchenko@susu.ru

Abstract – In this paper, we describe a solution for monitoring of parking availability based on computer vision. It allows us to detect and track cars in a parking lot, while collected historical data helps us to predict availability status of parking during the day based on data mining techniques. Parking Monitoring System tracks availability based on an analysis of images from an outdoor surveillance camera and analyzes in real time the state of the parking complex. The system is developed to determine the number and location of available parking places and to inform the drivers. We provide an algorithm for image capture and analysis to recognize a car in a parking lot and to define parking spaces as either reserved or free.

Keywords - computer vision, object recognition, parking, OpenCV

I. INTRODUCTION

The growth of the population of major cities exacerbates the problem of traffic and, as a consequence, the provision of parking spaces. The transport system of the metropolis is enormous, requires constant monitoring, timely expansion and modernization. Organization of monitoring of such a complex system is difficult without the use of intelligent information systems.

The technical system, which carries out the collection, pre-processing of data at a low level, highlights the characteristics and laws, prepares a consolidated report and renders it for user is called intelligent monitoring system [1].

Intelligent monitoring of parking availability involves parking places detection and tracking of parked vehicles. Different types of sensors exist to gather information on the movement [2, 3]: video detectors, radars, infrared sensors, ultrasonic sensors, passive acoustic sensors, inductive loop sensors. Each type of sensor has its advantages and disadvantages. Nowadays CCTV systems are very common and there is a very wide range of tasks in this area. One of the fundamental problems is the problem of motion detection and tracking. Video detector can replace multiple inductive sensors which detect vehicles in several places and is also quite economical to maintain. However, computer vision algorithms that are used in these sensors have significant limitations in accuracy of detection of the vehicle. They are sensitive to the image distortion caused by weather events, shadows, overlapping objects, the day and night cycle and a surface stains on the camera lens [4].

This topic is relevant due to the need to monitor the status of a parking lot in real time and providing timely information to users. Timely information about the available parking spaces will enable drivers to save time searching for a parking space, as well as significantly increase the usability of parking. Such system would save time of drivers and indirectly reduce the carbon dioxide emissions produced by vehicles, due to the reduction of local traffic and by reducing the time required to find a parking space [1, 4].

We have proposed the development of a system for tracking the availability of parking spaces based on the analysis of digital images obtained with a camera in real time.

The work consists of the following parts. Part I provides an overview of principles of work of currently available systems for tracking moving objects and provides a brief overview of algorithms for movement detection. Part II contains a description and analysis of the system requirements. Part III shows the architecture of the system of monitoring of availability of parking spaces based on the analysis of images from the outdoor surveillance camera. Part IV is devoted to the implementation of the system using technologies such as OpenCV, C ++ and Angular.js. Part V presents the results of system testing. The conclusion provides overview of complete work.

II. FORMULATION OF THE PROBLEM

The task of tracking an object is defined as, given a position of its areas of interest at a given time, to find a position of its area of interest at subsequent times. The rectangular area of interest is a set of pixels of a digital image, outlining the desired object [2, 5].

The aim of this work is to develop a system for tracking the availability of parking based on the analysis of digital images obtained with digital camera in real time.

To achieve this goal it is necessary to solve the following problems:

- to study the features of the existing systems for tracking of moving objects;
- to study the features of the OpenCV - a library of image analysis [6];

- development of a system architecture for monitoring of parking availability based on the analysis of images from the camera of outdoor surveillance;
- implementation of the algorithm of transformation of an image obtained from a digital device;
- development, implementation and optimization of the algorithm for parking image analysis and parking space availability detection;
- implementation of a web application that shows the state of the parking according to a surveillance camera; testing of the application.

III. EXISTING SOLUTIONS ANALYSIS

The task of optimizing traffic flow helps to solve the service "Yandex.Traffic", developed by Yandex. "Yandex. Traffic" shows user an image of congestion. In order to achieve this a service collects data from various sources on the busiest streets, analyzes them and displays it on "Yandex.Maps" [7]. In most large cities where traffic jams are serious problem, the service expects to score jams - the average level of congestion. Technology service is set up so that the traffic information is also collected from users themselves. To participate in data collection, drivers are required: Internet-connected phone or tablet with a GPS-receiver and installed application "Yandex. Navigator" or "Yandex. Maps" with enabled traffic reporting. Every few seconds, the device transmits its geographical position, direction and speed of the computer system to Yandex.Traffic. All data is anonymous, ie does not contain any information about the user or his car.

In addition to its coordinates drivers can report more information about service failures, repairs or other road troubles.

To find a parking space, the Yandex has released the "Yandex. Parking" service which is designed to quickly find, select and pay for a parking space [8].

The program is integrated into the application infrastructure of "Yandex". Thus, on the map you can see all the nearest parking lots, and use the navigator to construct the optimal route to the selected parking space. With the help of Yandex, a user can quickly pay the parking fee from card, mobile phone account or electronic wallet. Also, the program issues a warning when the paid parking time ends. The program remembers the place where the user left the car in case it is lost. Every 10-15 minutes, the service updates data on the utilization of municipal and commercial parking lots. There is information about cost, hours of operation and other important factors [9].

TIBA company developed the line Parking Guidance System. The TSGS-30 is an electronic system, designed to assist drivers in finding quickly and easily a vacant parking space inside a parking lot. The TSGS-30 utilizes sophisticated technology, based on using Ultra-Sonic wave detector with Red-Green lights on top of each parking space. The TSGS-30 uses also electronic signage of directional arrows, guiding the driver to the nearest low-occupancy zone. Ultra-Sonic sensors placed above each parking space indicate if the space is occupied. The information is sent to the management server, which updates the electronic traffic signage accordingly. The driver will be directed to a low occupancy zone. At the appropriate zone the driver will see a green light that will indicate the available parking space [10]. The system is too complicated and can not be used in small parkings or parkings near residential buildings.

IV. ARCHITECTURE

The system of tracking of parking availability consists of two independent components: the system of capture and image analysis, and web-based interface.

The system of capture and image analysis receives a stream of frames from web cameras. Using image analysis algorithm the selection of cars in the parking lot and the definition employed / free parking spaces is made. Periodically, data on parking availability is recorded into the database.

The web interface allows users to monitor the status of parking. This web-based application is built using the architectural pattern MVC: customer request is processed by the controller, then the model extracts the required data from the database and generates a representation of the response (see. Figure 1).

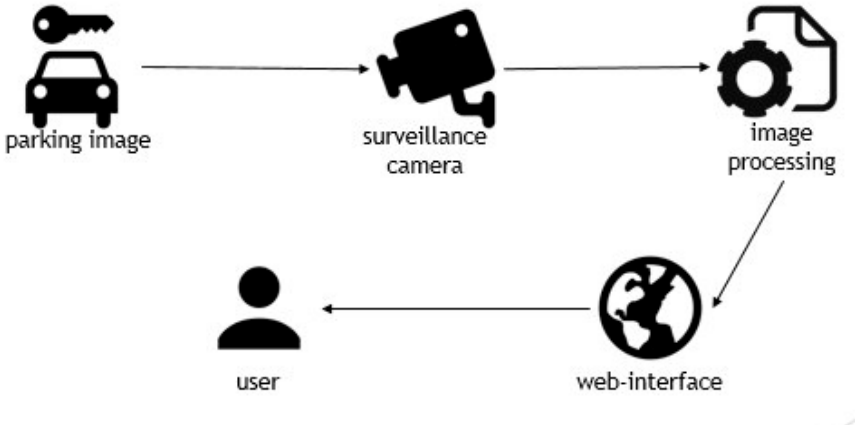


Figure 1: System components interaction diagram

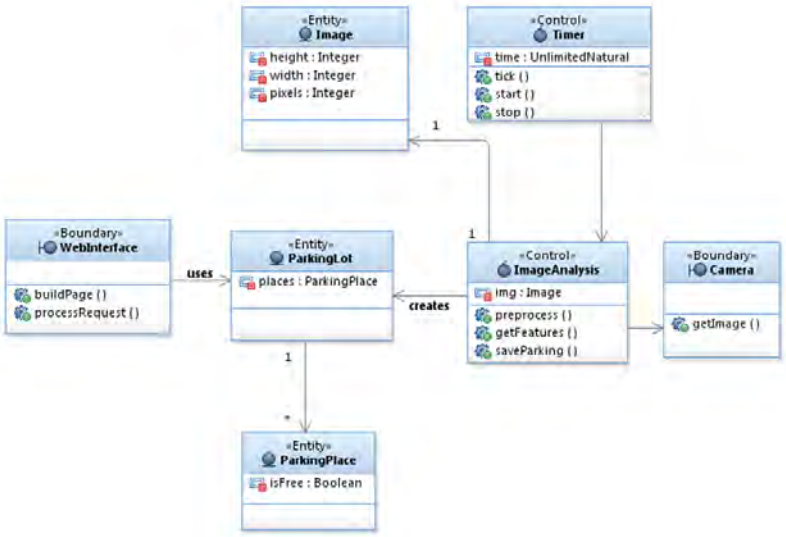


Figure 2: System components interaction diagram

During the analysis of the system requirements it was decided to allocate the following entity classes (see. Fig. 2):

- Image - the entity responsible for storing the image of parking in the computer's memory;
- ParkingPlace - essentially a parking space;
- ParkingLot - essentially a set of parking spaces.

The following are behaviour classes:

- ImageAnalisys - a class that is responsible for image processing and analysis;
- Timer - class which generates an event according to regular time intervals.

To ensure the interaction with the external environment of the system we identified the following interfaces classes:

- WebInterface - a class responsible for the operation of the web interface;
- Camera - the entity that provides access to an image from a surveillance camera

V. DEVELOPMENT

A. *Components*

The implemented system consists of two weakly coupled to each other components: a web server and an image analyzer.

The web server provides web GUI and API. Both are implemented as a web application, which provides direct access to the database. Image Analyzer with some frequency receives an image from the outdoor surveillance cameras, produces the transformation of perspective and compares it with the previous one, then locates the changed image area and updates the status of parking. New state of parking is then sent to the server.

The image analyzer is a separate application that interacts with the web server exclusively through the API.

B. *Image analyser*

Image Analyzer is an application that uses an OpenCV library [11]. After a predetermined interval



Figure 3. Image analysis result

image analyzer receives images from an outdoor surveillance camera. The transformation of the image changes the perspective of picture of parking so that cars are positioned perpendicular to the surveillance camera. This is done because calculations are simplified on a rectangular coordinate system, and because it is easier for an administrator to designate parking places on an image that went through a perspective transformation. The parameters of the transformation are initially set by an administrator. It is done by selecting 4 corners of a rectangular area on the parking lot (not on the image). The transformation is then calculated using those coordinates and applied to each image obtained from the camera [12, 13].

Next, using the resulting image comparison algorithm with the previous frame algorithm highlights potential changes in the state of parking. Each parking space that the algorithm for calculating the difference found to be changed then checked with the help of pre-trained neural network, to exclude the maximum number of false-positive results [14]. This information is converted into JSON format and sent via HTTP POST-request to the Web application. Presented here is the implementation of obtaining the difference between the two adjacent images:

```
std::vector<char> difference(const
std::vector<cv::Rect>& places, const
cv::Mat& transform, const cv::Mat& fst, const
cv::Mat& snd)
{
    std::vector<char> result;
        cv::Mat transformedFirst,
transformedSecond, greyFirst, greySecond, diff;
        cv::warpPerspective(fst,
transformedFirst, transform, fst.size());
        cv::warpPerspective(snd,
transformedSecond, transform, snd.size());
        cv::cvtColor(transformedFirst,
greyFirst, cv::COLOR_BGR2GRAY);
        cv::cvtColor(transformedSecond,
greySecond, cv::COLOR_BGR2GRAY);
        cv::absdiff(greySecond, greyFirst,
diff);
        for each (auto place in places)
        {
                auto cropped = diff(place);
                auto mean = cv::mean(cropped);
                bool changed = mean.val[0] >=
threshold;
                if (changed)
                {
                        changed =
*(classification_network.run(mat_to_array(croppe
d)));
                }
                result.push_back(changed);
        }
        return result;
}
```

The Image analysis result is shown in Fig. 3.

C. User Interface

The user interface is a web-application implemented in the Python programming language using Django web framework and Twitter Bootstrap Instruments [15, 16].

To identify ways by which you can access the Web application, there is a file ROUTES. It identifies the following ways:

- Home / parking displays all registered in the system;
- Parking / id / displays the status of parking numbered id;
- Parking / id / change / edit page displays the parking administration screen. The administrator can manually change the state of a parking space.

For each of these paths a controller and its presentation are defined.

When adding a new parking lot for the correct operation of the system administrator must perform the following steps:

1) configure video camera so that the angle was close to 90°;
2) for the correct execution of the transformation perspective, the administrator should note the angle of parking relative to the image using the mouse;
3) after performing the transformation of perspective administrator selects rectangular areas, designating parking spaces. After selecting each field must be set to be either free or busy.

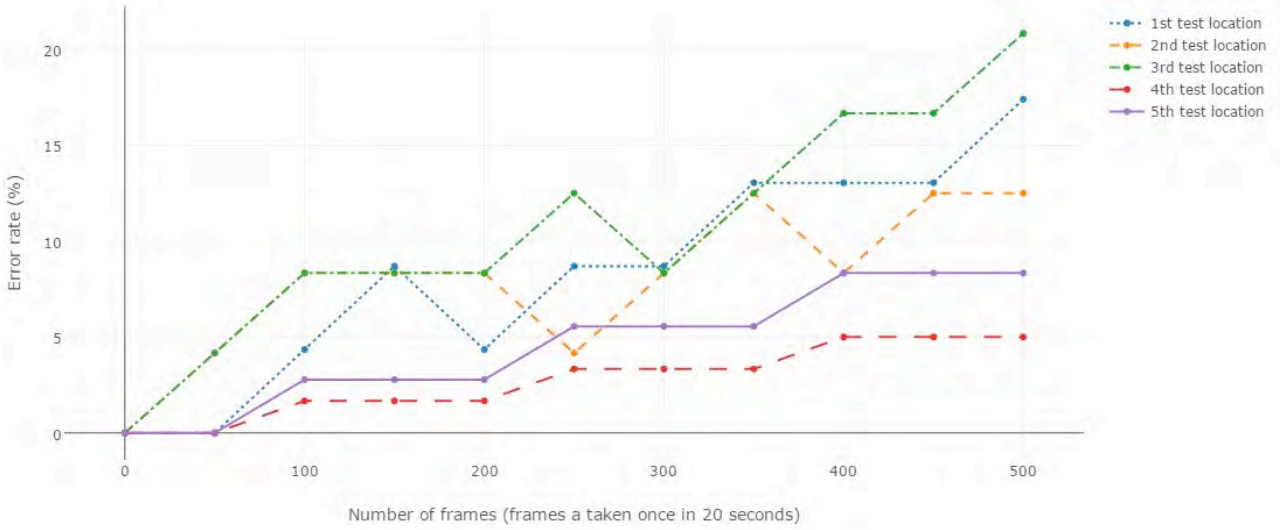


Figure 4: Image analysis testing

If the administrator has noticed an error in the calculation of free parking spaces, he can manually change the state of a parking space.

VI. TESTING

Testing of image analyzer was done using images of various difficulty levels, which reflect the status of the parking in the city of Chelyabinsk during daytime. Number of parking spaces are in the range of 12 to 70. For the testing we used the service "Online Streets" from Intersvyaz [17].

Service "Online Streets" provides access to view video from surveillance cameras on the busiest sections of the city of Chelyabinsk in real time. On some cameras a parking is clearly visible. Monitoring of these parking lots was conducted for 7 days in real time.

For the test we used five most appropriate for surveillance large car parkings in Chelyabinsk.

The results of the algorithm can be described to be good. The graph comparing the performance of the algorithm in various parking lots is shown in Fig. 4. This graph shows a correlation between the number of mistakes of the algorithm and the number of frames that the algorithm has analyzed.

VII. CONCLUSION

This article presents the design and implementation of a Web application that analyzes the state of parking lot according to a surveillance camera.

The developed system has been tested on real digital images taken with cameras of Chelyabinsk and “Online Streets" from Intersvyaz [10]. Testing showed that results of algorithm were generally good.

References

[1] Mimbela L.E.Y. A Summary of Vehicle Detection and Surveillance Technologies used in Intelligent Transportation Systems [Online]. Available from: http://www.fhwa.dot.gov/policyinformation/pubs/vdstits2007/ [Accessed 08 Feb 2016]

[2] T. Gao, P. Wang, C. Wang, and Z. Yao, “Feature Particles Tracking for Moving Objects,” *J. Multimed.*, vol. 7, no. 6, pp. 408–414, Dec. 2012.

[3] C. Goerick, D. Noll, and M. Werner, “Artificial neural networks in real-time car detection and tracking applications,” *Pattern Recognit. Lett.*, vol. 17, no. 4 SPEC. ISS., pp. 335–343, 1996.

[4] Sivaraman S., Trivedi M. M. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis //Intelligent Transportation Systems, IEEE Transactions on. – 2013. – Т. 14. – №. 4. – С. 1773-1795.

[5] Fischler M. A., Firschein O. (ed.). *Readings in Computer Vision: Issues, Problem, Principles, and Paradigms*. Morgan Kaufmann, 2014.

[6] Bradski G., Kaehler A. Learning OpenCV: Computer vision with the OpenCV library. – " O'Reilly Media, Inc.", 2008.

[7] Yandex maps, [Online]. Available from: https://maps.yandex.ru/ [Accessed 08 Feb 2016]

[8] Yandex parking, [Online]. Available from: https://yandex.ru/support/parking/ [Accessed 08 Feb 2016]

[9] Yaprobki service description [Online]. Available from: https://company.yandex.ru/technologies/yaprobki [Accessed 08 Feb 2016]

[10] TIBA Parking Systems [Online]. Available from: http://www.tibaparking.com/tsgs-30-single-spot-guidance-system-2/ [Accessed 08 Feb 2016]

[11] OpenCV library, [Online]. Available from: http://opencv.org/.

[12] R. O. Duda, P. E. Hart “Use of the Hough transformation to detect lines and curves in pictures”, *Communications of the ACM*. 15(1), pp. 11-15. 1972.

[13] X. S. Zhou and T. S. Huang, “Relevance feedback in image retrieval: A comprehensive review,” *Multimed. Syst.,* vol. 8, no. 6, pp. 536–544, 2003.

[14] W. K. Pratt, *Introduction to digital image processing*. CRC Press, 2013.

[15] Django web-framework, [Online]. Available from: https://www.djangoproject.com/ [Accessed 08 Feb 2016].

[16] Twitter Bootstrap Documentation, [Online]. Available from: http://getbootstrap.com/ [Accessed 08 Feb 2016]

[17] Street online sevice from Intersvyaz, [Online]. Available from: http://www.is74.ru/home/streets/ [Accessed 08 Feb 2016]

Laboratory Model of an Elevator: Control with Three Speed Profiles

A. Jozić, T. Špoljarić and D. Gadže
University of Applied Sciences - Department of Electrical Engineering, Zagreb, Croatia
andrej.jozic1992@gmail.com, tomislav.spoljaric@tvz.hr, davor.gadze@tvz.hr

ABSTRACT - In this article a control of laboratory model of an elevator with three different driving speeds is described. Physical model that simulates the elevator consists of an elevator car driven with small DC motor, track with 6 mechanical switches used for position of an elevator car and control section: speed, direction and safety relays with Siemens PLC S7-300 as a main control device. The purpose of this type of control is to show the use of a different speed profiles for simulation of comfort elevator ride.

I. INTRODUCTION

Since its invention an elevator has been a challenge for engineers. One of the major problems in control design of an elevator motion was a comfort ride requirement. Physical conditions, passenger interaction and safety rules needed to be taken into account for a quality ride and the definition of comfort for passengers using an elevator.

Automatic control of an elevator consists of controller and actuating elements that operate according to information provided by sensors such as position sensors and speed measurement devices. Algorithm that is provided in control device has a complexity level proportional to the type of control design. Performance rating includes information such as floor-to-floor time, speed, travel positions and car jerk rates. Usually, a PLC device is used as a controller, and actuating elements are relays, variable speed drives and motors connected with ropes to a car.

Position reference control in such control design operates according to car position and uses position sensors that operate in binary manners (On/Off). More complex controls require also speed sensors (resolvers or encoders) and its algorithms are also more complex. They include car driving methods based on mathematical algorithms and are based on feedback control with information on position from sensor, speed (either as first derivative of position or direct information from speed sensor), acceleration (first derivative of speed) and jerk rate (first derivative of acceleration).

For the purposes of laboratory exercises in a course of plant automation design a small laboratory model of an elevator was introduced. Using only binary type of signals, this model provides good teaching ground on position control while giving students appropriate information on basic elements of automated elevator car driving. Several speed profiles in such design serve as an introduction to passenger comfort basics. Simplicity of a model is acquired by using DC motor speed control with discrete type: only three speeds were used. PLC device as an elevator controller is used for adequate student learning in LADDER programming and introduction to process computing.

II. LABORATORY ELEVATOR MODEL – AN OVERVIEW

DC motor with belt transmission was chosen as a driving force for an elevator car. DC motor is used to convert electrical into mechanical energy, and a belt transmission strapped to motor shaft is used to secure linear mechanical movement of elevator up or down, depending on the order given from a control device. Speed change of a DC motor is maintained with help of electromechanical relays. These relays are used as an actuating devices for speed selection, safety and drive direction. In table I., an overview of devices used for completing a laboratory model is described.

TABLE I. DEVICES USED IN LABORATORY MODEL OF AN ELEVATOR

Device designation	*Device type*	*Assignment*
-M1	DC motor	Elevator car drive
-KG, -KD	Electromechanical relay	Direction switching
-KOG, -KOD		Top/bottom level
-K1, -K2, -K3		Speed selection switching
-S1, …, -S6	Microswitch	Position information
-S7, -S8	Pushbutton	Safety release switch
-D1, …, -D5	PLC	Controller device

Devices used in model shown on Fig. 1 are:

- 9VDC motor, shown on Fig. 3,
- 24VDC electromechanical relay with four C/O contacts, shown on Fig. 3,
- Siemens S7-300 PLC device with two digital signal modules (DI and DO module), shown on Fig. 1,
- Electronic printed circuit board (PCB) with two pushbuttons, two electromechanical relays (safety top/bottom indication and release switches, 2 change-over (C/O) contacts) and electronic components for voltage step regulation (indirect speed discrete control via three steps: 3V/6V/9V), as seen on Fig. 2.



Figure 1. Overview of a laboratory model of an elevator



Figure 2. PCB board



Figure 3. Electromechanical relays and DC motor

III. Description of Electrical Connections

A. *Digital outputs for movement direction and speed change*

Direction of the elevator car is selected via relays –KG (up) and –KD (down). Coils of these relays are connected to PLC's digital output module. Electrical connection is made in a way that safety measures are respected: if a car is in top position, relay –KOG indicates top position and, by disconnecting, negates the possibility of further upper movement. Respectively, a bottom position safety relay operates in a same manner.

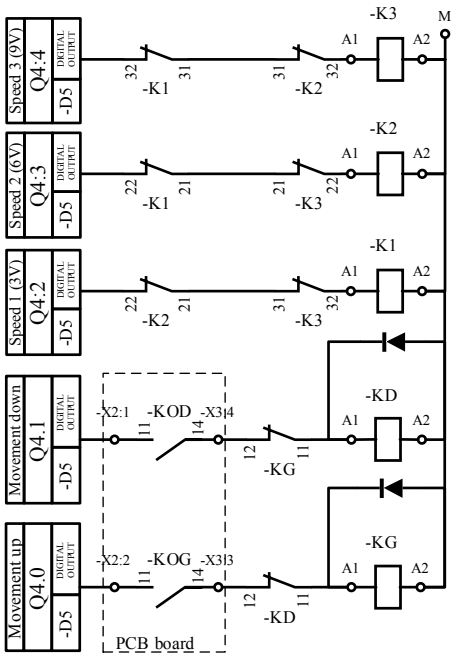


Figure 4. Schematic diagram for electrical connection of elevator's direction and speed selection

Elevator drive speed is selected with electromechanical relays –K1, –K2, –K3. They operate exclusively, which means that when a lowest speed is selected (–K1, voltage of 3V applied to motor) two other speed profiles (–K2 and –K3) cannot be operative until a request for speed profile change is applied by controller device. When selecting a medium speed profile, voltage of 6V is applied to motor, and 3V and 9V speed profiles are excluded. Also, when the highest speed profile is selected via controller, lowest and medium speed profiles are excluded and a 9V voltage is applied to motor. Since a small DC motor is selected for elevator car driving, simple speed change of rotor through three steps is obtained by discrete regulation of motor's armature voltage. Speed profiles operate independently of direction. Conditions for elevator car motion include activation of two relays: one for direction, and one for speed profile selection. Schematic diagram for movement direction and speed profile selection is shown on Fig. 4.

B. *Digital inputs for signalling elevator car level*

Position information is acquired through contacts of microswitches –S1, …, –S6, shown on Fig. 5. When the elevator car approaches the switch of a certain floor level, the microswitch clicks and its NO (normally open) contact closes. Previous level is released at a same time and its NO contact opens. These contacts are connected to PLC's digital input module and serve as an elevator car position information. Schematic diagram for microswitch electrical connection with PLC's digital input module is shown on Fig. 6.



Figure 5. Position of microswitches for upper three floors on a model

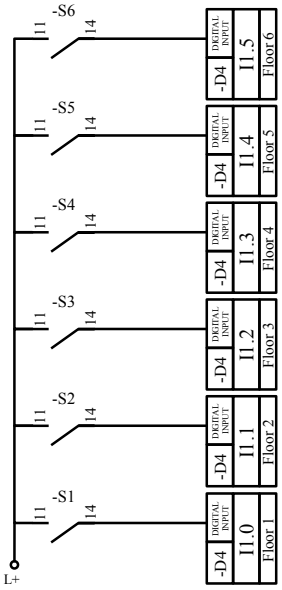


Figure 6. Schematic diagram for electrical connection of elevator's position indication (floors 4, 5 and 6)

C. Digital inputs for safety measures

Safety of the model is acquired with relays –KOG (top level indication, forbids the elevator car movement up), –KOD (bottom level indication, forbids the elevator car movement down). Coils of these relays are connected to top and bottom microswitches as shown in Fig. 7. Pushbuttons –S7 and –S8 serve as release switches: when the elevator car moves out of the forbidding position, clicking the pushbutton releases the system out of the safety state. Safety signals are connected to digital input module of the PLC.

D. Digital inputs for movement signals

Two digital inputs are used for acquiring information about elevator car direction. NO contacts from relays –KG and –KD are connected to digital inputs module and are used only for information: if a relay fails to activate after controller request, input signal will not be activated, and an error will be reported. Electrical connection is shown in Fig. 7.

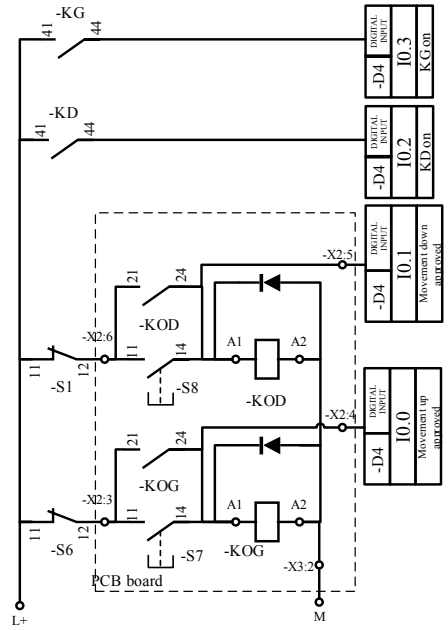


Figure 7. Schematic diagram of electrical connection for movement signals

IV. Three Speed Profiles: Ladder Program Flow Chart

In PLC program flow works through cycles. It begins with setting a PLC in run mode. Every cycle repeats then until a PLC is set in stop mode, or is turned off. Program flow for described elevator begins with input signal acquisition.

A. Input signal acquisition

In a first network of LADDER code MOVE block was used as a tool of input value. An activating input signal of PLC is determined and appropriate value is sent to memory word MW0. Input values are signals from microswitches. By activating, a switch will send an input signal to PLC by detection of accessed elevator level. Inputs are defined as I1.0, I1.1, I1.2, I1.3, I1.4, I1.5. Accessed input activates its MOVE block which stores a value of accessed floor (i.e. value 1 for floor 1, value 2 for floor 2, etc.) to memory data block MW0. Memory bit M100.0 gives confirmation of successful move operation First network's flow chart is shown in Fig. 8.

Second network proposes a similar program logic, but with few modifications. Another PLC inputs (I2.0, …, I2.5) are commands given by passenger and they present a floor on which he wishes to access. A MOVE block is defined in similar way as in floor detection and serves as position reference. To avoid confusion in execution of a program a different memory address MW2 is used for storing values of referenced floors. Memory location M100.1 is used as confirmation for entering the right value in memory location MW2. Status bit M100.1 is also used as SET command for reading status, so the passenger cannot give two commands at the same time and is reset at the end of the cycle. Position referencing is shown in Fig. 9.

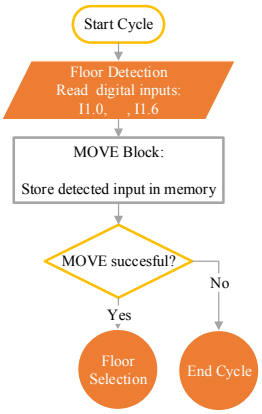


Figure 8. First LADDER network flow chart: floor detection

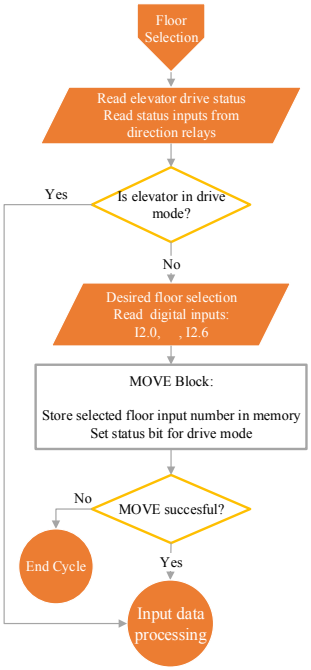


Figure 9. Second LADDER network flow chart: desired floor acquisition

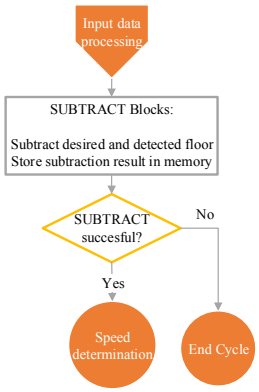


Figure 10. Third LADDER network: SUBTRACT blocks

B. *Input data processing*

Third network introduces SUBTRACT block to program. This block subtracts values of reference and real floor position and stores them in memory location. Two SUBTRACT blocks are used and this gives later a desired direction information. First block subtracts desired (MW2) from detected floor level (MW0), and a result is stored in memory (MW4). Second block subtracts detected (MW0) from desired floor level (MW2), and a stores result in another memory address (MW6). This way two same values are stored but with different sign. Direction of an elevator can be activated by determining a positive sign in these two values. Subtraction operation is shown on Fig. 10.

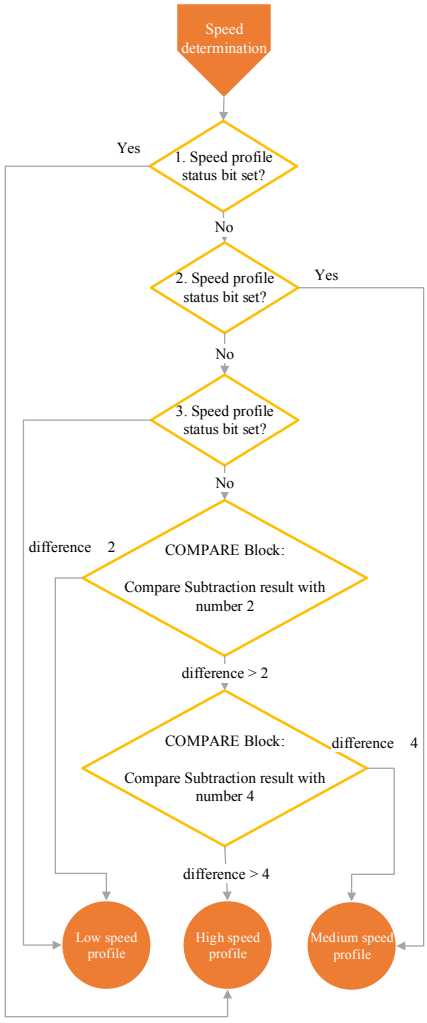


Figure 11. LADDER code flow chart for speed profile determination

C. *Determination of adequate speed profile*

A difference between real and referenced floor is needed for speed profile determination. In this LADDER network COMPARE blocks are used. First, all speed profile status bits are being read. If one of these bits is set, program flow goes to adequate speed profile. If these bits are reset, initial program run uses COMPARE blocks. These blocks compare values stored as subtraction results (selected – detected floor difference) with values two and four (Fig. 11). Low speed profile activates only relay for first speed through output Q4.2. Medium speed profile activates also an output Q4.3 for second speed, and high speed profile activates also a third speed relay through output Q4.4.

Logic of activation of appropriate speed relay works according to two conditions: distance from initial floor (acceleration mode) and distance from selected floor (deceleration mode). Details on activating appropriate outputs through acceleration or deceleration are shown on Fig. 12, Fig. 13, and Fig. 14. for all three speed profiles.

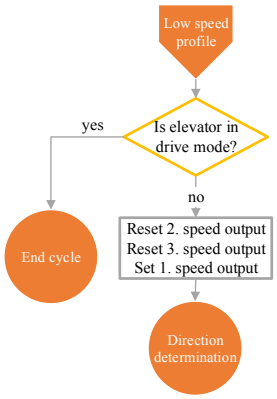


Figure 12. Program flow chart for low speed profile

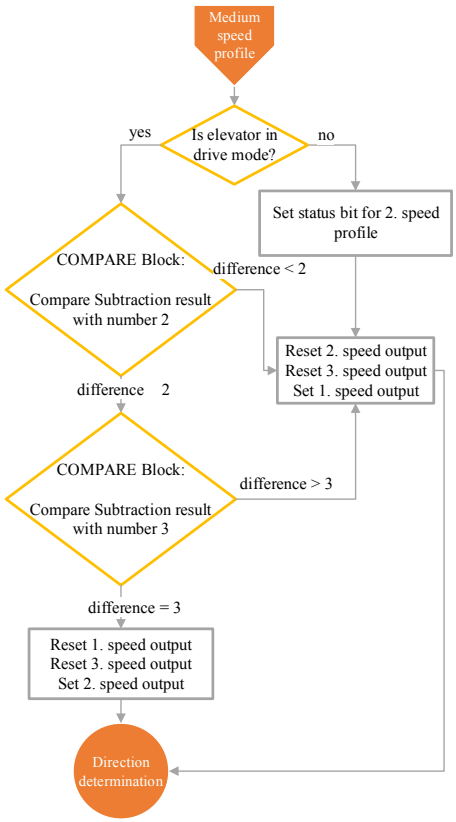


Figure 13. Program flow chart for medium speed profile

D. Determination of an elevator driving direction

In last lines of LADDER code COMPARE block is used to activate coil of a relay for upper direction through digital output –Q4.0 (moving up) or a coil of relay –KD through digital output –Q4.1 (moving down). Movement is set by difference greater or equal than one (value of 1). The use of differences throughout the program is best demonstrated in this last code since motor would not be activated without appropriate direction determination. This logic is shown on Fig. 15 through flow chart.

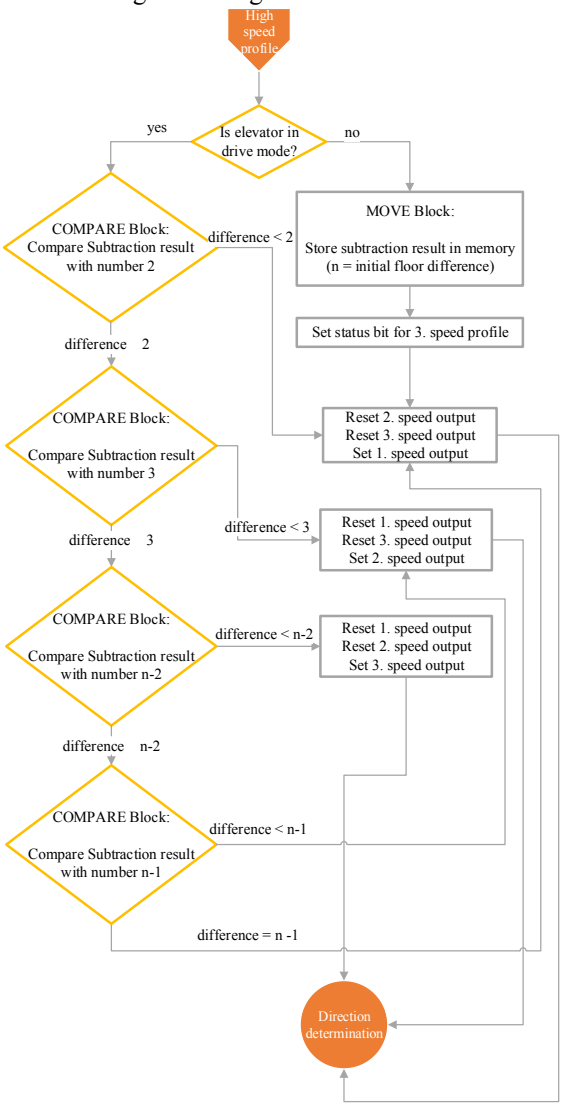


Figure 14. Program flow chart for high speed profile

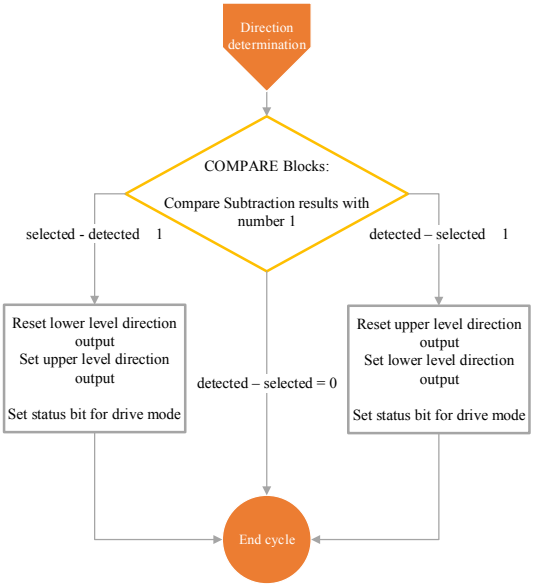


Figure 15. Program flow chart for selection of elevator direction

Program cycle ends with COMPARE block. If selected and detected floors are same, status bits and outputs are being reset and elevator stops at the desired location. If not, program cycle repeats until such state is obtained through LADDER program execution. After reaching the desired floor, program cycling flows only through reading inputs, until a different floor level.

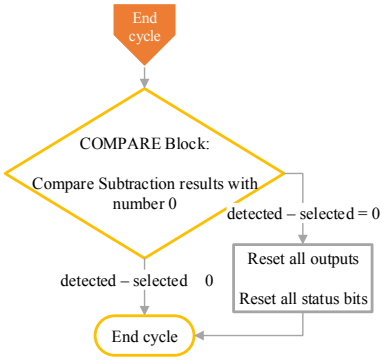


Figure 16. Program end cycle flow chart

E. SCADA settings

SCADA system is used for giving orders to the program, and signalling the current floor. Graphical interface for SCADA shown on Fig. 17 is made in WinCC Flexible® software and consists of signalling lamps showing a real floor position and buttons presenting a desired floor reference. For the purposes of this article, interface was made for PC-PLC connection via Ethernet protocol and can be run through WinCC Runtime.

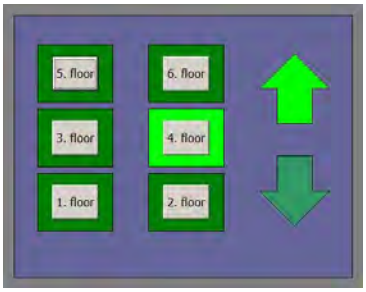


Figure 17. Runtime example of SCADA graphical interface in WinCC Flexible®

V. TESTING AN ELEVATOR – SPEED PROFILE AND ELEVATOR POSITION GRAPHS

The results of testing are shown in the next three graphs for three different differences between reference and real position.

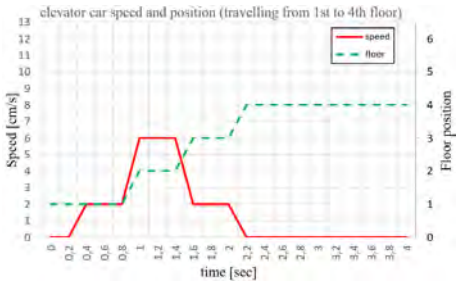


Figure 18. Time diagram of a first test: travelling from 1st to 4th floor (2. speed profile)

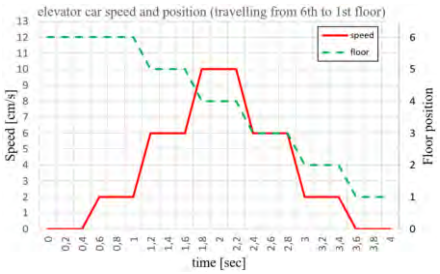


Figure 19. Time diagram of a second test: travelling from 6th to 1st floor (3. speed profile)

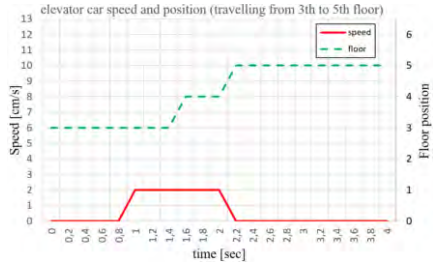


Figure 20. Time diagram of a second test: travelling from 3rd to 5th floor (1. speed profile)

Graphs of testing results (shown on Fig. 18, Fig. 19 and Fig. 20.) show successful operation of an elevator according to difference between selected and detected floor level in both directions and for all three speed profiles (first speed was 2 cm/s, second was 6 cm/s, and third was 10 cm/s).

VI. CONCLUSION

In this article an automation process for elevator with different speeds is proposed. Since a comfort ride is defined by slow acceleration in starting and slow deceleration in arriving, three profiles in speed are used to simulate quality elevator driving. A control technique for LADDER program for such operation is proposed through program flow charts, and a model containing only binary type of signals is built for that purpose.

Additional development on a proposed model can be done by implementing an analogue output signal connected to analogue-to-PWM converter for motor drive, but this involves also the expansion of a proposed PLC device with one such analogue module.

For the purposes of student learning in simple process computing this model, along with a proposed programming request, has proven to be a good challenge in practice for students.

REFERENCES

[1] Jozić A., "Laboratorijski Model Dizala sa PLCom S7-300 i tri profila brzine", završni rad br. 1766, TVZ, Zagreb, 2015.

[2] Boulter B. T., "Elevator Modelling and DC Drive Speed Controller Design – Case Study", ApICS LLC., USA, 2000.

[3] Ford P. J., "Electric Elevator Drive with Position Control", thesis for BD, LSU, USA, 2012.

[4] Gadže D., "Bilješke s predavanja iz kolegija Automatizacija Postrojenja", interna skripta, TVZ, Zagreb, 2011.

[5] "Siemens SIMATIC Step 7 Programmer's Handbook", www.plcdev.com, 2011.

Security and privacy in an IT context – a low-cost WIDS employed against MITM attacks (concept)

N. Poljak, M. Ševo, I. Livaja
Veleučilište u Šibeniku, odjel Menadžmenta, Trg Andrije Hebranga 11, 22000 Šibenik, Hrvatska
npoljak@gmail.com, marko-10@hotmail.com, ilivaja@vus.hr

Abstract – Contemporary rates of WiFi (802.11) adoption have turned the technology into an oft-underestimated source of potential security breaches. Even though the lay user base typically feels protected by employing basic encryption (first WEP, then WPA and currently WPA2), the truth is that it is trivially easy to perform a multitude of attacks on typical home and small business setups. This paper will cover the danger of a MITM[1]) attack as well as fairly straightforward steps one might take to prevent such intrusions by implementing a basic WIDS[2] setup with off-the-shelf hardware and open source software.

I. INTRODUCTION

WiFi has become ubiquitous both in homes and business environments. While larger business may have protocols and standards with regards to handling wireless security, even most tech-oriented subjects feel adequately secured by employing basic WPA2 encryption and rotating the key at random intervals.

However, because there is no requirement for physical access in order to breach a perimeter of a wireless network, such setups are extremely vulnerable to Man-in-the-middle (MITM) attacks.

II. MAN-IN-THE-MIDDLE ATTACKS

MITM (Fig. 1) is a type of eavesdropping attack perpetrated by a malicious individual, organization, or even a state-sponsored entity.

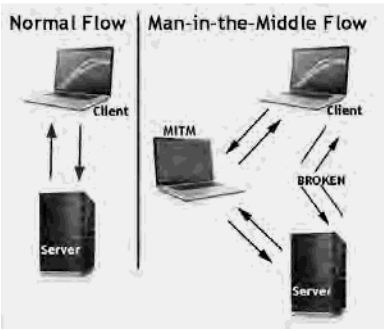


Figure 1. Diagram of basic MITM mechanics

As per Wikipedia:

"In cryptography and computer security, a man-in-the-middle attack […] is an attack where the attacker secretly relays and possibly alters the communication between two parties who believe they are directly communicating with each other. […] This is straightforward in many circumstances; for example, an attacker within reception range of an unencrypted Wi-Fi wireless access point, can insert himself as a man-in-the-middle."[3]

Indeed, it is quite trivial for a malicious agent to perform a MITM attack regardless of the type of encryption employed by the network operators. The lay mainstream idea of encryption being a safe modality of ensuring privacy is only true insofar as we're in the domain of small, private networks. In such networks it's possible to safeguard the knowledge of the encryption key. Even in perfect circumstances, however, it is emminently feasible to crack even WPA2 encryption – especially via GPU assisted computation. As soon as one's domain is a medium-sized home or business network, the encryption key is no longer sacrosanct and as such can be freely utilized by an outside attacker. This, in fact, is *always* the case with rudimentary publicly shared WiFi, such as in cafes, restaurants or even hotels. Even while sparing ourselves from a large mobile data bill, we should take into account the risk of MITM attacks while connecting to such public hotspots.

A. Step-by-step 'Man-in-the-middle'

The first step for an attacker is to set up a rogue AP (Fig. 2).

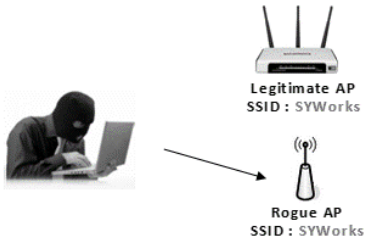


Figure 2. MITM 1/3 – setting up a rogue Access Point

[1] Man-in-the-middle
[2] Wireless Intrusion Detection System

[3] https://en.wikipedia.org/wiki/Man-in-the-middle_attack (as of 2015/07/13)

Ideally, the rogue AP should be positioned physically close to target clients to make sure there are no problems with signal quality. Later on, signal strength can also be a factor when the rogue AP „competes“ with the legitimate AP. The rogue AP's SSID and encryption are set to match the legitimate AP. However, the client may in most cases remain connected to the legitimate AP, regardless of the signal competition by the rogue AP.

The attacker proceeds to ensure that the legitimate connection between the client and AP is dropped (Fig. 3). This is accomplished through signal jamming (indiscriminate broadcasting on multiple channels) and a *deauth*[4] flood.

A deauth flood entails sending deauthentication frames – typically using open source software such as Aircrack-ng[5] (this tool can also be used to crack an encryption key in the first place). Once the connection is dropped, it is not difficult to ensure a good chance that the client will switch to the rogue AP.

It is important to continue to provide a clear, strong, functionally identical alternative signal – including MAC spoofing - to the one from the legitimate AP, which, due to the jamming implemented earlier, now has a weaker effective signal to the client.

Unbeknownst to the afflicted user (who may or may not experience a simple signal drop-off), the target device now connects to the rogue AP (Fig. 4).

At this point the malicious party can reliably intercept, analyze and/or modify all incoming and outgoing data traffic.

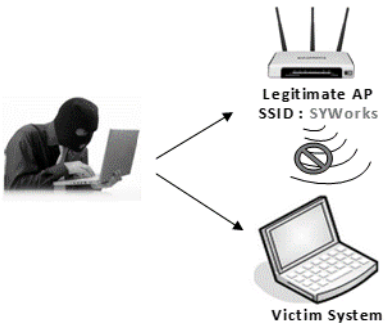


Figure 3. MITM 2/3 – Signal jamming and ‘deauth’ flood

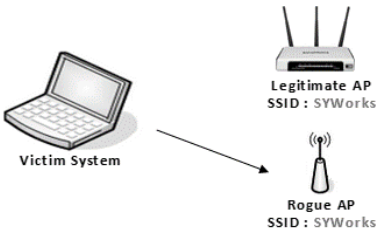


Figure 4. MITM 3/3 – Target device is connected to rogue AP

[4] Deauthentication
[5] http://www.aircrack-ng.org/

III. WIRELESS INTRUSION DETECTION SYSTEM – 'WIDS'

There are multiple approaches to defense against MITM attacks, but a system fully transparent to clients is always preferable. Ergo, one should employ a form of WIDS.

Put simply, WIDS (Fig. 5) is designed to detect potential attackers and vulnerabilities in the vicinity of a particular WiFi setup.

Again, as per Wikipedia:

‘A wireless intrusion detection system (WIDS) monitors the radio spectrum for the presence of unauthorized, rogue access points and the use of wireless attack tools. The system monitors the radio spectrum used by wireless LANs, and immediately alerts a systems administrator whenever a rogue access point is detected. […]
In addition to intrusion detection, a WIPS[6] also includes features that prevent against the threat automatically. For automatic prevention, it is required that the WIPS is able to accurately detect and automatically classify a threat.’

WIDS can be expected to detect any combination or all of the following:

- Rogue APs
- DoS attacks
- MAC spoofing
- Misconfigured APs

WIDS/WIPS systems are a growing requirement for large enterprises and organizations and as such are often implemented by default. Costs are not prohibitive – provided as SaaS, prices reportedly drop to a few “dollars per floor per month”.

For private and small-business use, however, it is neither cheap nor simple.

In the next part we will explore a simple way of setting up an Intrusion Detection System (IDS) with off-the-shelf equipment and free, open source software.

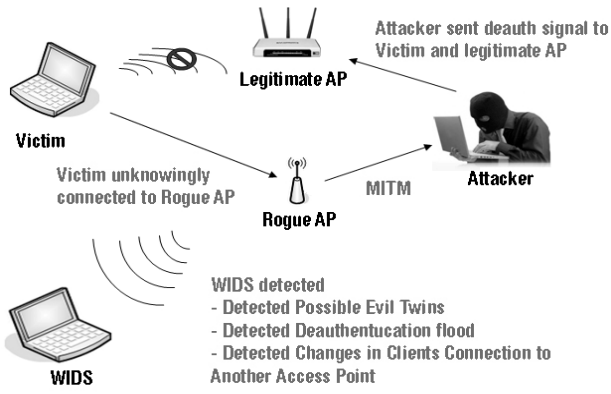


Figure 5. WIDS explained

[6] Wireless Intrusion *Protection* System

IV. HOMEMADE 'WIDS'

Required software:

- OpenWRT
- Kismet
- Snort
- Wireshark

Required hardware:

- One or more OpenWRT-capable APs (preferably dual band)
- Server capable of running required utilities

OpenWRT is an operating system designed to run on a multitude of embedded devices for routing network traffic. It provides enhanced capabilities for many routers and APs. During our process, the original router firmwares are replaced with the latest version of OpenWRT (currently Chaos Calmer v15.05)

The OpenWRT project initially came into being due to the Linksys WRT54G router (Fig. 6), the firmware for which the company built using publicly available code licenced under the GNU GPL.[7] Under the terms of the license, Linksys was obligated to also provide the source code of its end-product under the GPL, which enabled further innovation by third party, independent developers.

Running OpenWRT allows us to utilize the tools necessary for network monitoring.

The first software to set up is Kismet, which is the most widespread open source wireless monitoring tool. Kismet analyzes packets in the 802.11 (a/b/g/n) layer. Users may implement pseudorandom weighted channel-hopping lists to avoid sequences of overlapping channels, but also to prevent a malicious party from avoiding detection through concurrent channel hopping.

Kismet (Fig. 7) consists of three modules:

- kismet-drone
- kismet-server
- kismet-client

Kismet-drone is installed on all relevant APs with which we want to analyze network activity in the vicinity. Each drone is set up to forward collected data to the server, which collates and processes the data, and logs it if so instructed. Since it has a single purpose, Kismet-drone is a lightweight piece of software designed to run on the most basic devices.



Figure 6. Linksys WRT54G series

[7] GNU General Public License

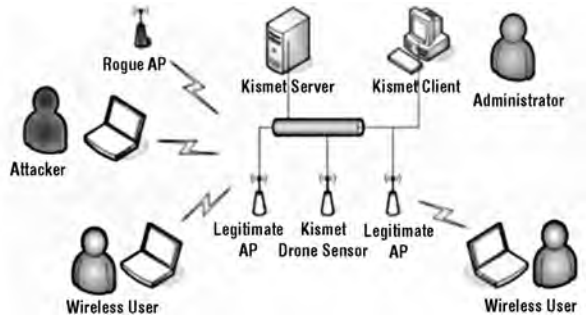


Figure 7. Typical Kismet architecture

Kismet-client is used as the interface to the server.

If we instruct Kismet to save a detailed log file, we can instruct other tools such as Snort or Wireshark to perform further analysis for threat patterns. To avoid MTBF and wear-and-tear surprises, one should avoid simple storage via potentially available SD slots. NAS is the preferred option, using reliable hard drives.

The Kismet server, after connecting to its drones, supports a wide range of alerts right out of the box. The alerts themselves may be externally interpreted in lieu of (and along with) raw data. Some of these are:

- APSPOOF – used to detect conflicting or spoofed access points. The user inputs a list of allowed MAC addresses for a given SSID, thus detecting potential attacks utilizing spoofed AP MAC's
- CHANCHANGE – any unauthorized channel change by a listed AP triggers an alert. An attacker may spoof a legitimate AP by MAC and SSID, but on a different channel. This alert detects such activity.
- CRYPTODROP – detects unauthorized reductions in encryption by surrounding APs
- DEAUTHFLOOD – detects 'deauth' flooding
- DHCPCLIENTID – detects potential DHCP DOS attacks, aiming to exhaust the DHCP pool
- LONGSSID/MSFBCOMSSID – detects oversized SSIDs (32+ bytes), which may be poorly handled by certain drivers, thus exploiting vulnerabilities
- PROBENOJOIN – detects active scanning tools (such as NetStumbler), which send network discovery probes, but fail to connect to any responding network

The Kismet-server itself can be run on practically any older, unused office computer, preferably with Linux. Given the advances in miniaturization, a Kismet-server can be run on a plethora of cheap hardware, from single-board computers (like Raspberry Pi) to, conceivably, even entry–level Android smartphones.

Since the aforementioned Kismet alerts are frequent, with a potential large majority of false positives, we need clever heuristics to achieve the desired result. For this, we can use additional tools to filter out specific threats, according to user-specified rules. One such tool is Snort.

Snort is a free, open source WIDS/WIPS utility. Depending on the configuration it can operate in three modes: sniffer, packet logger, network intrusion detection.

Sniffer mode reads network packets and displays them on the console, packet logger logs specified/unfiltered packets to disk and NID mode actively analyzes network traffic against a set of rules predefined by the user.

This, combined with a set of well-positioned sensors/drones, should provide us a with a reasonable detection grid that a hands-on administrator would have no problems managing.

V. CONCLUSION

Wireless security is yet to become a major factor for small businesses and home users. Wireless intrusion detection systems can provide adequate control over a particular surrounding, yet the costs, while quite manageable for large enterprises, are typically prohibitive for smaller businesses and home users. Using cheap off-the-shelf hardware and open source software, a typical sysadmin can easily implement a rudimentary detection system which, while not perfect, can provide a plethora of crucial information in times of need.

REFERENCES

[1] J. Kee, T. Proffitt, “Remote access point/IDS“, SANS institute, 2011.

[2] J. Murray, R. Wanner, “An inexpensive wireless IDS using Kismet and OpenWRT“, SANS Institute InfoSec Reading Room, 2009.

[3] “Snort Overview”, Snort Users Manual 2.9.7, http://goo.gl/bVfbFi

[4] “Wireless Intrusion Detection Systems”, Symantec, 2010/11/02, http://goo.gl/ILdisz

[5] “WIDS: WLAN security and protectiong guide“, TechTarget, 2015/12, http://goo.gl/sDHMcc

[6] M. Kershaw, “Wireless Intrusion Detection“, SharkFest 2013 presentation, https://goo.gl/dihHtt

[7] Kismet documentation, Kismet Wireless, http://goo.gl/l5KUMR

[8] Wikipedia: Kismet (software) – as of 2016/02/20, https://goo.gl/xCGrgm

Use of HLA During Customer Flow Simulation in a Polyclinic

Josef Brozek*, Jan Fikejz**Vojtech Samotan** and Lumir Gago***
*Department of information technologies, University of Pardubice, Pardubice, Czech Republic
**Department of software technologies, University of Pardubice, Pardubice, Czech Republic
***Department of process control, University of Pardubice, Pardubice, Czech Republic
mail@jobro.cz

Abstract - This article covers application of a distributed simulation for creation of a simulation model of a polyclinic. The goal was to monitor the individual patient flows (further referred to as "customer flows") during day-to-day operation, during exceptional situations and above all, to study the effects of exceptional situations on the system. This project is specific in terms of framing several bachelors' theses, which overlap into practice. The article briefly covers the chosen technology, motivation behind the selection of the technology, the principle of data collection and processing, methods of validation and verification, and the principle of creation of alternative scenarios and their evaluation. A discussion on the topic of suitability and advantages of using HLA instead of a common monolithic simulator or simulation tool is also a part of the article.

Keywords: HLA, customers flow, distributed simulation, simulation, JAVA, federate, federation, IEEE 1516:2010.

I. INTRODUCTION

The topic of the work is the creation of a simulator where the modeled domain is a simple logistic system. The particular application domain is a simulation of customer flows in the polyclinic in DvůrKrálovénad Labem. Within the building of this medical facility customer flows heading for certain surgeries were monitored. In addition to the study of the logistic system and its behavior within the simulator alternative scenarios, there were also partial, which were done by students as a key element for deepening their knowledge.

The article presents a case study of its own. Important milestones for its implementation are described, and in the conclusion, results of the case study are presented to the reader.

In the process of creating the simulation, Java programming language was used in which the federatesare realized within a distributed simulation according to a standard HLA (IEEE1516:2010). A RTI, which was used, is a solution from the Swedish company Pitch Technologies (therefore pRTI). Part of the study was realized in their trial version (i.e., via two federates) and the rest by a licensed version, which supports the run up to ten federates. Statistical data processing was carried out in MS Office.

II. STATE OF THE ART

Issues in dealing with similar case studies are relatively well known. Although this work is specific, it is desirable to identify similar solutions. Solutions that deal with the same issues, or address same problem, are categorized according to the important methods used, or according to the extent, so that the relevance of these resources would be more visible.

Solutions mentioned in the following chapters are categorized, so it would be possible to differentiate their focus and breadth. Mentioned solutions entirely (or its component parts) serve as an inspirational model for our own solution.

A. Complex case study

The complex case studies are often extensively presented by scientific teams and their financial demands are astronomical. The case study, which is represented in this case, had minimum cost (time manager and several employees and PhD students and a few hourly consultations), which amounted approximately 500 €. Most of the work was done by students to obtain skills and experience, the realization of their final work.

Particular examples of complex studies may be: [1], [2], [3], [4].

B. Simulators, study without input or output data analysis

While it may seem that publishing partial simulation or case studies, which are not supported by proper numbers - and so not possible to verification - does not make sense, the opposite is true. Similar studies can be found and they can be used for testing new methodology or simulation cores. The simulation solutions are often of small range, so the validation would be intuitive.Similar solutions are above all: [5], [6] and [7]. The Basic principle should be also found in [8],[9],[10],[11].

C. Use HLA to the simulators

High Level Architecture or HLA is a software architecture for building distributed simulations, therefore a combination of partial simulation into one large simulation. Individual partial simulations are combined into federates which mutually communicate through a Run-Time Infrastructure (RTI). RTI provides individual federate services and is the only possible way to communicate between federates, which is carried out in XML language. Additionally, a RTI takes care of synchronization in time. For communication between federates and RTI, it is necessary to create Ambassador, which allows you to call functions RTI and also provides a callback. Each federate contains two Ambassadors: RTI

Ambassador and Federate Ambassador. For more details see [12],[13],[14],[15], [16],[17],[18].

HLA is determined by several prescriptions and the usage of this architecture in the professional community is relatively expanded, even though originally it is a military standard.

III. CUSTOM SOLUTION

The goal of the work was to create a working model of polyclinics and to create alternative scenarios with changes in the baseline scenario. By changes of a baseline scenario, a broken elevator for example, means that everyone would have to take the stairs. Another alternative was to extend some parts of the building or move some offices. The actual realization was difficult, because the study itself was allocated with minimal resources. Also, the urgency to obtain input data, on which the model could be established, was time challenging and difficult to organize. Usable input data was available only from individual doctors, but they all considered data about traffic of patients confidential and therefore it was not possible to use this data.

The polyclinic is shown in figure number 1. From the figure it is clear that the polyclinic is a four-story building. Each floor has two wings. The connection is implemented through one staircase and one elevator. The picture also displays the positions of individual surgeries. On the top floor there is also a private apartment for the buildings owner.

The following chapters describe the method of collection, evaluation and verification of data and also creation of a building model. A special chapter will be reserved for transmission of the results to the customer. For more information see [3],[19].

A. Data collection

The data collection was realized by students within the polyclinic. Measurement was taken by five students and data was collected for one week. A Longer period was not needed, since it was found that every day of the week was similar to the next. Doctors and their offices are busy every day of the week. Differences can be observed only in the individual times of the day because of the varying set of surgery hours for individual doctors.

Before the data collection was realized, there was a complex preparation, during which it was specified that the data will be measured. For all measured data were prepared tables were prepared. Also there was clearly specified methodology for measuring future events.

Generally, the data were gathered by elementarization of clients (patients) movement. Thus, if the patient went into surgery in the left wing of the second floor, each time interval was measured for him particularly. In this case, the time interval from entering the building to the arrival on the stairs. The following time interval was the time spent by walking up the stairs to the first floor. The next time interval was time spent by walking up the stairs from the first to second floor. The following time intervals describe the transfer of the patient on the second floor and the time he spent in the waiting room of a particular doctor.

B. Data evaluation and data validation

All data was collected continuously, evaluated, and processed into a format that would be suitable for future applications in the Java programming language. In picture number two, we see an example of data processing in a histogram.

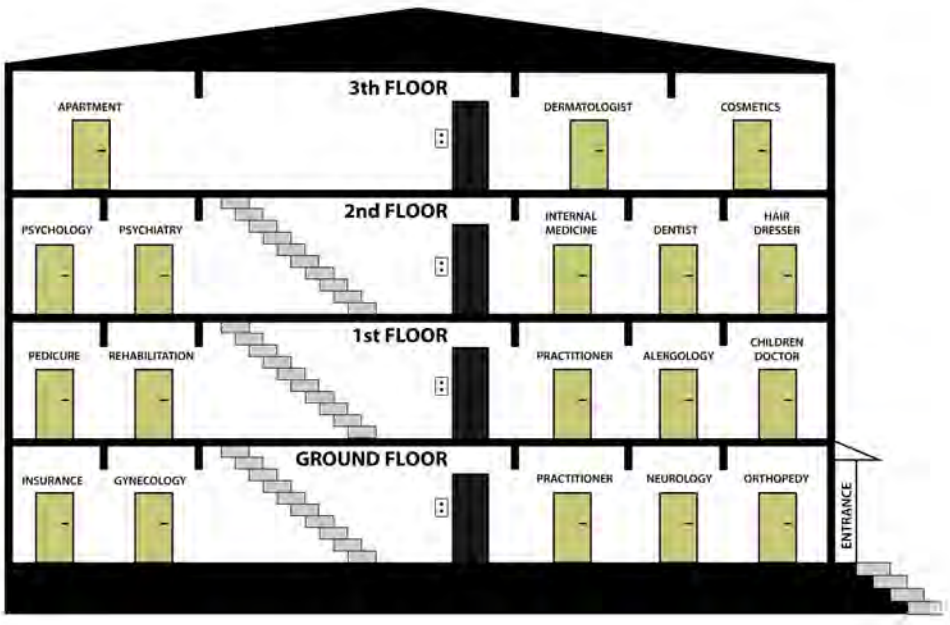


Figure 1. Polyclinic



Figure 2. Time required to take stairs to 1. floor

Statistical analysis was complicated, and it had to be very specific,because of the extreme measured values. In most cases it is possible to ignore the extreme values, or incorporate them into the total formula, because it is very often a measurement error, or possibly an acceptable level of abstraction. However, in this particular case, it is not possible to ignore these facts. Extreme cases are often the result of monitoring the movement of partially immobile patients or seniors. These extreme time values are the key for the correctness of the research.

Processing of the collected data was carried out in several steps, depending on the type of data. After a first analysis and sorting of the collected data (e.g. in the form of histograms) was required statistical analysis for a future generator that operates on the principle of random data in the frame of empirical probability. The creation of a suitable empirical probability improves our own simulation model.

The processed data was then analyzed immediately. Specifically, there was a comparison of measured data processed into histograms, and the results of the random number generator designed in the program, with the actually measured data values. The result of every test can be seen from figure number three. The graph proves that there are only slight variations in results.

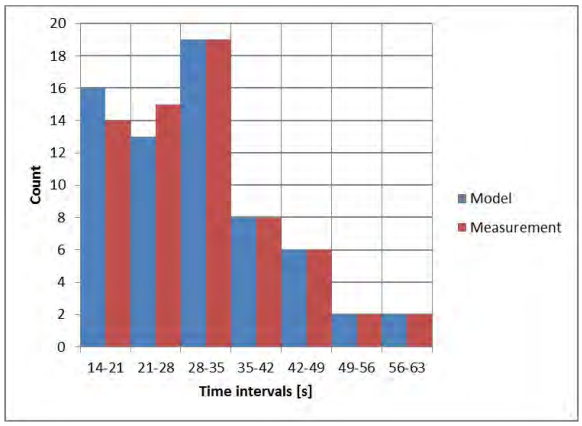


Figure 3. Analysis of input streams

C. Creating a simulation model

A program used to create a simulation model was designed in the programming language Java. The used development environment was NetBeans. Each story of the building is represented by a federate of distributed simulation in HLA. Communication between federates is provided by Pitch RTI. The biggest pitfall of the practical work was to develop and properly validate the generator which provides random numbers based on statistically processed empirical data. The seed a of random numbers generator was set up according to research in the polyclinic and mathematical distributions.

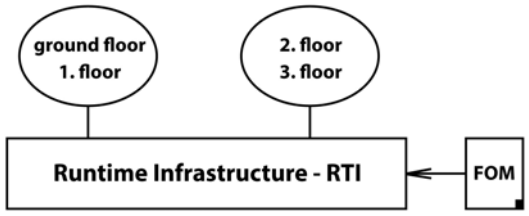


Figure 4. The structure of the first distributed simulator

The simulator was built twice. The first simulator contained only two federates and served primarily to validate the technology. The advantage of this solution was, outside the relative simplicity, that it also allowed only the trial version from Pitch company to be used, and so the students could code out of specialized laboratories. The structure of the first distributed simulator is shown in figure number 4. First federate ensured the ground floor and first floor, plus there were calendars (resp. Pseudo-random number generators), which took care of the customer input streams. The second federate took care of the upper half of the building.

Conceptually, this solution of simulation would not be appropriate. The trial version has thus been extended to the full version of pRTI. The second solution had four federates, thereby achieving extremely simple procedures for validating solutions. Four students studied a simulation run in such a way that each of the students followed the progress of the simulation at just one simulation federate. The second solution is shown in figure number 5.

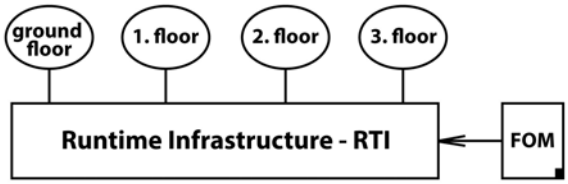


Figure 5. The structure of the second distributed simulator

D. Model verification

The whole model has been properly verified and validated. Verification was performed with the help of external experts in the field. Validation was then carried out at a several levels. Besides the already mentioned validation of input simulation experiment was done with known (measured) input data. The outputs from the simulator and outputs of the real system are similar. The values of the model and measured values can be viewed in figure number 3 and in tables 1 and 2. Figure number 3 reflects primarily the settings of the input streams in time. Comparison of the values of the model and the actual measurement is very positive (the unit variance is only between two histogram groups).

In the table number one we can see the individual percentage of customer flows heading into individual floors of polyclinics by using stairs. It is clear that the generator is well designed and really reflects the reality (for a fixed level of significance and after stabilization, thus after 35 replications).

Table 1: Probability of a usage of stairs

Staircase	Model[%]	Measurement [%]
Ground floor	35,75	37,78
1st story	30,13	29,30
2nd story	16,63	15,48
3rd story	17,50	17,44

Another variable investigated in the context of verification is the probability of usage of an elevator. In table number two percentages of these probabilities are seen (for a fixed level of significance and after stabilization, thus after 35 replications).

Table 2: Probability of a usage of an elevator

Elevator	Model[%]	Measurement [%]
1st story	3,5	3,82
2nd story	6,6	5,41
3rd story	9,1	8,58

E. Alternative scenarios

The primary goal of work was not to create a simulation model of the existing system, but the creation of alternative scenarios. Alternative scenarios focus mainly on the impact of the issue:

1. Emergency situations (evacuation, repair elevator)
2. Object extension
3. Waiting rooms extension
4. Expansion of some surgeries and moving other surgeries
5. Detection of risks and bottlenecks of the system

Creating alternative scenarios are usually only reflected in changes of model parameters. The most difficult was processing emergency situations. It was necessary to add methods to generate random distribution for implementation of the various alternative scenarios, the likelihood of traffic of individual floors, or surgeries, is changed by using alternative scenarios. The scenarios were defined on the basis of systematic observation in polyclinic and also on the evaluation of patient care.

F. Results presentation

The outcome of this work is the proof that using the software demonstrator is possible to perform a simulation study that meets all requirements for specialized work. Thanks to the work of students, the costs were very low. Students gain valuable experience through work. This approach can have an impact that leads to cooperation between private entities and universities in the Czech Republic; as the most professional work ensures the best qualified personnel.

A specific customer without professional education in the field of simulation and does not need to use simulation models for their own use, and does not expect a standard final report. The amount of statistical data and tables of this customer discourage and confuse them.

To present the results to the customer was the key moment, when the production team decided to build the simulation tool in quality animator that faithfully represented the layout of the building. All scenarios immediately became clear and easily understandable for the customer. The entity that moves in the building, forming a queues and occupying waiting rooms and are being processed (go inside clinics and leave). All these processes were unambiguously animated. The results have become very easily understandable for the layperson who could appreciate all the advantages of a simulation study.

In later conversation with the customer, it was confirmed that he was very surprised by what can be managed and predicted by a simulation study. Especially because of this customer and his positive experience there appeared demands for similar small-case solutions.

IV. CONCLUSION

This actual case study is particularly interesting because it was primarily carried directly by the students, despite its scale and quality, it remained low-budget study (if we neglect the need of background).

The second interesting thing is that the customer of the small project has mostly no interest in the final report in the form of multi-page report, and complex statistical reports.A high quality animation of the situation is preferred. Any person appreciates clear, simple and easily interpreted results in the form of clear charts.

Another great benefit for the students was the usage of HLA, which greatly improved knowledge in the field of modeling and simulation.

The whole solution can be regarded as successful. This case study usefully serves for practical purposes, while allowing some students to develop a great knowledge in comparison to just studying.

ACKNOWLEDGMENT

Specialized article created on the basis of the theses from Lukas Base [6] and Vojtech Samotan [7], with the help of team - the authors of this article. Special thanks to consultants, especially Jaroslav Marek, Ph.D., head of the department of mathematics and physics at the Faculty of Electrical Engineering and Computer Science, University of Pardubice, who contributed significantly in the methodology of processing the input data.

The article was written with the support of specialized team ASOTE: Application of software technologies which is funded through the project under the student grant competition at the University of Pardubice.

V. REFERENCES

[1] S. Nageshwaraniyer, K. M. Kim and Y. J. Son, Optimal blast design using a discrete-event simulation model in a hard-rock mine, vol. 67, Mining Engineering, 2015, pp. 47-53.

[2] E. Bottani, G. Ferretti, R. Montanari and M. Rinaldi, Analysis and optimisation of inventory management policies for perishable food products: A simulation study,vol. 9, International Journal of Simulation and Process Modelling, 2014, pp. 16-32.

[3] S. M. Saad and J. C. Aririguzo, Simulating the integration of original equipment manufacturers and suppliers in fractal environment, vol. 7, International Journal of Simulation and Process Modelling, 2012, pp. 148-158.

[4] V. Boschianc, M. P. Fanti, G. Iacobellis, G. Georgoulas, C. Stylios, and W. Ukovich, A model based decision support system for logistics management, 25th European Modeling and Simulation Symposium, EMSS 2013, 2013, pp. 364-369.

[5] E. Ucan and S. Nas, Analysing Istanbul Strait Maritime Pilot Capacity by Simulation Technique, Journal of Navigation, 2015, pp. 1-13.

[6] A. D. Schenkman and M. Xue, Bow-echo mesovortices: A review, vol. 170, Atmospheric Research, 2016, pp. 1-13.

[7] P. H. Shaikh, N. B. M. Nor, P. Nallagownden, I. Elamvazuthi and T. Ibrahim, Intelligent multi-objective control and management for smart energy efficient buildingsm, vol. 74, International Journal of Electrical Power and Energy Systems, 2016, pp. 403-409.

[8] J. Sklenar, Teaching variance reduction in MATLAB, 26th European Modeling and Simulation Symposium, EMSS 2014, 2014, pp. 11-16.

[9] L. B. Rainey and A. Tolk, Modeling and simulation support for system of systems engineering applications, Hoboken, New Jersey: John Wiley & Sons, Inc., 2015.

[10] B. P. Zeigler, H. Praehofer and T. G. Kim, Theory of modeling and simulation: integrating discrete event and continuous complex dynamic systems, 2nd ed. San Diego: Academic Press, 2000.

[11] B. K. Choi and D. Kang, Modeling and simulation of discrete-event systems, Hoboken, N.J.: John Wiley & Sons Inc., 2013.

[12] J. Brozek, A. Kavicka and Ongo, High level architecture virtual assisant framework, Pardubice, 2014

[13] P. Slavicek, Distribuovanesimulacniprostredi, Dizertacniprace. Vysokeucenitechnicke v Brne, 2006.

[14] S. Kozesina, Distrbuovanasimulacepodlestandartu HLA pro pouziti v smulacniknihovne J-SIM, Dimplomovaprace. FakultaaplikovanychvedUniverzita Plzen, 2005.

[15] THE HLA TUTORIAL, (2016). [online] Available at: http://www.pitch.se/images/files/tutorial/TheHLAtutorial.pdf [Accessed 27 Apr. 2016].

[16] D. Chen, L. Wang and J. Chen, Large-scale simulation: models, algorithms, and applications, Boca Raton, FL: CRC Press, 2012.

[17] F. Kuhl, J. Dahmann and R. Weatherly, Creating computer simulation systems: an introduction to the high level architecture, Upper Saddle River, NJ: Prentice Hall PTR, 2000.

[18] R. M. Fujimoto, Parallel and distribution simulation systems, New York: Wiley, 2000.

[19] A. Law and W. Kelton, Simulation modeling and analysis, 3rd ed. Boston: McGraw-Hill, 2000.

Revealing the structure of domain specific tweets via complex networks analysis

Edvin Močibob, Sanda Martinčić-Ipšić, Ana Meštrović
Department of Informatics,
University of Rijeka,
Radmile Matejčić 2, 51000 Rijeka, Croatia
Email: {emocibob, smarti, amestrovic}@uniri.hr

***Abstract*—In this paper we explore the relation between different groups of tweets using complex network analysis and link prediction. The tweets were collected via the Twitter API depending on their textual content. That is, we searched for the tweets in English language containing specific predefined keywords from different domains. From the gathered tweets a complex network of words was formed as a weighted network. Nodes represent words and a link between two nodes exists if these two words co-occur in the same tweet, while weight denotes the co-occurrence frequency. The Twitter search was repeated for four different search criteria (API queries based on different tweet keywords), thus resulting in four networks with different nodes and links. The resulting networks were subjects to further network analysis, as comparison of numerical properties for different networks and link prediction for individual networks. This paper shows the tweet scraping process, our approach to building the networks, the measures we calculated for them, the differences and similarities between different networks we built and our success in predicting future links.**

I. Introduction

Twitter is a popular online social network created in 2006 that enables user to send publicly visible messages called "tweets". One of the main characteristics that distinguishes Twitter from other online social networks is the limit on tweet length. Twitter user are allowed to send tweets that have a maximum of 140 characters. Hence, Twitter is often categorized as a micro-blogging platform. It is estimated that in 2015 Twitter had over half a billion users. [1]

Because of its popularity, user-base size and vast amounts of tweets, Twitter has been studied in the context of person-to-person relations [2], user influence [3], economic predictions [4], predictions of political elections [5], conversational practices [6] and trends discovery [7].

Another important research domain related to Twitter is sentiment analysis. In [9] Pak et al. automatically collect from Twitter a corpus and perform linguistic analysis on it. Then they build a sentiment classifier able to determine positive, negative and neutral sentiments for a document. There has been reported research in automatic classification of tweets regarding their sentiment [8]. [10] gives a detailed revision of the field of sentiment analysis with Twitter in focus. Research by Agarwal et al. [11] examines sentiment analysis on Twitter data. In it the authors introduce POS-specific prior polarity features and explore the use of a tree kernel to eliminate the need for laborious feature engineering. In [12] Kouloumpis et al. investigate the utility of linguistic features for detecting tweets sentiment using a supervised approach, while also leveraging existing hashtags in building training data. Wang et al. [13] present hashtag-level sentiment classification which aims to automatically generate the overall sentiment polarity for a given hashtag in a certain time period.

The following papers use the complex network analysis approach to Twitter data. Villazon et al. in [14] look at Twitter as a complex network, calculating the cluster coefficient, power law and average path length for it. [15] presents a model for describing the growth of scale-free networks. The model is applied only after checking that Twitter is indeed a scale-free network, and for that purpose the mentioned paper proposes a new heuristic method of finding the upper bounds of the path lengths instead of computing the exact length.

In our approach we use complex networks analysis to reveal the structure of domain specific tweets. The motivation of our research is to detect weather networks constructed from different tweets domains have different structural properties. More precisely, the goal of this research is to determine whether (and which) complex network measures can distinguish between networks of tweets with "positive" and "negative" aspects. Possible applications of proposed approach can be in the domain of sentiment analysis. Furthermore, link prediction enables anticipation of positive or negative attitude propagation on Twitter.

We collect positive tweets in English language using keywords with positive polarity (e.g. joy, happiness, ...) and negative tweets using keywords with negative polarity (e.g. anger, fear, ...). Then we perform the global and local complex network analysis where we compare results for four obtained networks. On the global level we use a standard set of network measures (e.g. diameter, average path length, clustering coefficient). However, for the local level analysis we apply a node selectivity measure encouraged by our previous findings [16]–[18] for which we show that it is an important measure for language networks analysis and differentiation.

In the second Section we present the network measures used in our research. In the third Section we describe how we construct the tweet networks. The results and discussion are given in the fourth Section. Finally, the fifth Section contains conclusions and directions for the further research.

II. NETWORKS MEASURES

Complex network is a graph with non-trivial topological features (e.g. high clustering coefficient, low distances, heavy-tailed degree distribution, etc.). It can be represented with a graph G, defined as a pair of two sets $G = (V, E)$; the first set V consisting of vertices and the second set E consisting of edges. N as the number of vertices in V and K as the number of edges in E. In the domain of network analysis, the vertices are referred as nodes and the edges are called links.

Network analysis can be classified by the following three levels: macro-scale or global level, meso-scale level and micro-scale or local level. In weighted complex networks every link connecting two nodes u and v has an associated weight w_{uv}. A node degree is the number of links directly connected (or incident) to that node. The set of nodes incident to a node v is denoted as $\Gamma(v)$. The number of network components is represented by ω. Next, we present network measures that will be used in the following sections.

The average network degree is the ratio of the number of links to the number of nodes. For undirected networks we multiply this ratio by 2 since undirected links always have two incident nodes:

$$\langle k \rangle = 2\frac{K}{N}. \tag{1}$$

Network strength is simply the sum of all link weights in a network:

$$S = \sum_{u,v \in V} w_{uv}. \tag{2}$$

For the average network strength we divide a networks strength with its number of nodes:

$$\langle s \rangle = \frac{S}{N}. \tag{3}$$

Node selectivity for a node v corresponds to the sum of weights of all incident links divided by that nodes degree (denoted as $\deg(v)$):

$$e(v) = \frac{\sum_{u \in \Gamma(v)} w_{uv}}{\deg(v)}. \tag{4}$$

Average network selectivity is the sum of all individual node selectivities divided by the number of nodes:

$$\langle e \rangle = \frac{\sum_{v \in V} e(v)}{N}. \tag{5}$$

Network density is represented as the ratio between the number of existing links and the number of all possible links:

$$d = \frac{K}{N(N-1)}. \tag{6}$$

Average path length for a network, where d_{uv} denotes the number of links lying on the shortest path between $u, v \in V$, is computed as following:

$$L = \sum_{u,v} \frac{d_{uv}}{N(N-1)}. \tag{7}$$

The network diameter represents the longest shortest path in a network ($u, v \in V$):

$$D = \max(d_{uv}). \tag{8}$$

The network radius denotes the shortest $\epsilon(v)$, where $\epsilon(v)$ is defined as the maximum distance between $v \in V$ and any other node:

$$R = \min(\epsilon(v)). \tag{9}$$

Network transitivity where possible triangles are identified by the number of triads (two links with a shared node):

$$T = 3\frac{\#triangles}{\#triads}. \tag{10}$$

Average clustering coefficient, where $c(v)$ is the clustering coefficient for a node v, sums all the individual clustering coefficients and divides them by the number of nodes:

$$C = \frac{1}{N}\sum_{v \in V} c(v). \tag{11}$$

The global network efficiency is the reciprocal value of a networks average path length:

$$E = \frac{1}{L}. \tag{12}$$

In the context of link prediction we use the following measures.

Weighted Common Neighbors, adapted from [19], where weights of links connecting u and v to their common neighbors are summed:

$$CN(u, v) = \sum_{z \in \Gamma(u) \cap z \in \Gamma(v)} w_{uz} + w_{vz}. \tag{13}$$

Weighted Jaccard's Coefficient, adapted from [20], which divides the weighted Common Neighbors value for u and v by the summed weights of all links incident to u and/or v:

$$JC(u, v) = \frac{\sum_{z \in \Gamma(u) \cap z \in \Gamma(v)} w_{uz} + w_{vz}}{\sum_{a \in \Gamma(u)} w_{au} + \sum_{b \in \Gamma(v)} w_{bv}}. \tag{14}$$

Lastly, we present the link prediction precision as the ratio between the number of correctly predicted links and the total number of predicted links. That is, we divide the number of true positives ($|TP|$) by the number of true and false positives ($|TP| + |FP|$). [21]

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \tag{15}$$

III. Networks construction

The first step in constructing networks is the collection of data. Initially, we searched for four sets of tweets according to the following criteria: a) tweets associated to recent immigrant and war related events; b) tweets containing negatively polarized words; c) tweets associated to house pets and d) tweets containing positively polarized words. The subset of positive and negative polarized words is extracted from the sentiment lexicon in [22]. From now on we will refer to the networks built from their respective sets as: a) emo-net^a, b) emo-net^b, c) emo-net^c and d) emo-net^d.

For the data collection process we use Python in combination with the Python Twitter Tools package, which provides an easy-to-use interface for the official Twitter API. In the API request arguments we specified we are searching for a mix of recent and popular tweets in the English language. We scraped about 10000 tweets for each of four different queries, resulting in a dataset of 39882 tweets. It is worth to mention that the official Twitter API documentation states that the language detection is based on the "best-effort" principle [23].

In the text (tweets) preparation step first we eliminate stopwords[1], and from the remaining text we compute the 100 most frequent words for each of the four subsets. We selected top 100 words as the reasonable list which provides the best trade-off between computation time and link prediction results. Note that the former computation was case-insensitive and we used the list of English stopwords presented at http://www.ranks.nl/stopwords.

From the words of preprocessed tweets extended with the set of explicit keywords (e.g. joy, puppy) used for retrieving each of the tweets we form the nodes of the networks. Link between two nodes (words) is established if these two word appear together in the same tweet. Weight on the link represents words co-occurrence frequencies, that is, the number of tweets in which two high-frequency words from the top 100 list co-occurred. That makes the generated networks weighted and undirected. Hence, based on the high-frequency words, we generate four different networks for each of the four data sets.

We build 16 distinct networks from four datasets: the first network is built from 25% of the data, the second from 50%, the third from 75% and the fourth from 100% of the data in one dataset. We will denote those networks, respectively, as emo-net$_1^x$, emo-net$_2^x$, emo-net$_3^x$ and emo-net$_4^x$, where $x \in \{a, b, c, d\}$. That means we, as previously mentioned, generate a total of 16 different networks, four per each dataset.

Some other used Python packages not previously mentioned are NetworkX [24] and LaNCoA [25]. The first one is a popular Python tool for creating and manipulating complex networks. It also provides a rich collection of functions for studying complex networks on various levels. The LaNCoA toolkit provides procedures for construction and analysis of complex language networks.

[1]Stopwords are a list of the most common, short function words which do not carry strong semantic properties, but are needed for the syntax of a language (pronouns, prepositions, conjunctions, abbreviations, ...).

IV. Results

1) Global and local network measures: Here we present the computed global and local network measures for emo-net$_4^a$, emo-net$_4^b$, emo-net$_4^c$ and emo-net$_4^d$. Table I shows the calculated measures that were previously described in Section II.

TABLE I
Global and local network measures

Measure	emo-net$_4^a$	emo-net$_4^b$	emo-net$_4^c$	emo-net$_4^d$
N	101	101	103	104
K	3454	3958	2854	3848
$\langle k \rangle$	68.396	78.3762	55.4175	74
$\langle s \rangle$	1025.9406	830.505	747.0291	1310.25
$\langle e \rangle$	29.4867	24.0104	42.9054	44.7693
d	0.684	0.7838	0.5433	0.7184
ω	1	1	1	1
L	1.316	1.2162	1.4582	1.2816
D	2	2	3	2
R	1	1	2	1
T	0.7965	0.875	0.7774	0.8595
C	0.0088	0.0208	0.0532	0.0077
A	-0.1257	-0.0933	-0.0587	0.0442
E	0.7599	0.8222	0.6858	0.7803

The first visualization we present (Figure 1) is for the node degrees across all emo-net$_4$ networks. We see no major differences for node degrees across those networks.

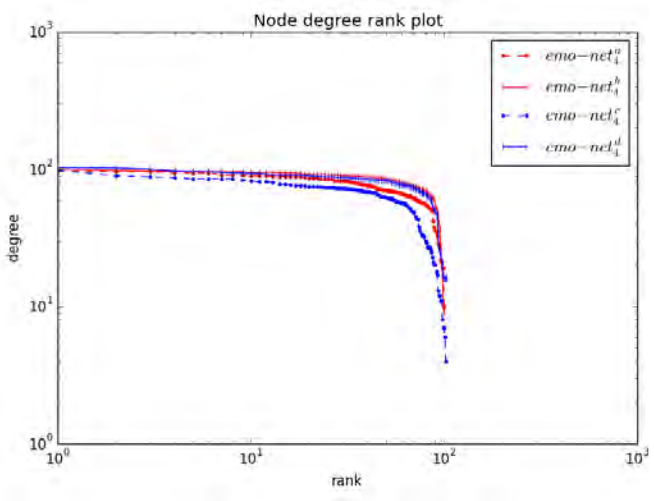


Fig. 1. Node degrees for all emo-net$_4$ networks on a log-log scale

Lets recall that emo-net$_4^a$ and emo-net$_4^b$ were based on data from queries with negative connotations. In contrast, emo-net$_4^c$ and emo-net$_4^d$ were based on queries with positive connotations. The most obvious difference between the first two "positive" and the last two "negative" networks in Table I is $\langle e \rangle$, which represent the value of average network selectivity. $\langle e \rangle$ is notably lower for emo-net$_4^a$ and emo-net$_4^b$ than for emo-net$_4^c$ and emo-net$_4^d$. Average network selectivity can be interpreted as how "heavy" the links across a network are. We

see how our positive networks have on average stronger ties between nodes.

In Figure 2 we visualize the node selectivities for the networks mentioned above. Note that the plot in Figure 2 uses a log-log scale.

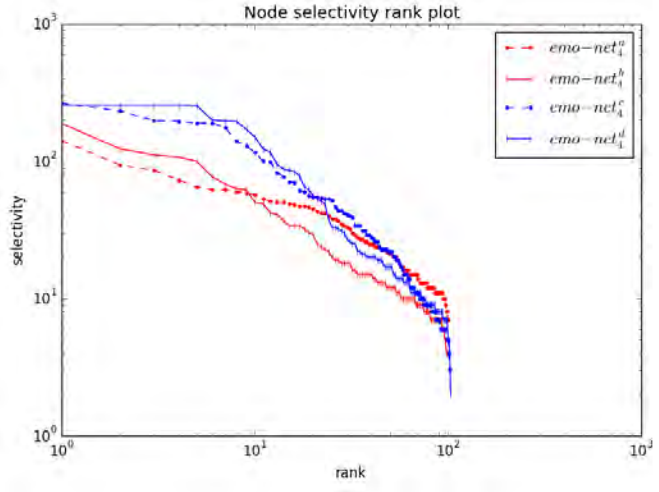


Fig. 2. Node selectivities for all emo-net$_4$ networks on a log-log scale

2) Link prediction: Next we present the results for the link predictions. Here we computed the most likely future links for emo-net$_1^x$, emo-net$_2^x$ and emo-net$_3^x$ where $x \in \{a, b, c, d\}$. The prediction were made using two measures: weighted Common Neighbors (Table II) and weighted Jaccard's Coefficient (Table III). The definitions of both measures can be found in Section II.

We will briefly describe the link prediction process which is the same for both measures. First compute the ranks for all non-existing links in emo-net$_i^x$, $x \in \{a, b, c, d\}$, $i \in \{1, 2, 3\}$. Generate the first set that contains the top n ranked non-existing links in emo-net$_i^x$ (n is the number of new links in emo-net$_{i+1}^x$). Next, generate the second set that holds links which appear in emo-net$_{i+1}^x$ but not in emo-net$_i^x$. Calculate the prediction precision by looking at the intersection of the first and second set.

TABLE II
PREDICTION PRECISION BASED ON THE WEIGHTED COMMON NEIGHBORS MEASURE

Network	emo-net^a	emo-net^b	emo-net^c	emo-net^d
(25%) emo-net$_1$	29.96%	45.92%	30.13%	26.84%
(50%) emo-net$_2$	24.57%	34.88%	21.43%	17.73%
(75%) emo-net$_3$	28.49%	27.49%	18.4%	13.45%

We see from Tables II and III that all predictions had a precision rate above 10%, with some going as high as 50%. The predictions are, by a large margin, most precise for emo-net$_1$ networks. Generally, those networks will not have all of the probable links already in them. With more data all the probable links are added. In most cases the prediction precision for networks with more links tends to fall. That is the only obvious trend for precision rates across query domains, network sizes and prediction measures.

TABLE III
PREDICTION PRECISION BASED ON THE WEIGHTED JACCARD COEFFICIENT MEASURE

Network	emo-net^a	emo-net^b	emo-net^c	emo-net^d
(25%) emo-net$_1$	35.88%	47.96%	37.95%	50.26%
(50%) emo-net$_2$	29.69%	37.72%	28.97%	41.14%
(75%) emo-net$_3$	12.85%	32.16%	30.06%	32.75%

V. CONCLUSION

In this paper we present how we construct multiple complex networks based on four different data sets. Each data set featured a collection of tweets gathered by predefined Twitter API queries. Two of those queries retrieved "negative" oriented tweets, while the other two gathered "positive" oriented tweets. We investigate global and local network measures across four query categories and compare them between "negative" and "positive" networks. In this paper we also predict future links for networks across all query domains. For that purpose we use networks built form a lower percentage of data and compare them with networks built from a higher percentage of the same data.

Regarding network measures, we found that the average network selectivity is the only measure that discriminates between "negative" and "positive" networks, favoring the positive ones. This preliminary results indicate that selectivity based network measures could be used in the Twitter sentiment analysis tasks.

The link prediction process gave no obvious patterns, except the higher prediction precision for networks built from the smallest amount of data. Also, for all of our networks the link prediction precision was above 10%. It should be noted that all our results are preliminary and a more complex analysis would be in order. Such analysis should primarily consider larger and more diverse data sets. Expanding the list of computed network measures would be also worth considering, along with community detection algorithms.

REFERENCES

[1] "Twitter - Wikipedia, the free encyclopedia," 20-Feb-2016. [Online]. Available: https://en.wikipedia.org/wiki/Twitter. [Accessed: 21-Feb-2016].

[2] B. A. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," arXiv:0812.1045 [physics], Dec. 2008.

[3] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in ICWSM, 2010, vol. 10, pp. 1017.

[4] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," Journal of Computational Science, vol. 2, no. 1, pp. 18, Mar. 2011.

[5] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," in Fourth International AAAI Conference on Weblogs and Social Media, 2010.

[6] D. Boyd, S. Golder, and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," in 2010 43rd Hawaii International Conference on System Sciences (HICSS), 2010, pp. 110.

[7] M. Mathioudakis and N. Koudas, "TwitterMonitor: Trend Detection over the Twitter Stream," in Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 2010, pp. 11551158.
[8] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, vol. 1, p. 12, 2009.
[9] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining.," in LREc, 2010, vol. 10, pp. 13201326.
[10] E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, and A. R. Montejo-Ráez, "Sentiment analysis in Twitter," Natural Language Engineering, vol. 20, no. 01, pp. 128, Jan. 2014.
[11] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment Analysis of Twitter Data," in Proceedings of the Workshop on Languages in Social Media, Stroudsburg, PA, USA, 2011, pp. 3038.
[12] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!," in Fifth International AAAI Conference on Weblogs and Social Media, 2011.
[13] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach," in Proceedings of the 20th ACM International Conference on Information and Knowledge Management, New York, NY, USA, 2011, pp. 10311040.
[14] J. Villazon-Terrazas, S. Aparicio, and G. Alvarez, "Study on Twitter as a Complex Network," in The Third International Conference on Digital Enterprise and Information Systems (DEIS2015), 2015, p. 54.
[15] S. Aparicio, J. Villazón-Terrazas, and G. Álvarez, "A Model for Scale-Free Networks: Application to Twitter," Entropy, vol. 17, no. 8, pp. 58485867, Aug. 2015.
[16] S. Beliga, A. Meštrović and S. Martinčić-Ipšić, "Toward selectivity based keyword extraction for Croatian news," In Proceedings of the Workshop on Surfacing the Deep and the Social Web, CEUR, Vol. 1301, Italy, pp. 1-14, 2014.
[17] D. Margan, A. Meštrović, and S. Martinčić-Ipšić, "Complex networks measures for differentiation between normal and shuffled Croatian texts," In Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, Croatia, IEEE, pp. 1819-1823, 2014.
[18] S. Šišović, S. Martinčić-Ipšić, and A. Meštrović, "Comparison of the language networks from literature and blogs," In Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, IEEE, Croatia, pp. 1824-1829, 2014.
[19] L. Lu and T. Zhou, "Role of Weak Ties in Link Prediction of Complex Networks," arXiv:0907.1728 [cs], Jul. 2009.
[20] H. R. de Sa and R. B. C. Prudencio, "Supervised link prediction in weighted networks," in The 2011 International Joint Conference on Neural Networks (IJCNN), 2011, pp. 22812288.
[21] Y. Yang, R. N. Lichtenwalter, and N. V. Chawla, "Evaluating Link Prediction Methods," Knowledge and Information Systems, vol. 45, no. 3, pp. 751782, Dec. 2015.
[22] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.", In: LREC, 2010, pp. 2200-2204.
[23] "GET search/tweets — Twitter Developers." [Online]. Available: https://dev.twitter.com/rest/reference/get/search/tweets. [Accessed: 21-Feb-2016].
[24] D. A. Schult and P. Swart, "Exploring network structure, dynamics, and function using NetworkX," in Proceedings of the 7th Python in Science Conferences (SciPy 2008), 2008, vol. 2008, pp. 1116.
[25] D. Margan and A. Meštrović, "LaNCoA: A Python Toolkit for Language Networks Construction and Analysis," in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on, 2015, pp. 16281633.

Counting Prime Numbers in Paralell - Faster by Reducing the Synchronization Overhead

Duraković Adna*, Pajić Enil*, Branković Ilvana*, Kušundžija Elma*, Karkelja Sead*
*Faculty of Electrical Engineering/Department for Computer Science, University of Sarajevo
Sarajevo, Bosnia and Herzegovina
(adurakovic4, epajic1, ibrankovic1, ekusundzij1, skarkelja1)@etf.unsa.ba

***Abstract*— This paper illustrates comparison of sequential and parallel execution of algorithm that counts prime numbers between 1 and N. Many systems rely on problems in number theory, where primes have an important role. Primes are used for encryption and hash functions too. In order to reduce the complexity of sequential algorithm execution, which is O(n^2), and also to achieve speedup, the algorithm was parallelized and some parts of it were additionally optimized by saving time on barriers. OpenMP tool is used as an application programming interface for parallelization. Sequential and parallel algorithm performances were measured on different architectures, based on Intel i3 and i7 processors, and the results, as well as the achieved speedup, are presented in the paper.**

***IndexTerms*—Primes, speedup, Paralellization.**

I. INTRODUCTION

It's known, since Euclid, that there is an infinite number of prime numbers[1]. There are also a lot of proofs that support this thesis. Mathematicians were always keen to investigate existence of an infinite number of primes. As time passed by, they came up with idea to define the function π(x) which calculates number of primes that are smaller than a certain number (value of x).

$$\boldsymbol{\pi(x) = \#\{p \leq x, where\ p\ is\ prime\ number\}}$$

For example, if we look at all primes smaller than 25, then we have following numbers: 2, 3, 5, 7, 11, 13, 17, 19 and 23. So, $\boldsymbol{\pi(25) = 9}$.

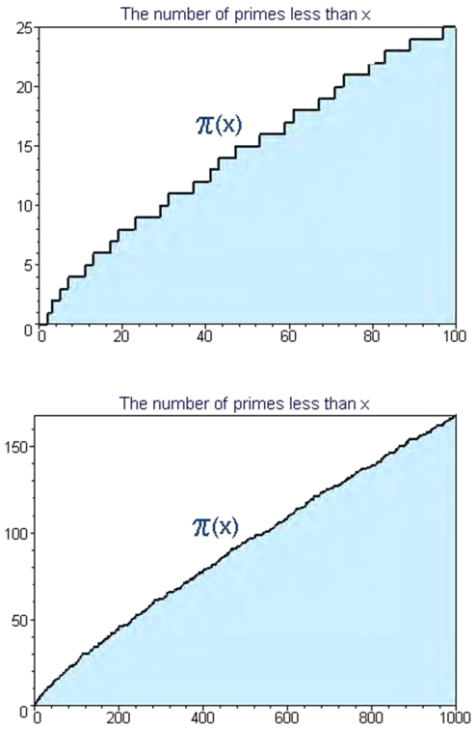


Fig. 1.Graph of function $\boldsymbol{\pi(x)}$ [3]

Before the invention of computers, mathematicians made a table of prime numbers. The most used one was D.N. Lehmer's[2] table of prime numbers (maximum value of x in table was 10,006,721).

[1]A prime number (prime) is a natural number greater than one that has no positive divisors other than one and itself.

[2] Derrick Norman Lehmer (1905.-1991.) was an American mathematician and number theorist.

	x	pi(x)
1	10	4
2	100	25
3	1,000	168
4	10,000	1,229
5	100,000	9,592
6	1,000,000	78,498
7	10,000,000	664,579
8	100,000,000	5,761,455
9	1,000,000,000	50,847,534
10	10,000,000,000	455,052,511
11	100,000,000,000	4,118,054,813
12	1,000,000,000,000	37,607,912,018
13	10,000,000,000,000	346,065,536,839
14	100,000,000,000,000	3,204,941,750,802
15	1,000,000,000,000,000	29,844,570,422,669
16	10,000,000,000,000,000	279,238,341,033,925
17	100,000,000,000,000,000	2,623,557,157,654,233
18	1,000,000,000,000,000,000	24,739,954,287,740,860
19	10,000,000,000,000,000,000	234,057,667,276,344,607
20	100,000,000,000,000,000,000	2,220,819,602,560,918,840
21	1,000,000,000,000,000,000,000	21,127,269,486,018,731,928
22	10,000,000,000,000,000,000,000	201,467,286,689,315,906,290
23	100,000,000,000,000,000,000,000	1,925,320,391,606,803,968,923
24	1,000,000,000,000,000,000,000,000	18,435,599,767,349,200,867,866
25	10,000,000,000,000,000,000,000,000	176,846,309,399,143,769,411,680

Fig. 2.Table of $\boldsymbol{\pi(x)}$ function values and corresponding x values [3]

Until 2007, mathematicians were manually finding a way of computing π(x) function. Xavier Gourdon was the first to write computer program for calculating π(x) function (maximum value of x was $\boldsymbol{4 * 10^{22}}$). The importance of the algorithm was in its use in hash functions and data encryption. Encryption is based on the theory of numbers, in which primes have important role. Some cryptographic algorithms such as RSA are affected by prime factorization. Practically, there is a public key which is in fact product of two large prime numbers and is used for encryption, and there is a secret key, that consists of prime numbers, for message decryption. Public key can be literally public, but only primes that are needed for decryption are known. It's almost impossible to guess those two prime numbers that form the public key. This algorithm is also used in integer and string hash functions. Both functions use prime number, which belongs to the specified interval.

II. ALGORITHM DESCRIPTION

The algorithm calculates number of primes in the interval from 1 to a given value. Time complexity of sequential program is $\boldsymbol{O(n^2)}$. Objective is to achieve speedup by parallelizing the algorithm. Speedup is achieved by parallelizing the parts of the program that can be parallelized, i.e. parts that are independent from each other and can be executed in parallel. OpenMP[3] is used as an application programming interface for parallelization. Optimal speedup achieved by executing a parallel algorithm on n processors is equal to:

$$\boldsymbol{optimal\ speedup = \frac{sequential\ execution\ time}{number\ of\ processors}}$$

In this case, the optimal speedup would be equivalent to:

$$\boldsymbol{optimal\ speedup = \frac{O(n^2)}{n}}$$

It should be noted that the performance of the program execution is affected by hardware architecture on which the program runs. In this experiment are used Intel Core i3-380M and Intel Core i7-3610QM processors. The processor i3-380M has 2 cores, operating at base frequency of 2.53 GHz, while i7-3610QM has 4 cores, operating at base frequency of 2.3 GHz. Both processors support Hyper-Threading technology, so each core of i3-380M and i7-3610QM can run two threads. Each core of i3-380M has 64KB of L1 cache, 256KB of L2 cache and 3MB of shared L3 cache, while each core of i7-3610QM has 64KB of L1 cache, 256KB of L2 cache and 6MB of L3 cache. Here after is listed pseudocode of the algorithm that calculates number of prime numbers.

Pseudocode :

```
n← inputNumber
prime← 1
total← 0
for(i=0 to n )
{
        prime← 1
        for( j=0 to i )
        {
                if(i mod j is 0)
                {
                        prime← 0
                        break
                }
        }
        total = total + prime
}
```

[3]OpenMP (**O**pen **M**ulti-**P**rocessing) is an application programming interface (API) that supports multi-platform shared memory multiprocessing programming in C, C++, and Fortran.

III. ALGORITHM PARALLELIZATION

As we can see from the pseudocode, it's necessary to check, for every number in the range of 1 and that number, if it's prime number, in order to define final number of primes in the range (the complexity of checking whether the number is prime or not is O(n)). Complexity of the algorithm shown in pseudocode above is O($\boldsymbol{n^2}$).).

We saved time by dividing independent jobs in for loop and assigning them to different threads. OpenMP, as a tool, gives various options for parallelization, if the used architecture meets certain requirements. $\#\boldsymbol{pragma\ omp\ parallel}$ directive is used to indicate block that suggests parallelization. After that, variable that represents shared variable within threads is defined with $\boldsymbol{shared(n)}$ directive, and in particular, with $\boldsymbol{private(i,j,prime)}$ directive it's emphasized that variables which iterate through the loop (i and j), as well as flag variable (that tells if number is prime or not), are private variables for every thread. Finally, in order to accomplish our parallelization goal, one more, specific directive $\#\boldsymbol{pragma\ omp\ for\ reduction}(+:\boldsymbol{total})$ is defined. It allows us to really execute parts of for loop on different threads, with an extra requirement that the total variable is special for each of the threads, but its value is added to the final value when all of the threads finish their execution.

Additional time savings could be achieved by parallelization of the inside for loop, however, this for loop manipulates the variable 'prime' which would be, in that case, shared within all the threads. This would mean that, if one thread changed the value of 'prime' variable, all the other threads should stop executing. This would cause loss of time spent on communication between threads more than achieving savings by dividing jobs on threads, so this part stayed sequential.

Algorithm with OpenMP directives as well as the optimized algorithm are given in addendums A and B, respectively.

IV. MAIN RESULTS

Various amounts of input data are used for algorithm testing; it's used from 100,000 to one million natural numbers for the search of prime numbers. Ten tests are performed in the specified range with a step of 100,000 (start with 100,000 and then go to 200,000, 300,000, …,1 million).

As it was expected, the sequential execution was the slowest, while the parallelized algorithm gave much better results. There are two parallel versions of the algorithm, one is more optimized, so instead of one variable, there is an array, with four or eight elements (depending on the number of threads), each storing results for different thread. In that case, time spent on synchronization is decreased or the need for a barrier is removed. There is an additional part that adds partial sum (4/8 array elements) and that part is exclusively sequential.

The following two tables show test results on i3-380M and i7-3610QM. They consist of sequential algorithm execution time, parallel algorithm with shared variable execution time and parallel algorithm with partial sums execution time columns. Additionally, results of parallel algorithm with shared variable execution on i3-380M, but with 2 threads instead of 4, are also included in the table below.

Algorithm (*s*) / Num. of elem.	**Sequential**	**Parallel with shared variable (2 threads)**	**Parallel with shared variable (4 threads)**	**Parallel with partial sum (4 threads)**
100 000	2.329	1.823	1.305	1.233
200 000	8.369	6.652	5.01	4.506
300 000	18.287	14.005	10.994	9.491
400 000	32.542	23.725	18.49	16.383
500 000	49.039	37.239	28.455	27.052
600 000	69.242	54.059	41.287	37.117
700 000	90.302	70.098	52.999	48.293
800 000	119.85	92.559	68.156	62.492
900 000	149.131	116.439	89.792	76.856
1 000 000	185.042	142.912	111.376	95.048

TABLE I. Optimisation results on i3-380M (execution time)

Results of parallel algorithm with shared variable execution on i7-3610QM, but with 2 and 4threads instead of 8, are also included in the table below. To save space, table shows results with a step of 100,000.

Algorithm (*s*) / Num. of elem.	**Sequential**	**Parallel with shared variable (2 threads)**	**Parallel with shared variable (4 threads)**	**Parallel with shared variable (8 threads)**	**Parallel with partial sum (8 threads)**
100 000	1.622	1.138	0.782	0.499	0.468
200 000	5.976	4.431	2.839	1.747	1.576
300 000	13.217	9.61	6.131	3.604	3.37
400 000	22.372	16.554	10.234	6.365	5.865
500 000	34.752	25.709	16.13	9.672	8.892
600 000	49.912	36.505	22.452	13.683	13.335
700 000	67.618	48.786	30.969	18.674	16.578
800 000	86.861	62.218	39.769	24.164	21.468
900 000	109.484	78.247	49.405	29.734	27.156
1 000 000	133.801	95.335	57.335	35.928	33.546

TABLE II. Optimisation results on i7-3610QM(execution time)

Achieved speedup of the fastest parallel code in relation to sequential achieved on i3 is satisfying and is about 1.94 for million elements. While achieved acceleration on i7 is about 3.98 for same number of elements. Regarding to the acceleration of parallel code, which is achieved by removing barriers (reduction of time for synchronization), code is faster for about 16 seconds for one million elements and achieved acceleration is 1.17 on i3, while on i7 code is faster for about 2 seconds and achieved acceleration is 1.07. The reason for this is presence of parts of the sequential algorithm that cannot be parallelized (like inside for loop and sum of partial results). It should be noted that for small inputs (numbers smaller than 100) sequential code is slightly faster than parallel version of code. This happens due to the additional overhead that occurs because of job distribution and synchronization between threads.

Speedup ratio, S, and parallel efficiency, E, may be used to provide an estimate for how well a code sped up if it was parallelized. For example, if f = 0.1 the speedup bound above predicts a 10 fold speedup in the limit. On the other hand, a code that is 50% parallelizable will at best see a factor of 2 speedup.[4]

Efficiency E in this case for i3 processor (N=4 threads) is

$$E = S/N = 1.94/4 = 0.485 \approx 0.5$$

and, for i7 processor (N=8 threads), efficiency is almost the same, as it's shown below:

$$E = S/N = 3.98/8 = 0.497 \approx 0.5$$

Program that scales linearly (S=N) has parallel efficiency 1. A task-parallel program is usually more efficient than a data-parallel program.[4]

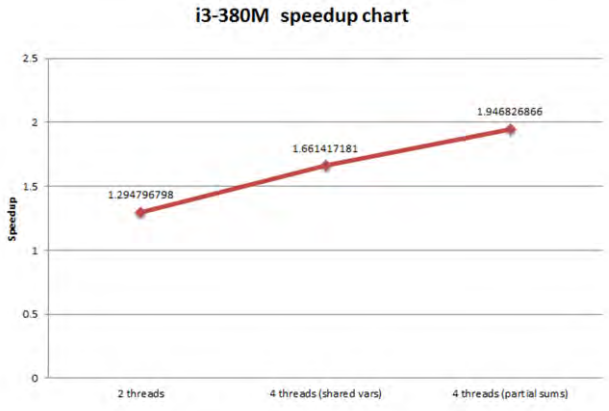


Fig. 3. Intel i3 processor speedup graph (considering n=1000000)

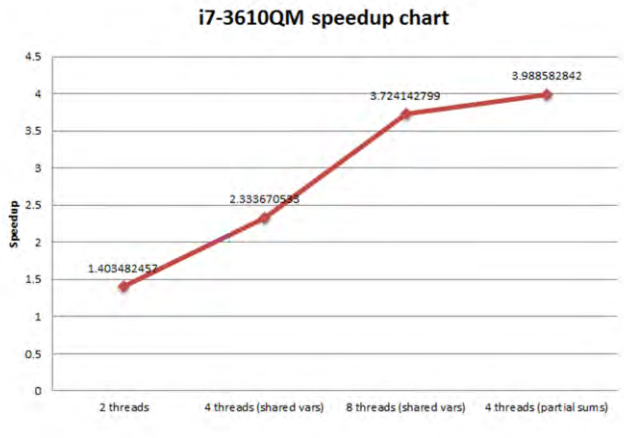


Fig. 4. Intel i7 processor speedup graph (considering n=1000000)

V. CONCLUSION

In this paper the focus is on comparison of sequential and parallel algorithm execution that counts number of primes from one to N (input value). Parallel execution is performed with different number of threads, with or without optimization. Conclusion is that it's possible to achieve speedup by parallelizing the algorithm. As it's shown in graphs above, the bigger the input number, the bigger is difference between sequential and parallel execution time. With big input values (bigger than 800000) and four threads, parallel execution time is almost two times smaller than sequential execution time. Additional time savings are made by eliminating the need for barrier (every thread got its variable) and avoiding the overhead caused by synchronization. This optimization is only visible for a very large input numbers. Great advantage of the chosen algorithm is independent task execution within a loop, which allows us to assign tasks to threads easier and also reduce a need for communication and synchronization between them.

VI. REFERENCES

[1] K. Mann, "The science of encryption: prime numbers and mod n arithmetic", Undated. [Online]. Available: https://math.berkeley.edu/~kpmann/encryption.pdf. [Accessed: 02- Nov-2015]

[2] People.sc.fsu.edu, "PRIME_OPENMP - Count Primes Using OpenMP", 2016. [Online]. Available: https://people.sc.fsu.edu/~jburkardt/c_src/prime_openmp/prime_openmp.html. [Accessed: 10-Nov-2015]

[3] C. Caldwell, "How many primes are there?",Primes.utm.edu, 2016. [Online]. Available: http://primes.utm.edu/howmany.html. [Accessed: 20-Dec-2015]

[4] Bu.edu, "Speedup Ratio and Parallel Efficiency : TechWeb : Boston University", 2016. [Online]. Available: http://www.bu.edu/tech/support/research/training-consulting/online-tutorials/matlab-pct/scalability/. [Accessed: 23- Jan- 2016].

A. ALGORITHM WITH OPENMP DIRECTIVES

```
int prime_number_par ( int n )
{
  int i,  j;
  int prime;
  int total = 0;
  int niz[4] = {0,0,0,0};

# pragma omp parallel \
  shared ( n ) \
  private ( i, j, prime )

# pragma omp for reduction ( + : total )
  for ( i = 2; i <= n; i++ )
  {
    prime = 1;

    for ( j = 2; j < i; j++ )
    {
      if ( i % j == 0 )
      {
        prime = 0;
        break;
      }
    }
    total = total + prime;
  }

  return total;
}
```

B. OPTIMIZED ALGORITHM

```
int prime_number_par ( int n )
{
  int i, j, k;
  int prime;
  int total = 0;
  int thread_num = omp_get_max_threads();
  int niz[thread_num];

  for ( k = 0; k <= thread_num; k++ ) { niz[k] = 0; }

# pragma omp parallel \
  shared ( n ) \
  private ( i, j, prime )

# pragma omp for
      for ( i = 2; i <= n; i++ ) {
        prime = 1;
for ( j = 2; j < i; j++ ) {
     if ( i % j == 0 )  {
       prime = 0;
       break;
      }
   }
    niz[omp_get_thread_num()] += prime;
  }

  for ( k = 0; k <= thread_num; k++ ) {
    total += niz[k];
  }
   return total;
}
```

Parallelization Challenges of BFS Traversal on Dense Graphs Using the CUDA Platform

Huma Milišić, Dina Ahmić, Hamdija Sinanović, Emina Šarić, Amar Asotić, Alvin Huseinović

Faculty of Electrical Engineering
University of Sarajevo
Sarajevo, Bosnia and Herzegovina
{hmilisic1, dahmic1, hsinanovic1, esaric1, aasotic1, ahuseinovic}@etf.unsa.ba

***Abstract*—This paper presents challenges encountered while parallelizing an existing sequential algorithm. A breadth-first search implementation in CUDA C++, of quadratic time complexity, is used. Even though BFS might seem like an easily parallelizable problem due to many independent iterations over graph vertices, there are other important aspects which need to be considered. Properties like granulation, communication and load balancing are thoroughly examined to find their exact impact in BFS. The implementation is profiled to find bottlenecks and problems which prevent its further optimization.**

***Keywords*—BFS (Breadth-First Search), CUDA (Compute Unified Device Architecture), GPU (Graphic Processing Unit), graphs, graphing algorithms, parallelization**

I. Introduction

Graphs are widely-used data structures that describe a set of objects, referred to as nodes, and the connections between them, called edges. They are a fundamental data representation method widely used in numerous fields such as intelligence analysis, robotics, social network analysis, and computational biology. These applications have traditionally required long periods of processing time due to their massive data-set sizes. Parallelism has usually failed to alleviate matters, because parallel speedup of these applications is severely limited by the random nature of their memory access patterns, which is a fundamental property of graph processing algorithms.

Breadth-first search (BFS) is of particular importance among different graph search methods and is widely used in numerous applications. BFS is a visiting strategy for all vertices of a graph. Given a distinguished source vertex s, BFS systematically explores the graph G to discover every vertex that is reachable from s. It is typically considered one of the most important graph algorithms, because it serves as a building block for many other algorithms including betweenness centrality calculation, connected component identification, community structure detection, and max-flow computation. Benchmark suites targeting graph applications include BFS as a primary element. Due to such importance, significant research has been conducted to efficiently implement a parallel BFS for a wide array of computing systems.

In this paper, a parallel implementation of BFS on a GPU using CUDA, handling large graphs up to 110 million edges is presented. The results for the speed-up obtained by the parallel algorithm over its sequential execution are shown.

II. CUDA Platform

Due to the inherently parallel nature of the algorithm and general availability of hardware, the CUDA programming model was chosen for the project. CUDA C++, with its roots in C and C++, is the most commonly used programming language for the platform. Regarding hardware, CUDA is supported by the majority of NVIDIA graphics cards, which have a high share in the graphics cards market. Thus, the chosen tools for the project are CUDA Toolkit 7.5, NVIDIA GeForce GTX 850M and NVIDIA GeForce 840M graphics cards.

III. Graph Representation

The graph is represented using an adjacency matrix A, whose rows are the adjacency lists Ai. The number of edges within the sparse graph is typically only a constant factor larger than n. The compressed sparse row (CSR) sparse matrix format is used to store the graph in memory consisting of two arrays. As illustrated in Figure 1, the column-indices array C is formed from the set of the adjacency lists concatenated into a single array of m integers. The row-offsets R array contains n+1 integers, and entry R[i] is the index in C of the adjacency list Ai [1].

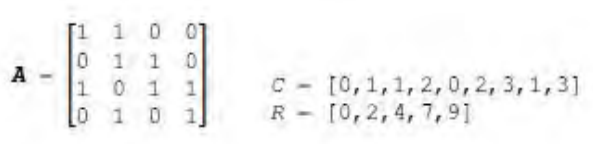


Fig. 1: Example CSR representation column-indices array C and row-offset array R comprise the adjacency matrix A

IV. ALGORITHM IMPLEMENTATION

Graph algorithms are challenging to implement on GPUs. The reason for this is the distinctiveness of every graph. Sparse graphs need to be handled differently than dense ones. The algorithm in this paper is made for dense graphs. The reason for this is that sparse graphs are hard to parallelize, because fewer vertices mean fewer parallelizable operations.

The traditional, sequential version of the breadth-first search uses a graph structure which keeps vertices and edges of the graph. The goal is to, given a starting vertex *s*, discover all edges reachable from *s*. The algorithm works by building a tree of all reachable vertices, with the starting vertex at the root. The tree is built iteratively, using a process called frontier propagation. Each level of the tree represents a frontier. Frontier propagation checks all neighbors of a vertex, and if they aren't visited, adds them to a new frontier. Frontiers can be stored in a queue structure. This structure can be an overhead on the GPU, so in this version of the algorithm it's eliminated. Instead, a vertex is checked to see if it belongs to the same frontier. However, such an algorithm has worse time complexity than the sequential implementation. If the number of vertices is V, and the number of edges E, then the time complexity of the sequential algorithm is $O(V+E)$. This is reduced to $O(V)$ for sparse graphs. On the other side, the parallel implementation, which maintains no frontier queue, has a time complexity of $O(VL+E)$. L is the number of levels in a tree. For sparse graphs, this is $O(V^2+V)=O(V^2)$. This is much slower than the sequential implementation.

Another parallel version of BFS uses matrices. Each frontier propagation is done through matrix-vector multiplication. The number of multiplications is L, which is the same as the number of levels. Total time complexity of this algorithm is $O(V+EL)$. $O(V)$ is the time needed to initialize the vector. In the worst case, the time complexity is $O(V^2)$. This means that this version of the algorithm is still slower that the sequential version.

The algorithm implementation used is called quadratic, because it does quadratic amount of work. The algorithm can be trivially parallelized, because work on each edge/vertex can be done independently. There are two approaches to this algorithm: vertex-centric and edge-centric. The algorithm works by visiting vertices and marking them as visited. First, every vertex and every edge are processed to find unvisited vertices. A vertex needs to be marked as visited if it is connected by an edge to an already visited vertex. As a consequence, all unvisited neighbors of the marked vertex need to be marked as visited as well. The work complexity is $O(V^2+E)$.

A. Sequential BFS Implementation

Graphs of the form G = (V, E) with a set V of n vertices and a set E of m directed edges are considered. The BFS problem is, given an undirected, unweighted graph G(V, E) and a source vertex S, to find the minimum number of edges needed to reach every vertex V in G from source vertex S. The optimal sequential solution for this problem takes $O(V+E)$ time since each vertex is labelled exactly once and each edge is traversed exactly once.

Algorithm 1 SEQ-BFS(G,v_0)

```
1: for each vertex u ∈ (G) − {v_0}
2:    u.dist ← ∞
3: end for
4: v_0.dist ← 0
5: Q ← {v_0}
6: while Q ≠ ∅ do
7:    u ← DEQUEUE(Q)
8:       for each v ∈ V such that (u,v) ∈ E(G)
9:          if v.dist = ∞
10:            v.dist ← u.dist + 1
11:            ENQUEUE(Q,v)
12:         end if
13:      end for
14: end while
```

Algorithm 1 describes a classical sequential algorithm for computing BFS, which uses a FIFO queue as an auxiliary data structure. The FIFO can be implemented as a simple array with two pointers to the head and tail of the items in the queue. Enqueueing an item consists of incrementing the tail pointer and storing the item into the array at the pointer location. Dequeueing consists of removing the item referenced by the head pointer and incrementing the head pointer [1].

B. Parallel BFS Implementation

The first CUDA BFS was published in [2]. In this paper their implementation approach is used.

Algorithm 2 host BFS(Graph $G(V,E)$, Source Vertex S)

```
1: Create vertex array V_a from all vertices and edge
   array E_a from all edges in G(V,E)
2: Create frontier array F_a, visited array X_a and cost
   array C_a of size V
3: Initialize F_a, X_a to false and C_a to ∞
4: F_a[S] ← true, C_a[S] ← 0
5: while F_a ≠ ∅ do
6:    for each vertex V in parallel do
7:       Invoke BFSKernel(V_a, E_a, F_a, X_a, C_a) on the grid
8:    end for
9: end while
```

The function BFS is an implementation of the Algorithm 2. It sets the number of threads per block depending on the number of vertices, and the number of blocks per grid depending on the number of threads per block and the number of vertices. After that, the function calls the kernel (Algorithm 3) to check if the assigned node is already visited. If the node isn't visited, the kernel removes it from the array of unvisited nodes and adds it to the array of visited nodes. The kernel then marks unvisited adjacent nodes, and updates their cost. In this case, cost is considered to be the number of the hops between two nodes. The described procedure is repeated until reduction of the array that contains unvisited nodes returns zero, which indicates that there are no more unvisited nodes left. This

implementation is not very efficient because it uses thrust reduction after each iteration of a loop.

Algorithm 3 kernel BFSKernel(V_a, E_a, F_a, X_a, C_a)

```
1: tid ← getThreadID
2: if F_a[tid] then
3:   F_a[tid] ← false, X_a[tid] ← true
4:   for all neighbors nid od tid do
5:     if NOT X_a[nid] then
6:       C_a[nid] ← C_a[tid] + 1
7:       F_a[nid] ← true
8:     end if
9:   end for
10: end if
```

Function BFS_1 implements Algorithm 4. It uses the boolean variable 'done' accessible from both, CPU and GPU to check if there are more nodes left to visit. It also calls a different kernel (Algorithm 5) which checks the list of costs for all nodes. If the kernel finds at least one node with a cost higher than or equal to the number of the current iteration, it updates the costs of the nodes and sets 'done' to false.

Algorithm 4 host BFS_1(Graph G(V,E), Source vertex S)

```
1: Create vertex array V_a from all vertices and edge
   array E_a from all edges in G(V,E)
2: Create cost array C_a of size V
3: Create managed variables iteration' and done'
4: Initialize C_a to ∞
5: C_a[S] ← 0, done ← true, iteration ← 0
6: while NOT done do
7:   done ← true
8:   for each vertex V in parallel do
9:     Invoke
       BFSKernel_1{, atomics, share, shareAtomics}
       (V_a, E_a, C_a) on the grid
10:  end for
11:  iteration + +
12: end while
```

Algorithm 5 kernel BFSKernel_1 (V_a, E_a, C_a)

```
1: tid ← getThreadID
2: if C_a[tid] = iteration then
3:   if done
4:     done ← false
5:   end if
6:   for all neighboors nid of tid do
7:     if C_a[nid] > iteration
8:       C_a[nid] ← iteration + 1
9:     end if
10:  end for
11: end if
```

Function BFS_1_Atomics (Algorithm 4) uses an integer variable instead of a boolean variable (Algorithm 6), so it can be updated using atomic bitwise operations, which is considered safer and faster than assignation.

Algorithm 6 kernel BFSKernel_1_Atomics(V_a, E_a, C_a)

```
1: tid ← getThreadID
2: if C_a[tid] = iteration then
3:   if done
4:     atomicAnd(done, 0)
5:   end if
6:   for all neighboors nid of tid do
7:     if C_a[nid] > iteration
8:       C_a[nid] ← iteration + 1
9:     end if
10:  end for
11: end if
```

Function BFS_1_Share (Algorithm 4) reduces the execution time by placing the iteration number in the GPU's shared memory (Algorithm 7). Performance is improved because a number of the current iteration is read multiple times and accessing a GPU's shared memory is faster than accessing a GPU's global memory.

Algorithm 7 kernel BFSKernel_1_Share (V_a, E_a, C_a)

```
1: tid ← getThreadID
2: shared localIt ← iteration
3: if C_a[tid] = iteration then
4:   if done
5:     done ← false
6:   end if
7:   for all neightbors nid of tid do
8:     if C_a[nid] > localIt
9:       C_a[nid] ← localIt + 1
10:    end if
11:  end for
12: end if
```

Function BFS_1_ShareAtomics (Algorithm 4) uses features from both BFS_1_Share and BFS_1_Atomics. The number of the current iteration is stored in a shared memory for faster access, and an integer variable updated using atomic bitwise operation is used to check if all nodes are visited (Algorithm 8).

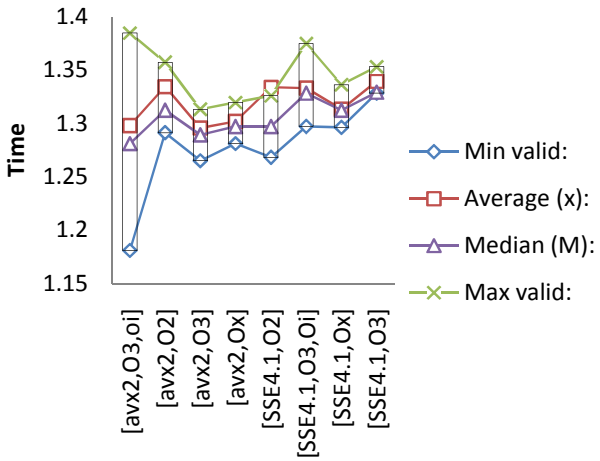


Fig. 2 Sequential code optimizations on GTX 850M

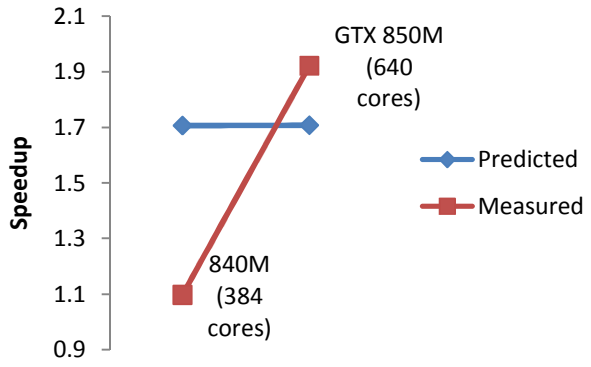


Fig. 3 Speedup Amdahl predicted vs measured

Algorithm 8 kernel BFSKernel_1_ShareAtomics (V_a, E_a, C_a)

```
1: tid ← getThreadID
2: shared localIt ← iteration
3: if C_a[tid] = localIt then
4:   if done then
5:     atomicAnd(done, 0)
6:   end if
7:   for all neighbors nid of tid do
8:     if C_a[nid] > localIt then
9:       C_a[nid] ← localIt + 1
10:    end if
11:  end for
12: end if
```

V. RESULTS

A. Benchmark

A test running program was created, that executes the mentioned implementations multiple times for a given graph and produces a log file with measured times. The log file was processed with MS excel.

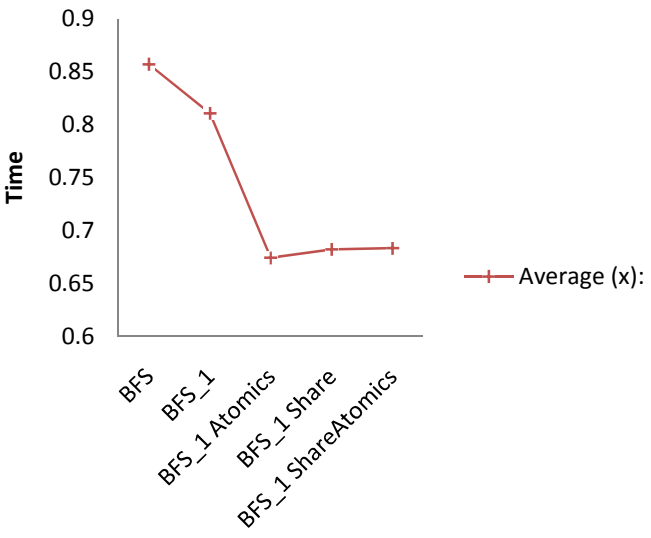


Fig. 4 Parallel optimizations on GTX 850M

TABLE I IMPLEMENTATION EXECUTION COUNT

Algorithm	Execution count
All parallel ones	200 x
Sequential [SSE4.1,O3], [avx2,Ox], [avx2,O3], [avx2,O2]	100 x
Sequential [SSE4.1,O2]	200 x
Sequential [avx2,O3,oi], [SSE4.1,O3,Oi], [SSE4.1,Ox]	300 x

The test benchmark graph was taken from [3], with properties shown in Table II**.**

Intel compiler 16.0 options used:

- O1 optimize for maximum speed, but disable some optimizations which increase code size for a small speed benefit
- O2 optimize for maximum speed (DEFAULT)
- O3 optimize for maximum speed and enable more aggressive optimizations that may not improve performance on some programs
- Ox enable maximum optimizations (same as O2)
- Oi[-] enable/disable inline expansion of intrinsic functions
- SSE4.1 May generate Intel(R) SSE4 Vectorizing Compiler and Media Accelerator instructions for Intel processors. May generate Intel(R) SSSE3, SSE3, SSE2, and SSE instructions and it may optimize for Intel(R) 45nm Hi-k next generation Intel Core(TM) microarchitecture.
- AVX May generate Intel(R) Advanced Vector Extensions (Intel(R) AVX), Intel(R) SSE4.2, SSE4.1, SSSE3, SSE3, SSE2, and SSE instructions for Intel(R) processors.

B. Experiment Measurement Results

Fig. 2 summarizes the results of sequential code optimizations. Results of parallel optimizations are shown in Fig. 4. Only average values are shown in Fig. 4 due to the values having low deviation compared to Fig. 2.

TABLE II: BENCHMARK GRAPH, GRAPH DETAILS TAKEN FROM [2]

Name	nlpkkt160
N, *vertice count* 10^6	8.3
M, *edge count* 10^6	221.2
Diameter $\bar{d}$	26.5
Graph description	3D PDE-constrained optimization

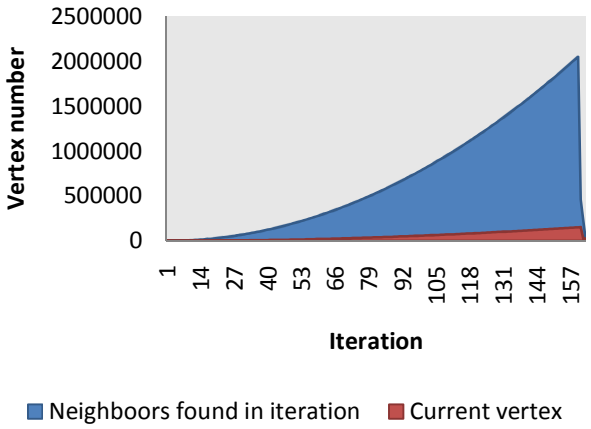


Fig. 5 Workload per iteration

The results are measured in seconds, with clock_t being used to measure the sequential performance and CUDA events being used to measure the parallel performance.

C. Analysis

1) Amdahl's Law

An OpenMP version of the algorithm was implemented with no regard to correctness to get a general behavior guideline.

Using the formula $\frac{1}{1-P+\frac{P}{N}}$, where P is the percentage of the code parallelizable and N is the number of workers used, P was calculated to be 41%. Results are presented in Fig. 3

2) Profiling CUDA Code

BFS as a parallel problem is highly dependent on the structure of the graph it is used on. The approach used calls for every vertex a worker, calculates whether the vertex needs to be considered in the iteration, then finds the neighbors. The static dispatching approach means that out of the 8345600 workers started in each iteration, on average only 51200 (0.613497%) are used. Since the number of possible neighbors for a given vertex is unknown in advance, the workers get tasks in highly variable sizes which leads to high load imbalance as seen in Fig.

In Table III a short overview of profiling results is presented. BFS consists of 3 kernel calls, two for thrust library kernels and one of the custom made kernel (Algorithm 3).

TABLE III: NSIGHT PROFILING RESULTS ON GTX 850M

Algorithm	Registers per Thread	Average Achieved Occupancy	Average Duration (μs)
BFS	40	0.62	1234
	32	0.01	13.99
	16	0.41	8136
			sum = 9384
BFS_1	11	0.49	4659
BFS_1 Atomics	11	0.52	3900
BFS_1 Share	10	0.52	3840
BFS_1 ShareAtomics	10	0.52	3886

VI. CONCLUSION

Graph algorithms are challenging to implement on GPUs and the reason for that is the distinctiveness of every graph. This paper has demonstrated parallel implementation of a breadth-first search algorithm on a dense graph and it was shown that an improvement can be made over the sequential implementation. With the parallel implementations a 1.9x speedup was achieved over the sequential implementation. The implementations are still slow for sparse graphs so other approaches need to be considered.

REFERENCES

[1] C. E. Leiserson, T. B. Schardl, *A Work-Efficient Parallel Breadth-First Search Algorithm (or How to Cope with the Nondeterminism of Reducers)*, 2010

[2] D. Merrill, M. Garland, A. Grimshaw, High Performance and Scalable GPU Graph Traversal, 2011.

[3] David A. Bader, Andrea Kappes, Henning Meyerhenke, Peter Sanders, Christian Schulz and Dorothea Wagner, *Benchmarking for Graph Clustering and Partitioning* In Encyclopedia of Social Network Analysis and Mining, pages 73-82. Springer, 2014. Available: http://www.cc.gatech.edu/dimacs10/downloads.shtml

[4] P. Harish, P. J. Narayanan, *Accelerating Large graph Algorithms on the GPU using CUDA*

[5] Gerassimos Barlas, *Multicore and GPU programming, an integrated approach*

[6] Lijuan Luo, Martin Wang, Wen-mei Hwu, *An Effective GPU Implementation of Breadht-First Search*, 2010

[7] ImageCFG [Online]. Available: https://robpol86.com/imagecfg.html

[8] Intel(R) C++ Compiler Help [Online]. Available: https://software.intel.com/sites/default/files/comment/1750288/icl12.1-options.txt

[9] Measuring Parallel Scaling Performance [Online]. Available: https://www.sharcnet.ca/help/index.php/Measuring_Parallel_Scaling_Performance

Buck converter controlled by Arduino Uno

H. Kovačević, Ž. Stojanović
Department of Electrical Engineering, Zagreb, Croatia
Polytechnic of Zagreb, Zagreb, Croatia
hkovacevi@tvz.hr, zeljko.stojanovic@tvz.hr

Abstract – The paper presents a buck converter controlled with microcontroller integrated on Arduino Uno board. The control is implemented by use of PI controller embedded on Arduino Uno board. Open loop control-to-output transfer function is obtained from measured step response. The PI controller and feedback divider transfer functions are synthesized to get desired loop gain. As a verification of analysis, input voltage step responses of regulated and unregulated buck converter are compared and improvements are identified.

I. INTRODUCTION

Buck converters are widely used in various types of electrical equipment and described at large in literature [1, 2]. The control of a buck converter is essentially important. There are many power converter control methods like linear control, fuzzy logic, predictive control, etc. [1, 3, 4]. Each of those has its advantages and disadvantages [3-5]. One of the most commonly used techniques of linear control is PI control because of its simplicity in design, implementation and understanding of operation [5, 6].

This article deals with a buck converter controlled by microcontroller ATmega 328 based on Arduino Uno board. It has been chosen because of its popularity and simplicity of use [7-10]. In this article, PI controller is realized by programming of Arduino Uno board. The converter is analyzed as a feedback system. Elements of feedback system are determined in order to achieve satisfying stability and time response. Simulation of input voltage step response for regulated and unregulated converter is performed. Programming code is shown in the appendix. Conclusions and proposals for next researches are done.

II. BUCK CONVERTER SCHEMATIC

Physical realization of the buck converter is shown in Figure 1. The basic circuit of the converter consists of a DC power supply E, switches $V1$ and $V2$, output filter L_d - C_d and load R_d. In the drive unit transistor $V3$ drives controlled switch $V1$. Resistors R_4 and R_5 are used for current adjustment. Capacitors C_1 and C_2 and resistor R_3 are used to improve switching characteristics. As a control device a microcontroller is used. The output voltage is sensed by microcontroller analog input pin A1. The driving transistor $V3$ is PWM controlled by digital pin D9.

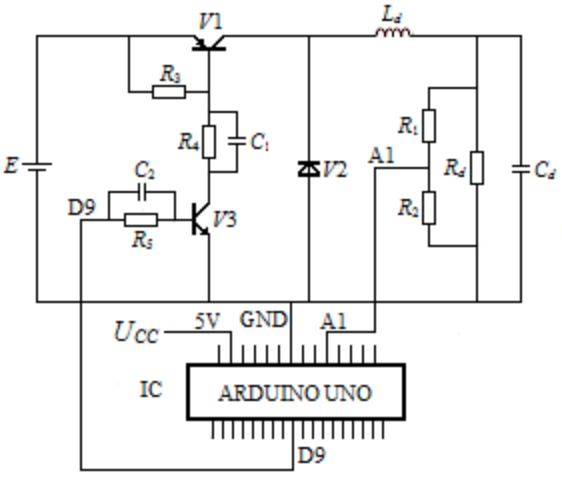


Figure 1 Buck converter schematic

A list of components and basic parameters of buck converter is shown in Table 1. The physical realization of buck converter is presented in Figure 2.

TABLE I. LIST OF COMPONENTS AND BASIC PARAMETERS

Component	Value	Component	Value
E	20 V	C_1	5,6 nF
$U_d(0)$	13,2 V	C_2	2,2 nF
$I_d(0)$	0,34 A	C_d	100 µF
$d = U_d(0)/E$	0,66	L_d	2,17 mH
R_1	470 kΩ	$V1$	MJE1501G
R_2	100 kΩ	$V2$	BC107B
R_3	27 Ω	$V3$	1N5408
R_4	220 Ω	Microcont.	ATmega 328
R_5	1 kΩ	Board	Arduino Uno
R_d	39 Ω	U_{CC}	5 V
Switching frequency	8 kHz	Sampling period	2 µs

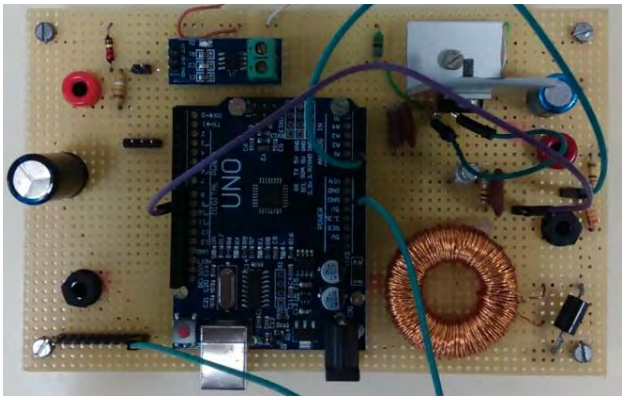


Figure 2 Physical realization of buck converter

III. BUCK CONVERTER CONTROL

Buck converter control can be realized with a digital or analog controller. A digital controller was chosen because of its parameter changes flexibility such as desired output voltage, switching frequency and PI controller proportional and integral parameters.

A. Block diagram of regulator system small-signal model

The buck converter is operating in voltage mode control. Feedback loop regulation of buck converter is shown in Figure 3. The output voltage $v(t)$ is measured by the microcontroller using a voltage divider with gain $H(s)$. The error signal

$$v_e(s) = H(s)v(s) - v_{ref}(s)$$

sets the duty cycle of PWM signal. Control algorithm was accomplished by using PI controller implemented in Arduino Uno board. For the analysis, whole schematic is presented by small-signal block diagram of the feedback system in Figure 4. List of all variables and functional block parameters of the feedback system and its descriptions is shown in Table 2.

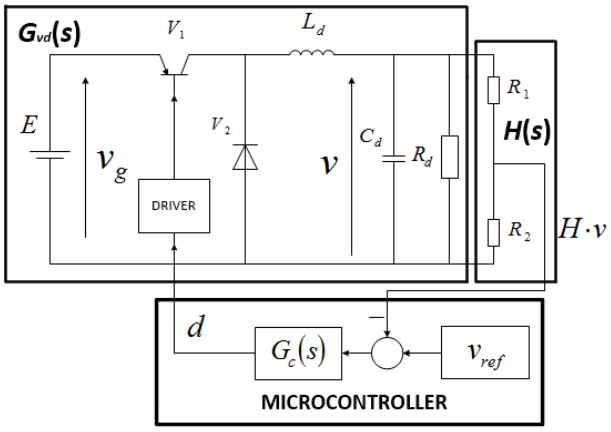


Figure 3 Feedback loop regulation of buck converter

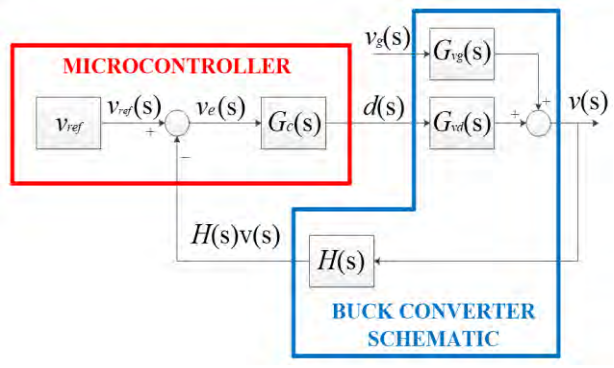


Figure 4 Small-signal block diagram of the feedback system

B. Obtained transfer functions of the buck converter

In order to control the converter successfully, transfer functions of all blocks have to be determined.

Buck converter open loop control-to-output transfer function

$$G_{vd}(s) = \frac{v(s)}{d(s)}$$

is determined experimentally by step change of the duty cycle $d(t)$ and measuring the transient of output voltage $v(t)$, Figure 5. The data is processed by System Identification Tool in Matlab. Obtained expression for control-to-output transfer function $G_{vd}(s)$ is shown in Table 3 and Bode plot is shown in Figure 7.

Buck converter open loop line-to-output transfer function

$$G_{vg}(s) = \frac{v(s)}{v_g(s)}$$

is determined experimentally by step change of the input voltage $v_g(t)$ and measuring the transient of output voltage

TABLE II. PARAMETERS AND VARIABLES OF SMALL-SIGNAL BLOCK DIAGRAM

Parameter	Description
v_{ref}	Reference input
v_e	Error signal
$d(s)$	PWM signal
$v_g(s)$	Input voltage
$v(s)$	Output voltage
$G_c(s)$	Transfer function of PI controller
$G_{vd}(s)$	Control-to-output transfer functions
$G_{vg}(s)$	Line-to-output transfer functions
$H(s)$	Sensor gain
$T(s)$	Loop gain $G_{vd}(s)$ $H(s)$ $G_c(s)$

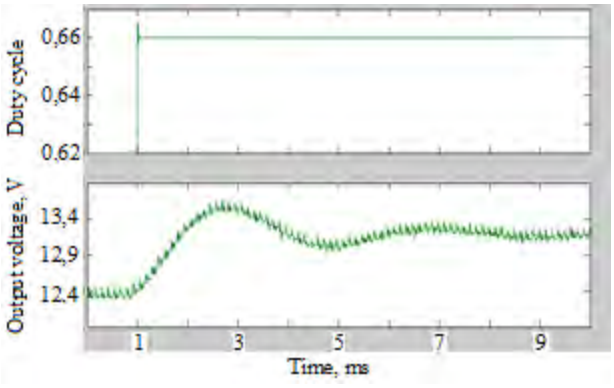


Figure 5 Unregulated output voltage step response on change of duty cycle $d(t)$

$v(t)$, Figure 6. Obtained expression for control-to-output transfer function $G_{vg}(s)$ is shown in Table 3. The shape of transfer function $G_{vg}(s)$ is the same as transfer function $G_{vd}(s)$ but they differ for constant gain.

Sensor gain of the output voltage divider is

$$H(s)=\frac{R_2}{R_1+R_2}$$

Criteria for choosing the values of output voltage divider are not to exceed maximum analog input voltage of Arduino Uno (5 V) and not to load the output significantly. Chosen sensor gain is shown in Table 3 and in Figure 7.

Reference voltage $v_{ref} = 2{,}3$ V is chosen to be in the middle of the range of analog input voltage of Arduino Uno.

Transfer function of PI controller is

$$G_c(s)=\frac{d(s)}{v_e(s)}$$

The parameters of PI controller are determined by the Bode plots using Simulink. PI controller parameters are chosen to

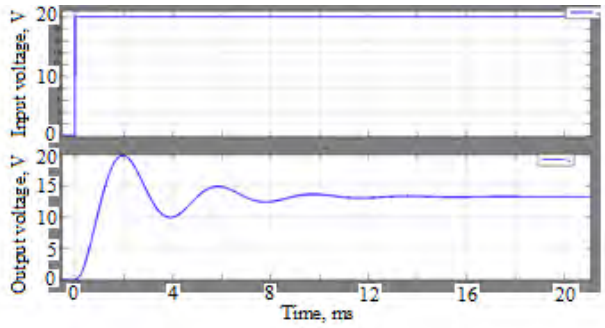


Figure 6 Free response of unregulated output voltage on step change of input voltage $v_g(t)$

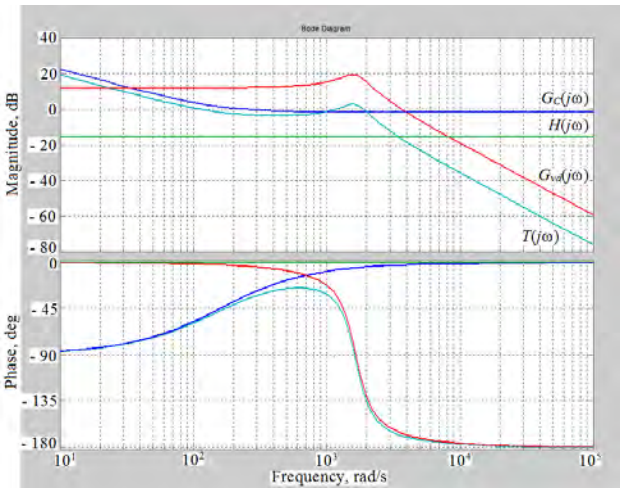


Figure 7 Bode plots of $G_{vd}(s)$, $H(s)$, $G_c(s)$ and $T(s)$

achieve phase margin of desired loop gain

$$T(s) = G_{vd}(s)H(s)G_c(s)$$

about 60°. PI controller's crossover frequency (determining the $G_c(s)$) is chosen to be 5-10 times lower than crossover frequency of the buck converters open loop system (derived from $G_{vd}(s)$). Obtained expression for transfer function of PI controller $G_c(s)$ is shown in Table 3. Bode plots of $G_{vd}(s)$, $H(s)$, $G_c(s)$ and $T(s)$ are shown in Figure 7.

IV. VERIFICATION BY SIMULATION

Simulation is done in Simulink. The simulation model is shown in Figure 8. Input voltage step response for unregulated and regulated output voltage is simulated.

Free response of regulated output voltage on step change of input voltage $v_g(t)$ from zero to full value ($E = 20$ V) is shown in Figure 9. Comparing this results to the unregulated output voltage response in Figure 6 it is visible that feedback prevents overshoot of output voltage.

TABLE III. EXPRESSIONS FOR PARAMETRERS AND VARIABLES OF SMALL-SIGNAL BLOCK DIAGRAM

Parameter	Expression
v_{ref}	2,3 V
$G_c(s)$	$G_c(s)=0{,}836\cdot\frac{6{,}33\cdot10^{-3}s+1}{6{,}33\cdot10^{-3}s}$
$G_{vd}(s)$	$G_{vd}(s)=\frac{3{,}99}{3{,}65\cdot10^{-7}s^2+260\cdot10^{-6}s+1}$
$G_{vg}(s)$	$G_{vd}(s)=\frac{0{,}66}{3{,}65\cdot10^{-7}s^2+260\cdot10^{-6}s+1}$
$H(s)$	0,175

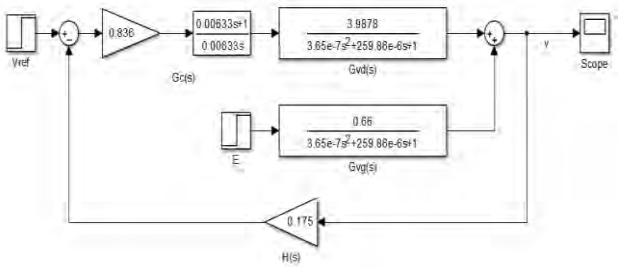


Figure 8 Simulation model for step changes of input voltage

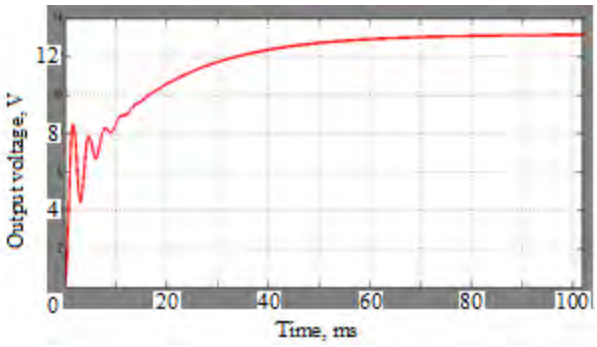


Figure 9 Free response of regulated output voltage on step change of input voltage $v_g(t)$

Simulation is also done for the case when converter is in steady state and there is increase of input voltage for 1 V. In this case, the waveforms of output voltages are shown in Figure 10 and Figure 11.

In a case of unregulated output voltage the steady-state error is significant and it can be calculated by multiplying change of input voltage by duty cycle. In the case of regulated output voltage the steady-state error is much smaller. Some of the parameters of quality of the transition process in the time domain read off from Figure 11 are shown in Table 4.

V. CONCLUSION

The buck converter controlled by Arduino Uno microcontroller board is realized. The converter control is implemented by PI controller which is built in Arduino Uno board. PI controller is synthesized to accomplish desired closed loop gain. Simulation of closed loop line-to-output step response validated proposed design. Impact of proportional and integral parameters of PI regulator on converter control is significant. In a future work, proposed design will be validated by measurements and compared with analog PI control.

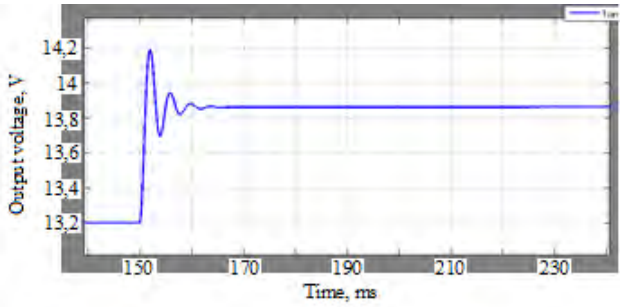


Figure 10 Unregulated output voltage by step of input voltage for 1 V

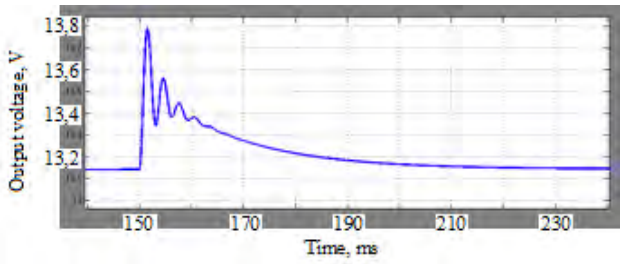


Figure 11 Regulated output voltage by step of input voltage for 1 V

TABLE IV. PARAMETERS OF THE QUALITY OF THE TRANSITION PROCESS IN THE TIME DOMAIN

Parameter	Overshoot, V	Peak time, ms	Settling time ms
Value	0,65	1,5	90

REFERENCES

[1] R. W. Erickson, D. Maksimović, Fundamentals of Power Electronics. 2nd ed. Kluwer, 2001

[2] I. Flegar, Elektronički energetski pretvarači, Kigen, Zagreb, 2010

[3] K. Wan, Advanced current-mode control techniques for dc-dc power electronic converters, Missouri university of science & technology, PhD Thesis, 2009

[4] P. Cortes, M. P. Kazmierkowski, R. M. Kennel, D. E. Quevedo, J. Rodríguez, Predictive Control in Power Electronics and Drives, IEEE Trans. on ind. el., vol. 55, no. 12, pp. 4312-4324, December 2008

[5] S. Verma, S. K. Singh, A. G. Rao, Overview of control Techniques for DC-DC converters, Res. J. Engineering Sci., vol. 2, no. 8, pp. 18-21, August 2013

[6] F. Tahri, A. Tahri, A. Allali, S. Flazi, The Digital Self-Tuning Control of Step a Down DC-DC Converter, Acta Polytechnica Hungarica, vol. 9, no. 6, 2012

[7] A. Sinha, M. Pavithra, K. R. Sutharshan, S. K. Sahoo, Arduino based pulse width modulated output voltage control of a dc-dc boost converter using Proportional, Integral and Derivative control strategy, Australian Journal of Basic and Applied Sciences, vol. 7, no. 11, pp. 104-108, September 2013,

[8] S. Dutta, S. Agarwal, S. Gadgil, J. N. Hemalatha, A Flexible and Reconfigurable Controller Design for Buck Conveters, AICERA 2014 - International Conference on Magnetics, Machines & Drives, July 2014

[9] Arduino Uno datasheet

[10] Atmega 328 datasheet



APPENDIX

```
#include <PID_v1.h>
#include <TimerOne.h>   // Including library

#define output_voltage A1
#define PWM 9

// Variable declaration
double Kp=0.836, Ki=132, Kd=0; // PI controller
parameters
double
output=0,desired_voltage=13.2,voltage=0,duty_cycle=0,

PID myPID(&napon, &output, &desired_voltage, Kp, Ki,
Kd, DIRECT); // Initialization of PID controller

void setup() {
  pinMode(PWM,OUTPUT);
  Timer1.initialize(125); // Frequency 8 kHz
  myPID.SetOutputLimits(0 , 1023); // PID controller limits
  myPID.SetSampleTime(0.0735); // PID sample time
  myPID.SetMode(AUTOMATIC);
}

void loop() {
    myPID.Compute();
    Timer1.pwm(PWM,output); // Duty cycle set up

    //Measuring output voltage
    voltage =((analogRead(output_voltage)/1023.)*5)/0.21;
}
```

Audio phonebook for the blind people

Goran Popović*, Una Pale*
* Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
goranpopovic@hotmail.com, una.pale@gmail.com

Summary - The technology can in many ways help visually impaired people, but problem of memorizing a large number of phone numbers for most of blind people is still not resolved. The aim of this study was to design, construct and test a device that would serve as a phone book for landline phones. The device is connected between the phone and the phone line and therefore can serve as a memory for phone numbers. It is important that it has a very simple interface and allows unobstructed use of the phone. The device was tested by visually impaired people and feedback showed that it could be used without any problem and without special training.

I. INTRODUCTION

Long distance communication has become a crucial part of life of the modern man. Without the Internet it is difficult to perform even simple personal tasks, but without phone it is almost impossible. The emergence of the SMS has brought a whole new dimension of life to the hearing impaired persons. At the first sight phones make communication of blind people easier, but only seemingly. In fact, in everyday life we use dozens, even hundreds of phone numbers. It is almost impossible to know them by heart. Today most of the people remember only a few phone numbers which they call frequently because they rely on the phonebook in their mobile phones where phone numbers are stored. Blind people do not have a possibility to write the number on the paper, nor in the mobile phones (most of them). Some people record phone numbers on audio cassettes or CDs in order to have a sort of an audio phonebook, and some, albeit rare have phone numbers written in Braille.

It is a myth that blind people, because of blindness have an excellent memory. This actually applies only to those who are systematically working to develop their mnemonics. Just like in the general population, among blind people exists a significant number of those who are not skilled with technical devices as well as those with poor memory skills. To these people it is a problem to call more than a few different phone numbers.

Based on research of Internet and scientific publications as well as consulting with blind people and their associations, it was concluded that today there is no available device that solves this problem. Also, almost all devices that are suitable for the blind (like speaking clocks, scales, blood pressure monitors, etc.) are several times more expensive [1] than the same devices for people who can see. Most of the devices are also even more expensive if the language of the device is not English.

The aim of this study was to solve this problem. The hypothesis is that this problem can be solved by constructing a device that can be connected between phone device and phone landline that would serve as an audio phonebook. It would have just about three buttons in order to be easy to operate. Buttons would enable scrolling through contacts which would be pronounced (the contact's name) and automatically dialed by pressing the "dial" button once the desired contact is found. Such device would allow blind people much easier and faster handling with the telephone device. One needs to type the phone number only once, in order to program the phonebook, which greatly facilitates the use, since there is no need to remember a lot of numbers, nor for creating any alternative phonebooks. On the other hand ordinary phones, often because of inadequate interfaces (like a lot of small buttons that are inadequately positioned) make entering the phone numbers difficult. The proposed device would reduce the use of phone keys to a minimum.

The two biggest problems that blind people have are calling the wrong number and the fact that the call is interrupted due to the long pause between entering two digits (more than eight seconds). Wrong number is entered mainly due to inadequate phone interfaces for blind people. Interruption of the call happens because people often must look for a phone number in their own phonebooks.

The second hypothesis is that such audio phonebook would eliminate wrong calls which are today present due to false memory or mistakes in operating the phone keys. The third hypothesis is that the audio phonebook could be used also by people with reduced cognitive and motor abilities.

II. EXISTING SOLUTIONS TO THE PROBLEM

Research has shown that there are two main groups of solutions to this problem: alternative phonebooks and mobile phones.

A. Alternative phonebooks

Auxiliary phonebooks that blind people use are usually CDs, audio cassettes or notes/address books in Braille which contain contact's information. Usage of any of the mentioned phonebooks is very slow and clumsy.

It is hard to learn Braille, complicated to use and digital versions are very expensive. The phones with big buttons or with Braille labels can be bought, but still the problem of remembering a phone number is not resolved because the landline phones have a small memory: 10, 20 or 30 numbers. Even if such phones would have a larger memory, the problem is remembering at which memory

location is the desired number, and there is no feedback except waiting for who will answer the call. In some countries one can call a phone assistant and ask for help, but more information about the phoned person are needed, which can be a problem for people with reduced cognitive abilities. In some countries the law requires that manufacturers of phone equipment need to produce devices that are adapted for the blind, visually impaired, deaf or people with other disabilities, the requirements that need to be fulfilled are usually minimal.

B. Mobile phone solutions

Blind people have several feasible ways to call a large number of different phone numbers, but they are adapted for people with advanced technical competences: mobile phones with talkback and mobile phones with voice recognition.

Mobile phone with talkback has a program running in the background that pronounces screen's content [2]. With that program blind user can select a phonebook, and find the desired contact. The advantage of this solution is that it uses existing, conventional models of phones ("smartphones"), phonebook size is practically unlimited, and the additional cost is only for the voice program. The disadvantage is that coping with a multitude of programs available on mobile phone, without tactile (and visual) feedback, is cognitively very demanding, and requires intense training which can be very hard to overcome for many people.

Phonebook with voice recognition [3] appeared on mobile phones even before the era of smartphones, and it is also accomplished by means of a computer program on a mobile phone. The advantages are the unlimited number of contacts, any phone can be used and the cost is only for getting certain programs. The disadvantage is that such programs do not yet have a high level of recognition, so it is necessary to use very distinguishable names of the contacts: words that do not sound similar. This means that it is necessary to remember recorded (voice print) names of individual contacts, which limits the number of possible contacts. Another problem is temporary or permanent changes of the user's voice: throat illness, teeth problems, stress as well as noisy environment.

None of the available methods is suitable for all blind people, especially the elderly people who recently got blind or people with reduced motor and cognitive abilities.

III. CONCEPT AND REQUIREMENTS FOR THE DEVICE

The basic idea is to construct a standalone device that can be installed between the landline wires and the telephone. With simple, intuitive and robust user interface the device would enable all blind people to make calls, especially the elderly and those with reduced motor and cognitive abilities.

A. Required functionalities

One of the main requirements is that the device does not limit the usage of phones and phone lines, regardless of the state of the device. This means that users can still make a phone call or answer an incoming call as if the device is not connected. Another requirement is the ability to save a large number of contacts (several hundred). Also, the user must be able to enter a new contact without anybody's help, at any time and at any position in the phonebook as he prefers. This means that moving, deleting and adding new contact anywhere in the phonebook must be possible.

The basic functionality is pronouncing the name of the contact that is selected. It is also important to be able to hear the phone number that is stored under each of the contacts. To use these options (finding the desired contact, adding a new contact, moving contact and deleting the contact) device must have voice instructions that guide the user through each function. Also, if there are a large number of contacts in the phonebook, functionality of rapid search is required and phonebook must be circularly shaped: after the last contact, comes the first, and vice versa.

B. User interface requirements

Several functions are mentioned that phonebook must have. For this reason, in order for operation to be simple, it is important that steps for each implemented functionality are logic and intuitive. Also, the physical interface must be easy to use.

The goal was to create a device with as few buttons as possible. Most important operation is „listing“, but also other operations such as calling, accepting and rejecting (e.g. just recorded name of the contact or entered phone number) and exiting (e.g. out of current option) are needed during the use of the device. All of operations can be implemented with one button, by different number of clicks on the button (one, two or three clicks). In such a way a total number of buttons can be reduced to only three buttons or even only one if encoder knob, which can be rotated and pressed, is used.

C. Flexibility requirements

Except mentioned functionalities, the goal is also that the device can be easily translated into any language. Contact names need no translation since they are entered by the user in the language he prefers. The only thing that needs to be translated into another language, are voice instructions that guide the user through the functionalities of the device. These voice instructions, as well as all the contacts, are stored on the SD memory card. Voice instructions are files in WAV format that can be recorded on any personal computer in a matter of minutes. In the same way contacts can be quickly and massively recorded, which is useful when initially storing a lot of contacts in an empty phonebook. This requires a separate program for the personal computer, for which a help of a sighted person is needed.

In addition, the goal is that the device is as cheap as possible and can be built by anyone. Therefore, the device is based on "open source" hardware so later anyone could make it, even the ones who have no engineering skills.

IV. DEVICE DEVELOPMENT

The device was built and its operation and functionality tested.

A. Hardware

General parts of the device are Arduino [5], DTMF encoder/decoder [6], speaker, microphone, memory (SD card), power supply, and the control circuit for the phone line as shown in Figure 1. The device is connected in parallel to the landline wires and the telephone. In that way device can "listen" what is happening on the line, in other words what the phone is sending, and does not disturb the line during the incoming or outgoing call when using the telephone. Device shares the line with the telephone and can send the phone number through the line instead of telephone, and then leave the rest of the call (speaking, transferring, hanging-up) to be handled by the phone. Device is electrically isolated from the landline wires and the phone through an audio transformer 1:1 with a 300Ω impedance which substituted the "existence" of the phone. Power is drawn through AC/DC adapter with voltage 7-12 V.

Figure 1 shows the circuitry of whole device.

B. Working principle

Today most telephones use DTMF [4] (Dual Tone Multiple Frequencies) standard, which defines combination of two frequencies that represent particular telephone digit/symbol. All phone devices that support DTMF dialing can generate 12 DTMF signals which represent digits/symbols: 1, 2, ..., 9, 0, '*' and '#'.

For the realization of proposed solution platform "Arduino" has been chosen, because it is "open source" and relatively cheap. The entire device is designed as a "shield" board which is just stacked on top of the Arduino board. In order to have phonebook that is easily transferable and editable, all contacts and corresponding telephone numbers are written in a single .txt (ASCII) file stored on the SD memory card. The phone numbers in the file are in the following format:

001-90981862526#########-018-002-
002-016632567##########-001-003-
003-003851676148########-002-022-
RRR-NNNNNNNNNNNNNNNNNNN-PPP-SSS-

First three digits (RRR) are the number that denotes the name of the file (.wav) where contact name audio recording is stored, but also the number of the row of number in .txt file. 20 characters (NNN…N) are reserved for the telephone number. If the number has less than 20 digits, the rest of the characters are replaced with symbol '#'. Last two three-digit numbers represent previous (PPP) and following (SSS) contact (its position in .txt file and the name of the .wav file). In such a way a linked list is created with pointers to next and previous contact.

V. DEVICE TESTING

One the device was built and its operability tested, in order to examine functionality and suitability of user interface for the visually impaired, the device was given to blind people to test it. In cooperation with the Croatian association of the blind functionality and user interface was tested by 11 blind or visually impaired users. It was planned to test the device on more people, but not all were able to participate in the available time frame. Six examiners were older than 50 years, and five were younger. Although the group of 11 examiners is relatively a small group from which statistically significant results cannot be obtained, we believe that the homogeneity of the results can confirm quality of functionality of the proposed solution. The review of similar studies with the blind [7-10] revealed that the number of the examiners was also small (under twenty examiners).

A. Testing procedure

1) Initial poll

Firstly, the initial survey was conducted, which was aimed to find out the person's profile, their habits of using the telephone, and their opinions on problems about devices adapted for blind people.

The survey consisted of three parts: personal data, issues related to the current way of using the telephone, and general issues related to the life of the blind people. In personal data, the most important data were: age of the person, whether the person is visually impaired or blind, and for how long. Also, it was important to know whether the person through their employment and life was engaged with computers and other complex technical devices, and whether they use Braille. For general questions, examiner's attitudes about the prices of devices for the blind and visually impaired, and the sensitivity of manufacturers and engineers for the blind and visually impaired people in the development of new devices were addressed.

Part of the survey, related to the current way of using telephone, comprised of first determining in which category the examiner may be put. Several categories were suggested:

- classical mobile phone with keyboard
- mobile phone with talkback
- remembers phone numbers by heart and types every number on the phone
- personal phonebook (on paper, cassette, CD, computer, etc.) but types the numbers every time
- landline phone with memory using pre-stored numbers only or predominantly

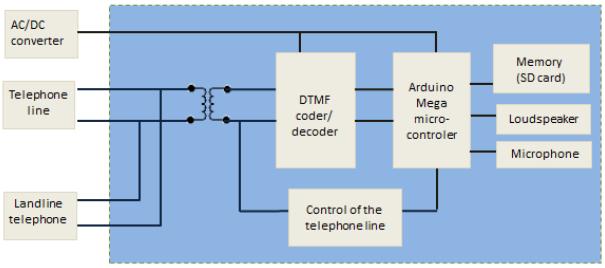


Figure 1. Basic electrical sheme of device

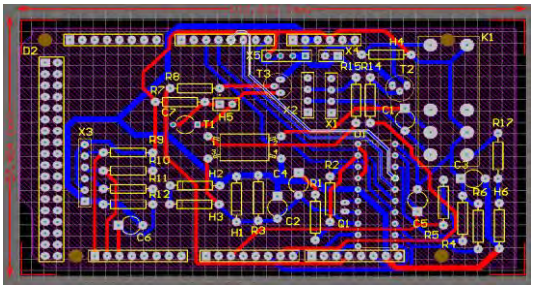


Figure 2. Electrical design of device

Important information were the number of the telephone numbers which examiners remember, number of the telephone numbers which examiners have in phonebook, average time needed to call a contact or add new contact in the phonebook, rate of calling a wrong contact, problems with current phonebook and telephone interface and autonomy they have when using the telephone. Furthermore, if they have been unsatisfied with current way of calling, problems which occur to them and their ideas for solving that problem were recorded in the poll. Also, participants were asked to evaluate if there is a need to create device that would allow blind people to call others easier, as well as what would be the acceptable price of such a device relative to prices of other devices for the visually impaired.

2) Testing the device

Testing consisted of three parts: calling the contact, entering a new contact and deleting the contact. For each of these three parts, the examiner would first call, enter and delete contact with assistance of the mentor. This was repeated until the examiner was prepared to do it without the help of the mentor. The amount of time for calling, entering and deleting the contact was measured, along with the number of questions examiner had and the number of attempts. Afterwards the time required for the examiner to find a person in the phone book and call it, to record a new contact, and to delete recorded contact was measured. Both of the interfaces were tested in order to determine which interface is better and why.

3) Closing poll

Closing poll was conducted after the testing. The aim was to evaluate examiner's satisfaction with the device, as well as the views of the examiners about offered device as a solution to problems which blind have when using the telephone. Their advices were collected for improvement of the device. They also had to decide which interface they find better, easier and more logical/intuitive.

B. Testing results

1) Initial poll results

In this section just few of the most important feedbacks are discussed. The number of contacts that people have in their phonebooks depends on the calling method.

Average number of contacts for people with smartphones is more than 80 contacts, whereas for people with ordinary mobile phones around 6. We wanted to know whether examiners need other people's help when making a call or editing the phonebook, and for which operation. 73% of examinees can do everything by themselves. Mostly senior examiners who are blind for less than five years need help of others. They need help rarely, when entering contact (e.g. contact in phonebook on mobile phones, speed dial in older mobile phone with keyboards or a landline phone). Based on the answers of the examiners, average time for creating new contact in the phonebook for mobile phones with keyboards is more than two minutes, while with smartphones average time is under one minute.

Average time to call a contact using their current calling device was 10-30 seconds. Half of the examiners dial wrong number very rarely, almost never, while the other half dials wrong number in average several times a month. A reason for mistakes is lack of audible feedback which number is pressed on older phones, and lack of saying the name of the contact selected when listing the phonebook or the name of the contact called when using the speed dial. Also, a common mistake is simultaneously pressing two buttons, or having too long pause (> 8 sec) between the two subsequent number entries during which line disconnects.

In total, 91% of examiners agreed that there is need for new solutions which would ease telephone usage among the visually impaired. However, examiners who are accustomed with smartphone stated that new solutions are needed mostly by elderly and people who are visually impaired for a relatively short time. All agreed that device manufacturers and engineers are not sufficiently sensitive to the needs of the blind. Also, all stated that the prices of the devices adapted to visually impaired are too high. Examiners estimated that the price of the audio phonebook they have tested should be in range of 200-400kn (33%), from 400- 700kn (44%), and 700-1000kn (22%).

As problems with their current way of using the phone they mentioned: the size of the keys, the smoothness of the keys, the screen contrast, the complexity of phone interface, the missing feedback that they are calling the right number, short maximum pause allowed between two numbers when dialing, a small number of contacts one can have on speed dial or in memory of a landline phone, and the complexity of creating new contacts.

2) Device testing results

The time that examiners needed to call certain contacts from the phonebook was measured (three tries).

While calling the contacts, examiners were able to ask questions. Elderly people asked questions for about 7 minutes in total, while the younger people in average 3 minutes.

Figure 4 shows that the time for calling a contact is shorter for younger people (<50 g.): Around 20 seconds. For seniors, at first test that time was higher (37s), but in three attempts the time decreased. The goal was to determine the approximate time it takes for the average person to call a desired contact. 30 contacts were stored in the memory and contacts they needed to call in two consequent attempts were equally spaced (8 contacts). Based on the results for younger and older people, that time is around 25 seconds. Slight increase in time for younger examiners is practicaly insignificant and probably attributed to fatigue, distraction and other factors. Consequently, the user interface is simple enough to call a contact very quickly, regardless of age, previous experience and skills of the user.

The time needed to erase a contact was also less for younger people. Time needed to erase a contact is longer than time needed to call the contact because it takes more steps to erase the contact than to call it. Entering a new contact lasted even longer, because in addition to entering a telephone number, user must record name of the contact which was very unusual to the examiners in the beginning. First attempts took around 90 seconds for younger examiners and about 150 seconds for senior examiners. In both cases, the time decreased in the following attempts.

The time (for calling, entering and deleting contact) tend to fall with number of repetitions, especially for the elderly. It can therefore be concluded that people can rapidly learn to use the device.

3) Closing poll

All examiners have answered that the device had fulfilled their expectations. They also said that they could use the device in their homes and that it has all the functionalities they need. Furthermore all said that the device is enough intuitive to be used. Examiners who are already using smartphones said that they do not need this device, which is understandable. As advantages over the current way of using the phone they mentioned short calling time, insurance that they are calling right contact, the ability to edit the order of the contacts by the priority, and the ease of entering a new contact.

In total 73% of the examinees opted for the interface with rotational knob, while other 27% opted for interface with three buttons.

Examiners were asked to provide as many proposals to improve the interface as they could think of. Mostly the proposals were oriented to new functionalities, such as a sound signal before recording of the contact's name starts, the option of saving an incoming number, identifying incoming calls by saying the name of the caller, etc.

VI. CONCLUSION

The blind examiners have concluded that the device with its interface and functionalities is fully adapted to their needs, and the overall impression was positive. The first hypothesis was confirmed: it is possible to make a device that would help all blind people.

The designed and developed audio phonebook for the blind people can facilitate the usage of the phone for blind. The device solves the main problems that the blind people have when making a phone call: stores a large number of contacts, completely eliminates the risk of dialing the wrong number or being interrupted because of slow dialing. This confirms the second hypothesis.

Furthermore, several advices for further improvements of the device were collected. The advices were mainly oriented to new functionalities, and not to the modification of existing functionalities or interface.

Except the feedback on the device itself, a profile of blind people was created, their needs, problems, current solutions. An overview on part of the community of blind people, to whom this device could help, was acquired. Those people are ones who did not manage to get used to the new technologies, especially on touchscreen. To those people the device would not only enable easier dialing of a phone number, but would also solve the problem of need to memorize a lot of contacts. In this research, more than half of the examined people have belonged to that category. This also indicates that the third hypothesis could be confirmed, but a carefully selected group of examiners is required in order to get scientific proof. They might now be able to delete or add new contacts, depending on severity of their cognitive impairment but calling could be obtained.



Figure 3. Final design of the device, two interface versions: buttons and rotational knob

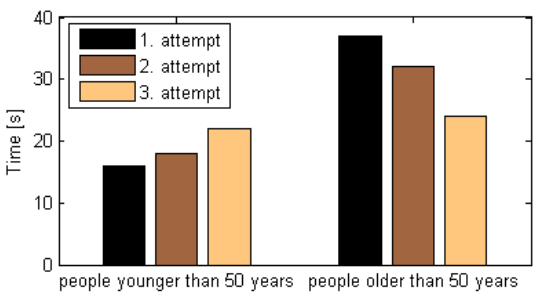


Figure 4. Calling time for two age groups

Additional feedback obtained from the examiners indicates the need for a similar user interface on other devices they use in everyday life, which represents a potential market niche at the global level.

In further work more examiners would need to be included in the research in order to get the statistically significant results.

ACKNOWLEDGEMENT

We would like to thank Predrag Pale for the initial idea of the device, as well as for support and help during the work on the project. We also thank the Croatian association of the blind: "Hrvatski savez slijepih i slabovidnih" for assistance in carrying out the tests. Thanks go also to all examiners that participated in the research and tested the device.

LITERATURE

[1] http://www.savez-slijepih.hr/hr/clanak/sljepoca-teskoce-i-dodatni-troskovi-njome-prouzroceni-455/, (7.4.2015.)

[2] http://www.androidcentral.com/what-google-talk-back, (7.3.2015.)

[3] R.Yousef, O.Adwan, M.Abu-leil."An enhanced mobile phone dialler application for blind and visually impaired people" , International journal of engineering and technology, 2 (4) (2013) 270-280

[4] D.Nassar. „DTMF Encoding and Decoding", Circuits (radio electronic), Dec.1986

[5] Arduino Mega 2560: http://www.arduino.cc/en/Main/arduinoBoardMega (12.4.2015.)

[6] MT8880 datasheet: http://pdf1.alldatasheet.com/datasheet-pdf/view/77086/MITEL/MT8880.html (17.4.2015.)

[7] M.Sečujski, D. Pekar. „Evaluacija različitih aspekata kvaliteta sintetizovanog govora", Fakultet tehničkih nauka Novi Sad; AlfaNum d.o.o., Novi Sad

[8] http://www.savez-slijepih.hr/hr/clanak/ii-metodologija-istrazivanja-1440/ (28.1.2015.)

[9] M.Sečujski. „Ka automatskoj sintaksnoj analizi rečenice na Srpskom jeziku", Fakultet tehničkih nauka Novi Sad

[10] K. Nenadić. „Relaksacija kao pomoćna metoda peripatološkog programa", Sveučilište u Zagrebu, Edukacijsko rehaibilitacijski fakultet, 1999.

HEART RATE VARIABILITY ANALYSIS USING DIFFERENT WAVELET TRANSFORMATIONS

Una Pale*, Florian Thürk**, Eugenijus Kaniusas**
* Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
** Vienna University of Technology, Institute of Electrodynamics, Microwave and Circuit Engineering, Vienna, Austria
una.pale@gmail.com

Abstract - Heart rhythm is modulated by the autonomic nervous system in order to adapt to changing cardiovascular loads. The resulting heart rate variability (HRV) is considered as one of the most promising non-invasive markers of the activity of the autonomic nervous system. In this work, the possibility to analyze HRV with wavelet transform was investigated. Specifically, three methods of wavelet transform (CWT, CWTFT and DWT) were compared. DWT showed to be the best one because of the simplicity of decomposition, and quality of reconstruction, high time resolution as well as the shortest computation time and lowest CPU power consumption. DWT showed to have high noise robustness, excellent temporal and good frequency resolution and small influence of the mother wavelet type. ULF, HF and LF components of HRV were assessed using DWT transform. LF/HF ratio was extracted from HRV signal during voluntary breathing apneas and compared to the respiration signal. An increase in LF/HF ratio corresponding to the times of breathing apnea could be observed, indicating that the proposed method could reveal specific changes in the HRV and that LF/HF ratio could serve as a good indicator of breathing apnea.

I. INTRODUCTION

Heart rate variability (HRV) refers to the variation of the intervals between consecutive heartbeats over time. Since the heart rhythm is modulated by the autonomic nervous system, HRV is considered as one of the most promising non-invasive markers of the activity of the autonomic nervous system. HRV power spectrum can be separated into three frequency bands with physiological importance: the ultra-low frequency (ULF) component, low frequency (LF) component and high frequency (HF) component (see Table 1).

Today's most common analysis methods of HRV are spectral analysis techniques using the Fourier transform, or fast Fourier transform (FFT), which assumes that the signal is stationary in time. The problem with such techniques is the lack of temporal resolution. The exact time instant of HRV changes remain therefore unknown. To overcome this limitation, time window frames are often used, so that small segments of the signal are analyzed, for example as in Short time Fourier transform (STFT). However, time-frequency resolution depends on the width of the window used. As a consequence, higher temporal resolution means lower frequency resolution and vice versa. From this reason Wavelet transform as a method which performs time- frequency analysis of non-periodic signal is studied, in order to asses ULF, HF, LF components of HRV.

Table 1. Frequency bands and corresponding frequencies

Ultra-low frequencies (ULF)	0.003 – 0.004 Hz
Low frequencies (LF)	0.04 – 0.15 Hz
High frequencies (HF)	0.15Hz – 0.4 Hz

II. HEART RATE VARIABLITY

Heart rate (HR) or heart pulse is a term that describes frequency of heart beats and is calculated by 1/RR (*60 to get bpm) for each two consecutive R peaks in the ECG signal. Normal HR at rest ranges from 60 to 100 bpm, but it varies according to body's physical needs and activity states (physical exercise, illness, stress, and sleeping, eating etc.).

HR is regulated by sympathetic and parasympathetic input to the sinoatrial node. Sympathetic nervous system (SNS) increases HR, e.g. by releasing hormones, and is activated during stress. On the other hand parasympathetic nervous system (PNS) controls routine functions of the body and mainly decreases heart rate. As a consequence, continuous interplay of SNS and PNS can be measured by the HRV. A static HRV indicates inhibited regulatory mechanisms and a potential pathology.

It is normal that HR changes in time, but the speed of changes (HRV) is what is very interesting to observe. It was shown that PNS is related to the power of HF band, reflecting short-term regulatory mechanisms, while on the other hand SNS activity (and partly PNS) is related to the power of LF band, indicating mid-term regulatory mechanisms. ULF band is related to very slow oscillations and indicates long-term regulatory mechanisms [1]. SNS activation presents a slow increase of HR meaning also reduction in HRV, while PNS activation causes fast decrease of HR and increase in HRV. A reduction of SNS and PNS activity on the other side will result in reverse effects. Increased activation of SNS can be caused by stress, physical activity, standing, 90˚ tilt etc. while activation of PNS can be attained by controlled respiration, cold stimulation of the face or rotational stimuli [1]. Physical exhaustion is indicated both by decrease in PNS and increase in SNS, while on the other side positive stress is indicated as increase in both PNS and SNS. SNS plays an important role in causes of arrhythmias and PNS reduces possibilities of arrhythmias thus having protective role. Altogether, PNS activity is sign of healthier people, while its reduced activity can indicate some type of dysfunction [1].

LF/HF ratio is sometimes introduced to mirror sympathetic-parasympathetic balance. During normal uncontrolled breathing, respiration seems to influence HRV for less than 10%, but controlled respiration increases this influence up to almost 50% [2]. During apnea or suspension of breathing, heart rate changes with frequencies in range from 0.01 to 0.04 Hz. This means that power of ULF band of HRV increases. Also at the same time LF/HF ratio is increased[2] as a consequence of higher sympathetic processes. Similar effect happens during sleep of patients with sleep disordered breathing, where increased LF/HF ratio reflects not only sympathetic dominance but also reduced parasympathetic control [3].

III. RELATED RESEARCH

In the past 20 years there were several papers focused on analysis of wavelet transform, its properties and possible application to HRV analysis. Their approaches were different since wavelet transform, unlike FFT, has many variations and possibilities to be analyzed and used. In [4] authors used a wavelet transform to build a simulated model of a HRV signal and to create an algorithm for HRV signal decomposition. For standard MIT database in [5] HRV data was analyzed on the basis of LF/HF ratio. Wavelet and Wigner Ville Transforms were used for data analysis. The aim of [6] was to examine a set of mother wavelet functions and its orders for implementation in HRV analysis and to highlight the benefit of this transform relating to today's methods. The power distributions in each of different levels and types of wavelets were analyzed. LF/HF ratio was studied on a subject during deep breathing test, and thus an idea of [6] was similar to the present work, but with a different approach. Whereas [6] researched a multitude of mother functions and power ratios, the present paper observes smaller number of functions but other wavelet transform properties were studied too, using various methods of analysis not only the power distribution. In conclusion of [6] more extensive evaluation was suggested, since their results have been obtained from a limited number of subjects. In [7] HRV was analyzed using wavelet and cosine packets. A comparison was made on the same database with results based on the short-term Fourier transform method.

On the other side, numerous papers focus on the application of wavelet transform in healthcare and, in particular, on HRV analysis. In [8] CWT has been used to evaluate the effect of local anesthesia on HRV parameters. The major goal of [9] was to obtain a method which allows completely noninvasive distinguishing of the patients with different levels of coronary artery disease. Evaluation of sub-frequency regions of heart rate variability in supraventricular tachyarrhythmia patients was done in [10] using wavelet packet transform. Wavelet packet transform was also used in [11] for an analysis of sympathovagal balance in patients with major depressive disorder. In order to diagnose and detect diabetes automatically in [12] DWT decomposition was performed.

This is only short overview of papers that were interested in wavelet transform itself and its application on HRV evaluation. It is clear that this is very interesting area and that this technique could further improve understanding of the interactions of the autonomic control systems with the cardiovascular system. While these papers technically analyzed wavelets and showed certain applications in medicine, the presented work compared three methods of wavelet transform by evaluating them with various techniques. At the end the best method was used on HRV analysis. By applying DWT, information not only HRs frequency components but also their timely characteristics can be obtained which are of crucial importance to explain physiological and pathophysiological events. Therefore the results of this paper might serve as starting point or component of some future research.

IV. WAVELET TRANSFORM

The most common spectral analysis technique is the Fourier transform, which assumes that the signal is stationary in time. One problem with such technique is lack of temporal resolution, limiting the precise detection of transient events. In order to overcome this drawback, short-time Fourier transform was designed which is an adapted Fourier transform in a way that only short sections of the signal are analyzed at a time. This maps the signal into two-dimensional signals where it is possible to determine the time instant of frequency changes. However, this information can be obtained only with a certain resolution which is determined by the size of time window. Consequently, once a particular size for the time window is chosen, it stays the same for all frequencies. However, many signals require a more flexible approach where we can vary the window size to determine more accurately either time or frequency.

This can be achieved by the Wavelet transform [13] since it offers variable size of the window. Wavelet analysis allows the use of long time intervals to observe more precise low frequency information, and shorter intervals to observe high frequency information. The name Wavelet transform origins from the fact that waveforms, instead of sinusoid function in Fourier transform, are used. The mother wavelet is a waveform of effectively limited duration that has an average value of zero. Wavelet analysis does not use a time-frequency region, but rather a time-scale region. This is because wavelet analysis works by breaking up a signal into shifted and scaled versions of the mother wavelet. Scaling a wavelet means stretching (or compressing) it, with small scale as compressed signal and large scale as stretched signal. Scale is related to the frequency of the signal in a way that smaller scale means larger frequency and vice versa. Shifting the wavelet introduces a time delay. Wavelet transform works in a way that the mother wavelet is compared sequentially to sections of the original signal. The result is a set of coefficients that are a function of scale and position. These coefficients can be represented and used in many ways. For instance, inverse transform

can be performed, which can be used to reconstruct the original signal.

In following subsections three types of wavelet transform will be represented: continuous wavelet transform (CWT), continuous wavelet transform using FFT (CWTFT) and discrete wavelet transform (DWT).

A. *Continuous wavelet transform - CWT*

In continuous wavelet transform, the correlation of signal with mother wavelet can be calculated for any scale and shift of mother wavelet. CWT is calculated by smoothly shifting the scaled wavelet function along the signal while correlation with that signal section is measured.

B. *Continuous wavelet transform using FFT - CWTFT*

The problem with CWT is that it is redundant [13] and there is no unique way to define inverse. This means that it is not possible to reconstruct signals from coefficients. CWTFT uses FFT of the wavelet function in order to reconstruct signal. Not all wavelet functions can be used in CWTFT. The wavelet is required to be a real value function and that its FFT has support on only positive frequencies. Wavelets that satisfy this admissibility condition are called analytical wavelets.

C. *Discrete wavelet transform -DWT*

In order to speed up calculation, discrete wavelet transform was used. In DWT "dyadic" (which means based on factor two) scales are used. Even though only few scales are used to cover the whole area of frequencies, the transform is much more efficient and equally accurate. In each level of transform a signal is decomposed on "detail" and "approximation". Details are low scale (high frequency) components of signal, while approximations are high scale (low frequency). In each consecutive level, approximation is decomposed into detail and new approximation whose scale is smaller and frequency is larger. Iteration of this process results in wavelet decomposition tree. This process could be continued indefinitely in theory, but in practice it is limited with the resolution of the signal. When individual detail consists of a single sample, the end level of decomposition is reached.

V. RESULTS ON WAVELETS

A. *Comparison of methods*

The basic idea was to analyze three methods of wavelet transform in order to compare their practical advantages and drawbacks and to choose the most suitable for HRV analysis. Several methods were investigated; pure coefficients comparison, reconstructed signal comparison, sum of coefficients for each of frequency band of interest (ULF, LF and HF), energy distribution and calculation time comparison. Pure coefficients were compared only for CWT and CWTFT since DWT has only few scale levels. In HF frequency range there is almost no difference between CWT and CWTFT, but as approaching to ULF frequencies differences increase (Figure 1). Unfortunately, appropriate explanation for that was not found but it is believed that it is because of some constraints of CWT for very low frequencies. Also some differences could be caused by slightly different boundary frequencies for frequency bands.

For comparison of reconstructed signals, original signal was reconstructed using coefficients for each frequency band separately. This could only be done with CWTFT and DWT method. Again, trends and shape were similar for HF and LF, only for ultra-low frequencies (UFL) differences ware more visible, but even then main trends were similar. In order to compare all three methods at the same time, sum of all coefficients for every frequency band was studied. Curve shapes are the same for all three methods with only minor differences for HF and LF band. For ULF band larger difference for CWT is noticed which agrees with previous results.

Distribution of energy through different frequencies was studied in order to compare methods with Fourier decomposition of signal. The results have shown very good correlation between frequency values of peaks of energy distribution and FFT spectrum.

The last thing to consider was computation time for different methods. CWT method consumes very much time and CPU power. CWTFT takes only few (< 5min) minutes to compute even for signals with almost million samples in comparison to CWT which takes few hours (<3h) for the same task. On the other hand DWT is even faster than CWTFT and it takes under one minute to do the decomposition.

In conclusion, DWT is the most suitable method for the wavelet decomposition for several reasons: time and CPU power consumption, simplicity of decomposition and reconstruction and the quality of decomposition and reconstruction itself.

B. *DWT method analysis*

1) *Quality of decomposition*

Quality of decomposition was tested using artificial sinusoidal signals of known frequencies. Frequencies within different frequency bands were used. Energy distribution globally and how well the peak of maximal energy corresponds to the frequency of sinusoid was studied. Reconstruction signals using separately HF, LF and ULF frequency bands ware observed (their amplitude and period) too.

Classification to the correct frequency band from reconstructions could only be performed precisely for frequencies which are in the middle of the frequency band. As the frequency gets closer to the boundaries, other bands too detect these frequencies, but with lower amplitude and wrong period.

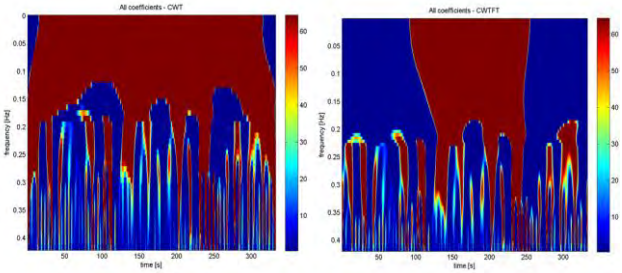


Figure 1. Coefficients comparison for CWT (left) and CWTFT (right).

In addition, the quality of detection based on energy distribution depends whether the frequency is closer or further from one of the frequencies corresponding to the scales that were used in decomposition (since DWT uses dyadic scales). Frequency detection was very precise for ULF and LF band, but less precise for HF band, since here, the frequency resolution is smaller and the error is larger.

2) Noise influence

Signal which consists of sinusoid waves of three different frequencies (each within one frequency band) was used. To this signal two different levels of Gaussian noise were added. Reconstruction and energy distribution for signal with and without noise were studied (Figure 2).

Results showed that noise didn't influence decomposition and reconstruction even when amplitude of noise was in range of signal amplitude (SNR 1:1). This is another very positive property of DWT transform.

3) Influence of chosen wavelet function

Finally, the influence of chosen mother wavelet function on energy distribution and reconstruction was studied. Four wavelets from Daubechie's family were used: "db3", "db5", "db8" and "db10" as shown in Figure 3. Result of comparison is that the difference is very small for LF and HF band, but for ULF there are noticeable differences. For LF and HF, reconstruction signals look very similar especially for "larger" Daubechie wavelet functions This can be explained by the fact that larger orders of Daubechie functions are more complex, but also more similar to each other than the ones of small order (for example Db1 is actually step function and Db3 is shown in Figure 3). For this reason it is suggested to use enough large Db wavelet functions (≥Db8) for precise calculations. It should be noted, that reconstructed signals in each band are not sinusoid functions as one may expect. This is due to the fact that borders of frequency bands in DWT are not very precise.

4) Temporal resolution

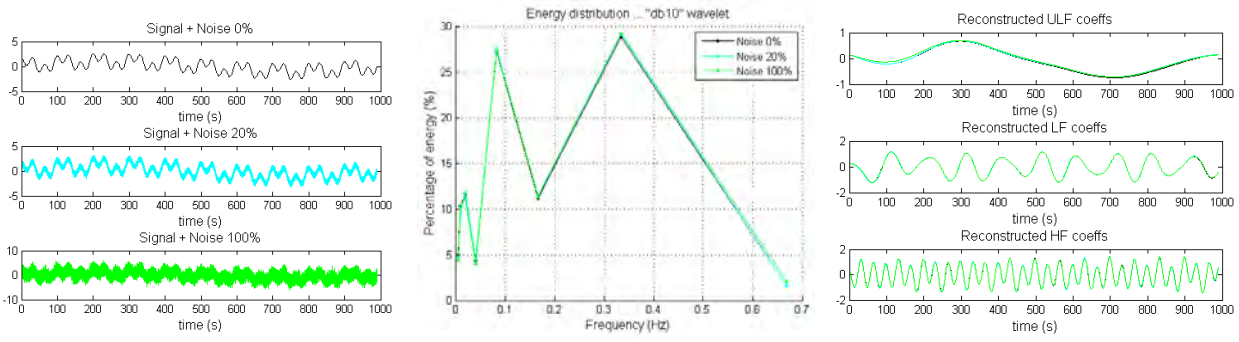
For this analysis, a signal with frequency variations was used. First part of the signal is sinusoid with frequency of 0.3Hz which is in HF band, then sinusoid of 0.1Hz (LF band) follows and at the end sinusoid of 0.01Hz (ULF band). It can be seen from reconstruction signals for different frequency bands that time detection of each sinusoid within input signal is very precise (Figure 4). Also energy distribution clearly shows peaks on frequencies values of original signal. This confirms DWT ability to analyze signal in frequency and time domain.

VI. RESULTS ON HRV

A five minute HR signal, sampled with 1000 sample/s, is used to present input and output data of HRV analysis using DWT method. In Figure 5a) original signal and DWT coefficients (absolute value) separately for ULF; LF and HF frequency band are presented. As previously mentioned, PNS is the major contributor to HF component. Slight disagreement exists in respect to the LF component. Some studies suggest that LF is a quantitative marker of SNS activity, while other studies view LF as reflecting both sympathetic and parasympathetic activity. Consequently, the LF/HF ratio is sometimes introduced to mirror sympathetic-parasympathetic balance and also to reflect the sympathetic modulations.

LF/HF ratio is obtained by dividing values of LF coefficients with HF coefficients. Absolute value is used because we are interested in "power" of the LF and HF band in time. After division, LF/HF ratio was averaged in time using median filter with a window size of 1000 samples which corresponds to one second. This way resolution of LF/HF ration is one second which is reasonable considering physiological modulations.

Respiration and HR signals were measured at the same time. Deep breathing or apnea was performed during recording. Such tests are usually used in clinical testing or calibrations. In the signal on Figure 5b) it can be noticed that LF/HF ratio is increased in times that precisely correspond to the times of breathing apnea from respiration signal which indicates that DWT can serve as LF and HF signal extraction from HRV and that such calculated LF/HF ratio can serve as a good indicator of breathing apnea. Moreover, voluntary apnea recordings from 18 subjects, it was shown that the proposed method could reveal specific changes in the HRV.



a) b) c)

Figure 2. Original sine signal with different levels of noise (a), energy distribution (b) and reconstruction (c).

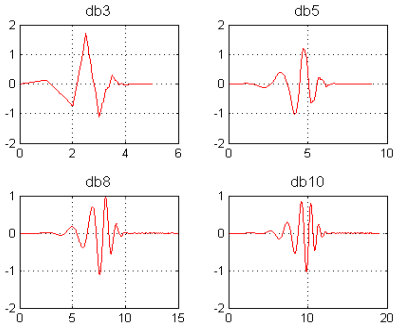


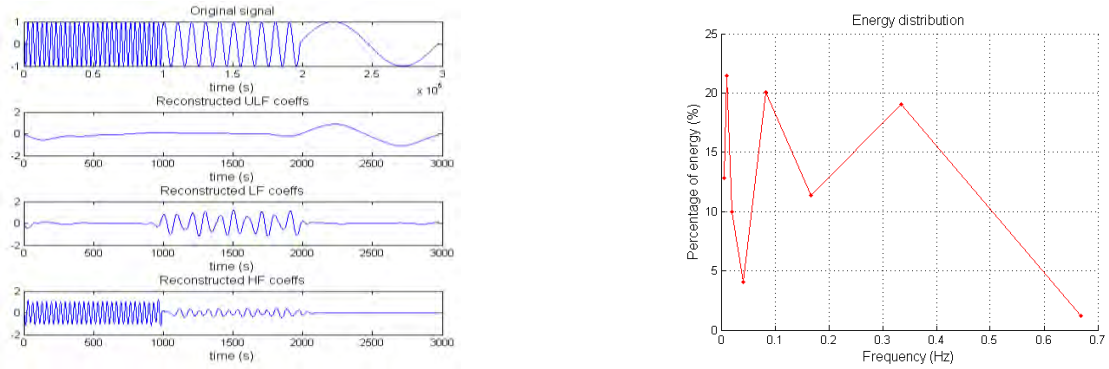
Figure 3. Different types of Daubechies wavelet function

VII. CONCLUSION

In this work, performance and possibilities of different wavelet transform for HRV analysis were investigated. Three methods of wavelet transform were compared using reconstruction for different frequency bands, sum of coefficients for every frequency band, energy distribution per frequencies and computational time needed to conduct decomposition. DWT showed to be the most suitable technique in terms of computational time, CPU power consumption, simplicity of decomposition and reconstruction and analysis quality.

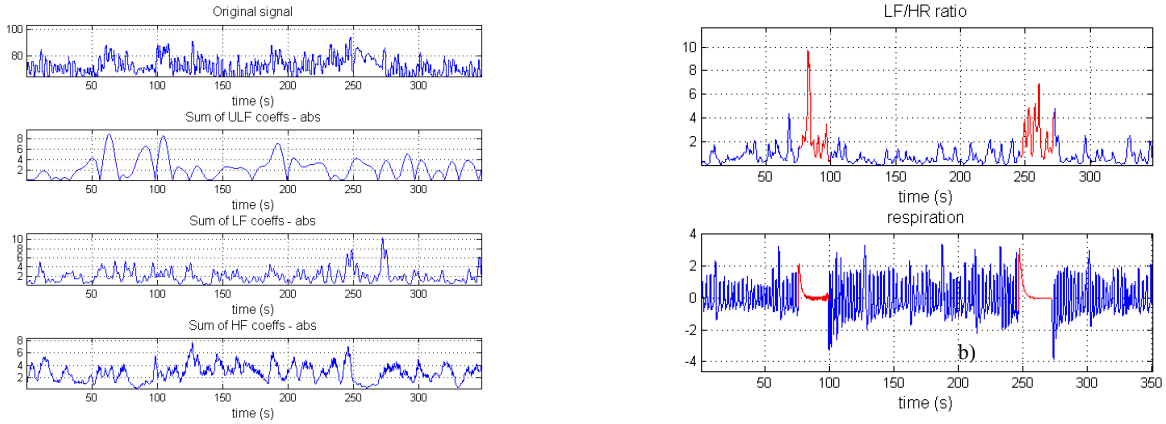
Quality of DWT method was further evaluated, observing noise influence, mother wavelet type influence and time resolution. Quality of detection based on energy distribution depends whether the frequency is closer or further from one of the frequencies corresponding to the scales that were used in decomposition. Even though the mapping to exact frequency bands is not perfect, reconstruction using coefficients from decomposed signal is very accurate. Noise didn't influence decomposition and reconstruction even when SNR was 1:1. In order to achieve precise decomposition and reconstruction some "large" (≥Db8) mother wavelet function from Daubechies family should be used. Temporal resolution was shown to be very precise, which confirms DWTs' ability to analyze signals in frequency and time domain.

Discrete wavelet transform showed to be a suitable candidate for fast and precise HRV analysis with potential application for scientific and clinical purposes.



a) b)

Figure 4.Test of temporal resolution a) Reconstructed signal b) Energy distribution



a) b)

Figure 5. DWT applied to HR signal a) Absolute value of sum of coefficients b) LF/HF ratio and corresponding respiration signal

ACKNOWLEDGEMENT

This work was done during Erasmus+ exchange on Technical University of Vienna. I would like to thank to the University of Zagreb and Erasmus+ program that gave me opportunity and financial support to be one of the exchange students.

LITERATURE

[1] E. Kaniusas: Biomedical sensors and signals I, Springer 2012.

[2] V. Demchenko,R. Čmejla, J. Vokřál: Analysis of heart rate variability during respiration

[3] I. Szollosi, H. Krum, D. Kaye, M. T. Naughton: Sleep Apnea in Heart Failure Increases Heart Rate Variability and Sympathetic Dominance, Sleep Journal, 2007.

[4] F. Yang, W. Liao: Modeling and decomposition of HRV signals with wavelet transforms, IEEE Engineering in Medicine and Biology Magazine 1997

[5] R. A. Deshpande, D. S. Ramdasi: HRV analysis using wavelet transform and wigner - Ville transform, Biomedical Engineering and Sciences (IECBES), 2012

[6] D. Singh, V. Kumar, M. P. S. Chawla: Wavelet Filter Evaluation for HRV Signal Processing, Advances in Medical, Signal and Information Processing, 2006. MEDSIP 2006.

[7] U. Wiklund, M. Akay, U. Niklasson: Short-term analysis of heart-rate variability of adapted wavelet transforms, IEEE Engineering in Medicine and Biology Magazine 1997

[8] K. Shafqat, S. K. Pal, S. Kumari and P. A. Kyriacou: HRV analysis in local anesthesia using Continuous Wavelet Transform (CWT), 33rd Annual International Conference of the IEEE EMBS

[9] E. Tkacz, P. Kostka, D. Komorowski: An application of wavelet transform and adaptive filters for decomposition of the HRV signals in the case of patients with coronary artery disease, Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE

[10] S. Bilgin, O. H. Colak, O. Polat, E. Koklukaya: Evaluation of sub-frequency regions of heart rate variability in supraventricular tachyarrhythmia patients, Computers in Cardiology, 2008

[11] S. A. Akar, S. Kara, V. Bilgic: Analysis of sympathovagal balance in patients with major depressive disorder using wavelet packet transform, Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2015

[12] U. R. Acharya, V. K. Sudarshan, D. N. Ghista, W. J. E. Lim, F. Molinari, M. Sankaranarayanan: Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform method, Knowledge-based systems, 2015

[13] L.G.Gamero, M. Risk, J.F.Sobb, A.J.Ramirez, J.P.Saul: Heart rate variability analysis using wavelet transform,IEEE Computeirs in Cardiology 1996.

[14] C.A.Garcia, A. Otero, X. Vila, D.G.Marquez: A new algorithm for wavelet-based heart rate variability analysis, 2014

[15] Task Force of the European society of cardiology and the North American society of pacing and electrophysiology: Heart rate variability - Standards of measurement, physiological interpretation, and clinical use, European Heart Journal, 1996

[16] M.Misiti, Y.Misiti, G.Oppenheim, J.M.Poggi: Wavelet Toolbox™ 4 - Getting Started Guide, The Mathworks, 1997

[17] M.Misiti, Y.Misiti, G.Oppenheim, J.M.Poggi: Wavelet Toolbox for use with Matlab, The Mathworks, 1997

Istraživanje *ransomware* napada i prijedlozi za bolju zaštitu

Marin Rak, Marinko Žagar
Tehničko Veleučilište Zagreb
marin.rak@tvz.hr, marinko.zagar@tvz.hr

Sažetak - *Ransomware* zlonamjerni programi napravili su veliku štetu organizacijama, ali i pojedincima. Prema predviđanjima za 2016. godinu, *ransomware* će biti jedna od najvećih prijetnjā u ovoj godini ne samo za tvrtke nego i za pojedince. Dosadašnje iskustvo pokazalo je da su najčešći vektori napada iskorištavali korisničko neznanje, ali i neažuriranje programskih paketa na napadnutim računalima. U ovom su radu predstavljeni rezultati istraživanja na napadnutim sustavima, vektori napada, razlozi koji su doveli do zaraze te prijedlozi za poboljšanje zaštite od ove vrste zlonamjernih programa.

I. UVOD

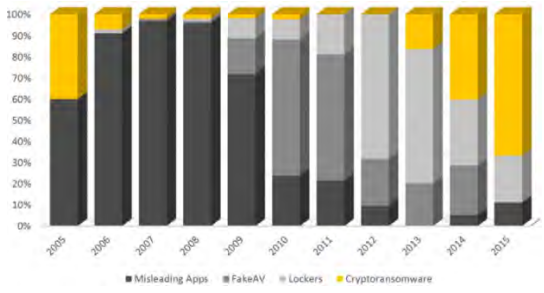
U vrijeme kada naši osobni i poslovni digitalni dokumenti, slike, filmovi i sl. počinju definirati i predstavljati naš život, bez kojih „kao“ da ne postojimo, tvorci zlonamjernoga koda osmislili su načine da upravo pomoću kripto *ransomwarea* otmu žrtvine podatke i zatim za njih traže otkupninu. Težište je ove studije slučaja, u kojoj se pokušalo odgovoriti na pitanje kako uspješno zaštititi poslovni informatički sustav od ove vrste zlonamjernoga programskoga koda, na povijesti napada *Filecoder* kripto *ransomware*a, vektorima zaraze sustava, mehanizmima djelovanja i na načinu naplate otkupnine.

A. Ransomware

Ransomware zlonamjerni kod dobio je ime kombiniranjem riječi *ransom* (otkupnina, ucjena) i *malware* (zlonamjerni program). Iz samoga imena vidljivo je da je riječ o zlonamjernom kodu koji za svrhu ima financijsku iznudu. *Ransomware* možemo podijeliti na dvije osnovne skupine koje su trenutačno u opticaju. Radi se o kripto *ransomwareu* i *locker ransomwareu* koji, premda imaju sličan cilj, u načinu funkcioniranja potpuno se razlikuju. Kripto *ransomware* za svrhu ima šifrirati osobne žrtvine podatke i time onemogućiti pristup datotekama i dokumentima, a *locker ransomware* cilja na zaključavanje računala kako se žrtva njime ne bi mogla koristiti. [1]

Povijest *ransomware* zlonamjernoga koda za financijsku ucjenu seže u 1989. godinu, kada je kreiran trojanski konj AIDS[1]. Iako AIDS nije postigao nikakve značajne uspjehe kao *ransomware* zbog ograničene razvijenosti kriptografije i internetskoga sustava naplate, započeo je eru *ransomware* zlonamjernoga koda. Godine 2011. pojavljuju se prvi *locker ransomware* zlonamjerni programi, koji su bili znatno agresivniji način ucjene jer su korisnicima zaključavali praktično sve funkcionalnosti grafičkoga sučelja operativnoga sustava, osim onih potrebnih da provedu uplatu tražene svote. Najčešće se uplata obavljala pomoću elektroničkih vaučera u iznosu od 150 do 200 dolara. *Locker ransomware* uglavnom se oslanjao na psihološku komponentu i društveni inženjering, plašeći korisnika zaraženoga sustava kaznama, nemogućnošću rada na računalu, lažnim policijskim prijavama s upozorenjima da je prekršio zakon i sl. Nedostatak *locker ransomware* bila je u žrtvinoj sposobnosti da, ako prepozna pokušaj prijevare, jednostavno reinstalira OS ili instalira sigurnosnu programsku potporu. [1][2]

Locker ransomware bio je nažalost samo korak prema kreiranju znatno opasnijih i za korisnike vrlo neugodnih kripto *ransomwarea* koji su se pojavili 2013. godine. Za razliku od kripto *ransomwarea* s kraja 80-ih godina prošloga stoljeća, novi kripto *ransomware* zlonamjerni programi imali su na raspolaganju moćne kriptografske algoritme poput AES-a, 3DES-a i RSA.[3] Pored toga, virtualna valuta *bitcoin*, kojom su se mogle provoditi anonimne financijske transakcije, pokazala se kao izvrsno rješenje za iznuđivače otkupnine. [4] Upravo zbog svoje učinkovitosti ova vrsta *ransomwarea* preuzima primat nad drugim vrstama *ransomware* zlonamjernoga koda. [1]



Slika 1. Omjeri različitih tipova *ransomwarea* u proteklih deset godina[2]

[1] AIDS - Aids Info Disk ili PC Cyborg Trojan, je trojan koji je zamijenio AUTOEXEC.BAT datoteku
[2] Symantec The evolution of ransomware.pdf

Locker ransomware zlonamjerni je programski kod koji preuzima kontrolu nad zaraženim računalom na način da najčešće blokira korisničko sučelje, omogućujući jedino minimalnu interakciju s prozorom samog *ransomware*a. Tako, na primjer, korisnik može u prozor zlonamjernoga programa samo upisati brojeve koda za otključavanje računala koji je prethodno otkupio od napadača. *Locker ransomware* benigniji je tip *ransomwarea* s obzirom na to da ne čini štetu računalnom sustavu ili podatcima u njemu, nego se uglavnom zasniva na psihološkom momentu i strahu korisnika. Katkada je dovoljno napraviti *restore* operativnoga sustava iz prethodne nezaražene slike (*Image*). Stoga je način na koji *locker ransomware* djeluje znatno manje destruktivan od kripto *ransomwarea*.

Kripto *ransomware* za razliku od *locker ransomwarea* ima sasvim drugačiji mehanizam funkcioniranja. Nakon što zarazi računalo, ova vrsta *ransomware* zlonamjernoga koda pokušava ostati pritajena dokle god ne pretraži računalo u potrazi za svim dokumentima koji bi korisniku mogli biti važni i korisni. Nakon skeniranja, *ransomware* kriptira pronađene datoteke u samo nekoliko minuta. U trenutku kad su sve datoteke kriptirane i onemogućen im je pristup, kripto *ransomware* završava fazu prikrivanja i obavješćuje korisnika da su mu podatci kriptirani, odnosno oteti, te traži otkupninu za otključavanje datoteka. [1][5]

II.KRIPTO *RANSOMWARE*

Kripto *ransomware* zlonamjerni program ne bira žrtve, nego pokušava zaraziti što više računala s ciljem da od što više žrtava napada iznudi otkupninu. Stoga se u cijelom svijetu broj žrtava napada kripto *ransomware* bilježi u milijunima. Mete napada podjednako su privatni korisnici, poslovni korisnici i razne javne institucije. Najranjivija skupina korisnika upravo su privatni korisnici, koji najčešće nemaju dovoljno znanja da se uhvate u koštac s prijetnjom. Također, vrlo je malo privatnih korisnika osviješteno u pogledu kreiranja sigurnosnih kopija svojih podataka. Međutim, privatni korisnici koji imaju sigurnosne kopije na prijenosnim tvrdim diskovima ili kućnim NAS (*Network Attached Storage*) rješenjima nisu potpuno sigurni od napada kripto *ransomwarea* s obzirom na to da neke inačice toga zlonamjernoga programskoga koda mogu kriptirati podatke i sigurnosne kopije koje se nalaze na dijeljenim mrežnim diskovima. [5] [6]

Kad govorimo o poslovnim subjektima, tvrtke kojima je poslovanje vezano uz digitalne podatke, a takve su u većini, uglavnom imaju implementirane sustave kreiranja sigurnosnih kopija i planove oporavka u slučaju katastrofe. No najčešće ta rješenja uključuju samo ključne podatke za poslovanje tvrtke te sigurnosne kopije poslužitelja, ali ne i korisničke podatke spremljene na lokalnim računalima zaposlenika. Iz svega toga razvidno je da ni tvrtke nisu pošteđene napada kripto *ransomwarea*. Kad govorimo o javnim institucijama, ni one nisu pošteđene napada s obzirom na to da se tvorci ove vrste zlonamjernoga programskoga koda osjećaju dovoljno sigurnima da su izvan dosega zakona. Tako imamo primjere kada su šifrirani i oteti službeni podatci i podatkovne baze policije, ministarstava, gradskih uprava, fakulteta i drugih institucija, od kojih su neke platile traženu otkupninu kako bi došle do svojih ključnih podataka. Rezultati istraživanja pokazuju da je vrlo velik udjel javnih institucija među zaraženim korisnicima u RH, tj. 85 % od analiziranih napada. Niti jedna institucija od analiziranih nije platila otkupninu. Otkupninu su plaćali uglavnom pojedinci ili manje tvrtke.

A. Ugrožene računalne platforme

Kripto *ransomwareom* uglavnom se napadaju računala sa Microsoft Windows operativnim sustavima zato što operativni sustavi temeljeni na Microsoft OS-u drže oko 90 % tržišta operativnih sustava. [7] Zabilježeni su i napadi na *Unix* platforme, iskorištavajući sigurnosne propuste u *JavaScript* jeziku, te su funkcionirali na razini internetskoga preglednika umjesto iskorištavanja sigurnosnih propusta operativnoga sustava. Takva vrsta napada nije bila toliko opasna za razliku od napada iskorištavanjem nedostataka *OS-a*. Međutim, na početku ožujka 2016. zabilježen je i prvi uspješan napad kripto *ransomwarea* na *Apple OS X* operativni sustav. Radilo se o *ransomwareu* nazvanom *KeRanger* koji je, iskoristivši važeći *Apple Developer* certifikat za instalaciju na lokalno računalo, uspješno kriptirao korisničke podatke. [8] Premda se radi o trenutačno jedinstvenom slučaju napada na Apple OS X platformu, novi napadi na taj operativni sustav sasvim su izvjesni.

Kreiranje *ransomwarea* vrlo je kompleksno na tržišno dominantnim mobilnim platformama *Google Android i Apple iOS*, koje također spadaju u skupinu *Unix-ida*. Te mobilne platforme naslijedile su od *Unixa* relativno visoku razinu sigurnosti operativnoga sustava. Kad govorimo o platformi *Google Android*, koja je znatno otvorenija od platforme *Apple iOS*, tvorci *ransomwarea*, radi restrikcija samoga operativnoga sustava, pribjegli su prije svega kopiranju ideja *locker ransomwarea*, koji je zapravo samo onemogućavao grafičko sučelje mobilnoga uređaja umjesto da enkripcijom otme korisničke podatke. [1] Međutim, 2014. pojavio se *Adroid.Simplocker* kripto *ransomware,* prvi kojim su se pokušali implementirati mehanizmi zaraze i enkripcije na *Android* mobilnoj platformi, ali s ograničenim mogućnostima šifriranja samo korisničkih podataka spremljenih na eksternoj *microSD* kartici ili emuliranoj SD kartici na internoj memoriji uređaja [9]

B. Vektori širenje ransomwarea

Glavni je problem sa širenjem *ransomwarea* u tome što se koriste zapravo svi aktualni vektori širenja zaraze, pa se zbog toga vrlo teško zaštititi od ove vrste zlonamjernih

programa. Dovoljan je samo jedan pogrješan klik, posjet sumnjivoj stranici ili sigurnosni propust u sustavu da dođe do zaraze i kompromitacije cijeloga sustava. Kreatori *ransomwarea* najčešće se koriste sljedećim načinima zaraze:

- *drive-by-download* tehnike
- distribucijom nepoželjne e-pošte
- *botnet* metodama
- društvenim inženjeringom i *malvertisementom*

Drive-by-download računalnog sustava inficira sustav posjećivanjem web stranice na kojoj se nalazi maliciozni kod (*Exploit kit*) koji, ukoliko pronađe sigurnosni propust u sustavu, isporučuje maliciozni sadržaj (*Payload*) koji inficira računalo. [10]

Distribucija nepoželjne e-pošte (*Spam*) koja sadrži maliciozni privitak, koji nakon preuzimanja i pokretanja zaražuje sustav malicioznim kodom, jedan je od najraširenijih načina kompromitacije računalnog sustava. [11]

Botnet metoda zaraze *ransomware* malicioznim kodom računalnog sustava koji je već dio *botnet* mreže podrazumijeva da zlonamjerna strana već ima potpunu kontrolu nad kompromitiranim sustavom te može u bilo kojem trenutku instalirati *ransomware* program.

Ipak najčešći je način zaraze korištenje društvenoga inženjeringa i *malvertisementa* [12] implementiranoga u obliku zlonamjernoga privitka ili internetske poveznice na zlonamjerni sadržaj, u nepoželjnoj (*Spam*) elektroničkoj pošti, što je pokazalo i provedeno istraživanje. Svi korisnici koji su prepoznali vektore napada zaraženi su putem društvenoga inženjeringa. Širenje zlonamjernoga koda društvenim inženjeringom, koji iskorištava neznanje, radoznalost i naivnost korisnika, nepresušan je i pouzdan izvor novih zaraza računala. Teme koje su se najčešće koristile u analiziranim napadima u zlonamjernim *spam* porukama bili su računi za električnu struju ili telekom operatera, obavijesti o povratu poreza, policijskim kaznama i sl. [13]

Kako bi na napadnute korisnike povećali psihološki pritisak, tvorci kripto *ransomware* često postavljaju krajnji rok za uplatu otkupnine, nakon kojega će podatci biti nepovratno izgubljeni odnosno uništeni. [14]

Rezultati analize posljednjih napada pokazuju da *ransomware* zlonamjerni programski kod napade u slučaju *drive-by* zaraze izvodi u dvije faze, slično taktici koju su proveli tvorci *Stuxneta*. [15] U prvoj fazi napada otimaju se podatci iz internetske povijesti (*History*) i dostavljaju napadaču. Nakon analize dobivenih podataka i uvida u to koje su stranice najposjećenije, napadač postavlja zlonamjerni kod upravo na tim stranicama jer je velika vjerojatnost da će upravo te stranice posjetiti najveći broj korisnika.

Ključni element svakoga uspješnoga kripto *ransomwarea*, a koji je specifičan za ovu vrstu zlonamjernoga koda, zapravo su mehanizmi i odabir enkripcije koji se koriste u procesu otimanja korisničkih podataka. Kripto *ransomware* podjednako se koristi i simetričnim i asimetričnim enkripcijskim algoritmima. Kada se primjenjuju simetrični kriptografski algoritmi, poput AES-a, isti se ključ koristi za kriptiranje i dekriptiranje datoteka. Kripto *ransomware* koji se koristi simetričnim algoritmom, kontaktira *C&C*[3] poslužitelja da mu isporuči željeni ključ za kriptiranje ili ga sam generira na zaraženom stroju te ga zatim pošalje *C&C* poslužitelju. Velika je prednost simetričnih algoritama u brzini kojom mogu šifrirati golemu količinu podataka zbog male dužine ključa (128- ili 256-bitni) uz minimalne procesorske i memorijske zahtjeve, a te su karakteristike ključne za uspješno prikrivanje kripto *ransomwarea*.

Za razliku od simetričnih algoritama, neki kripto *ransomware* zlonamjerni programi koriste se algoritmima javnoga ključa (*Public key encryption*) tj. asimetričnim algoritmima, koji za enkripciju upotrebljavaju dva ključa: javni i privatni. Kripto *ransomware* koji ima implementiran asimetrični enkripcijski algoritam kao što je RSA, koristi se javnim ključem za kriptiranje korisničkih podataka na kompromitiranom stroja, a privatni ključ posjeduje ucjenjivač koji je potreban da bi se kriptirani podatci dekriptirali. Ako korisnik plati otkupninu, napadač šalje privatni ključ žrtvi kako bi podatci mogli biti dekriptirani. Iz perspektive kibernetičkih kriminalaca, glavni je nedostatak korištenja asimetrične kriptografije u kripto *ransomwareu* sporost i veliki zahtjevi tih algoritama za procesorskim vremenom i radnom memorijom. Razlog tome je što se za enkripciju koriste vrlo dugačkim ključevima (do 4096-bitne dužine). Uz duže vrijeme šifriranja podataka i povećano opterećenje centralnog procesora, povećava se vjerojatnost da zlonamjerni kod bude otkriven prije nego što dovrši enkripciju svih datoteka.

U najnovijim inačicama kripto *ransomware* zlonamjernoga koda koriste se kombinacijom simetričnih i asimetričnih enkripcijskih algoritama. Najčešće se koristi AES simetrični algoritam, generiran lokalno na zaraženom računalu uz RSA ili ECDH (*Elliptic curve Diffie–Hellman*) asimetrične algoritme. Pritom se moćni i brzi simetrični AES algoritam koristi za šifriranje žrtvinih podataka, a zatim se sam AES ključ šifrira javnim RSA ili ECDH ključem koji je poslao *C&C* poslužitelj napadača. Tako je osigurano da korisnik zaraženoga računala ne može doći od simetričnoga AES ključa, premda se on nalazi u svakoj šifriranoj datoteci. Jedini način da se AES

[3] C&C - *command and control server*

ključ dekriptira jest da napadač, nakon uplate otkupnine, žrtvi pošalje privatni RSA ili ECDH ključ. CTB[4] kripto *ransomware* na još elegantniji način kombinira upotrebu simetričnih i asimetričnih algoritama za šifriranje. Naime, CTB kripto *ransomware* već sadržava u sebi javni ključ pomoću kojega šifrira lokalno generirani simetrični AES ključ. Tako ovaj *ransomware* može započeti kriptiranje datoteka i prije nego što kontaktira *C&C* poslužitelja. Tek kad je gotov sa šifriranjem podataka, ovaj kripto *ransomware* kontaktira *C&C* poslužitelja i obavješćuje napadača da je napad uspio. [1][3]

Većina kripto *ransomwarea* koristi se službenim, u operativni sustav integriranim enkripcijskim API-jima. Tako napadači ne moraju zlonamjernim programom propagirati dodatne datoteke niti osmišljavati svoja enkripcijska rješenja. Kripto *ransomware* zlonamjerni programi stoga se najčešće koriste Microsoft *CryptoAPI* ili *OpenSSL*. Usporedbom kripto *ransomwarea* prema vrsti enkripcije kojom se koriste i načinu implementacije vidljiv je napredak u programiranju novih inačica ove vrste zlonamjernih programa. [3]

Broj ciljanih ekstenzija odnosno tipova podataka za kojima se putem ovih zlonamjernih programa traga kreće se od 70 do 200. Najčešće se napadaju slike, audio-video datoteke i, naravno, razni *office* dokumenti poput .doc, .xls, .pdf. Najnovije vrste kripto *ransomwarea*, za razliku od prethodnika koji su primarno ciljali privatne korisnike, sve više napadaju tvrtke. Taj trend vidljiv je iz odabira ekstenzija, među kojima se sve češće nalaze i razne CAD, financijske i *backup* ekstenzije. [16] Razlog je u tome što je vrijednost podataka za bilo koju tvrtku vrlo velika, pa je vrlo vjerojatno da će iste biti spremne platiti znatno veću otkupninu za povrat kontrole nad datotekama.

III. KRONOLOGIJA I ANALIZA NAPADA

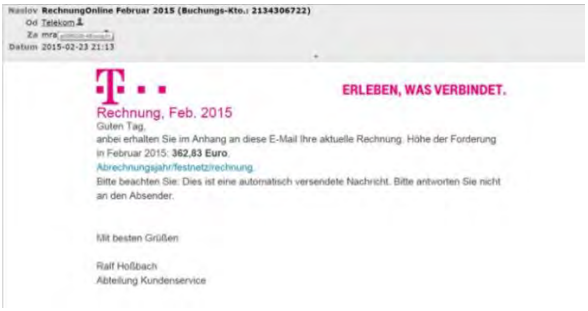
Kripto *ransomware* zlonamjerni programski kod pripada skupini trojanskih konja, jer mu je potrebna suradnja odnosno akcija buduće žrtve da se infiltrira u računalni sustav i provede svoju zlonamjernu aktivnost. Stoga su vektori širenja infekcije najčešći metodom *drive-by*, zaraženim privitcima e-pošte, instalacijom pomoću *trojan-downloadera* koji je već kompromitirao žrtvin računalni sustav te infiltracijom kroz RDP[5] servis.

Budući da se u slučaju analiziranih napada u RH uglavnom radilo o tipu *sleeper* (spavač) kripto *ransomwarea* koji zlonamjernu radnju nije proveo čim je zarazio žrtvino računalo nego je, skriven u sustavu i neaktivan, čekao unaprijed zadani termin za svoj zlonamjerni zadatak. Upravo je zbog takva ponašanja ovoga kripto *ransomware* zlonamjernoga koda vrlo teško sa stopostotnom sigurnosti utvrditi vektor zaraze sustava. Međutim, u slučaju promatranoga napada sve indicije upućuju na *drive-by* vektor zaraze.

Drive-by vektor zaraze računalnoga sustava inficira sustav posjećivanjem *web* stranice na kojoj se nalazi zlonamjerni kod. U više od 80 % slučajeva radi se o legitimnim stranicama koje su i same postale žrtve. Nakon pristupanja zaraženoj legitimnoj stranici, zlonamjerni kod preusmjerava računalo na stranicu pod kontrolom napadača, na kojoj se nalazi sofisticirani zlonamjerni kod (*exploit kit*) za skeniranje žrtvina računala u potrazi za sigurnosnim propustom. Najčešće se radi o sigurnosnim propustima u neažuriranim inačicama Adobe Flash i Oracle Java dodatcima za internetske pretraživače. U slučaju da zlonamjerni paket za traženje sigurnosnih slabosti sustava pronađe propust, iskorištava ga i na žrtvino računalo isporučuje zlonamjerni paket (*payload*) te ga zarazi zlonamjernim kodom.[13]

A. Kronologija napada

Kronologija napada na analizirane poslovne subjekte pokazuje upravo takav način zaraze korisničkih računala s kripto *ransomware* kodom. S adrese pt33951526@unizg.hr (Telekom), 23. veljače 2015. dolazi masovna *spam* poruka na sve ciljane korisničke adrese, naslova *RechnungOnline* Veljača 2015 (*Buchungs-Kto.: 2134306722*) tj. *online* račun za 2. mjesec 2015. u iznosu od 362,83 eura. U samoj poruci nalazi se poveznica *Abrechnungsjahr/festnetz/rechnung* s putanjom na *insidetechbd.com/RechnungFebruar2015* koju neoprezni korisnici otvaraju.


Naslov RechnungOnline Februar 2015 (Buchungs-Kto.: 2134306722)
Od Telekom
Za mrx
Datum 2015-02-23 21:13

ERLEBEN, WAS VERBINDET.

Rechnung, Feb. 2015

Guten Tag,
anbei erhalten Sie im Anhang an diese E-Mail Ihre aktuelle Rechnung. Höhe der Forderung in Februar 2015: 362,83 Euro.
Abrechnungsjahr/festnetz/rechnung
Bitte beachten Sie: Dies ist eine automatisch versendete Nachricht. Bitte antworten Sie nicht an den Absender.

Mit besten Grüßen

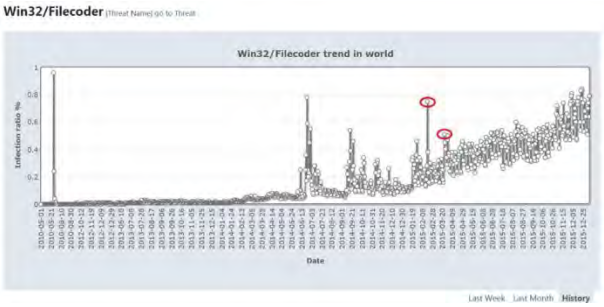
Ralf Hoßbach
Abteilung Kundenservice

Slika 2. *Spam phishing* e-poruka u kojoj se korisnike preusmjerava na zaraženu stranicu *insidetechbd.com*

Pritom su njihovi internetski pretraživači preusmjereni na zaraženu stranicu *insidetechbd.com*, koja se nalazi u gradu Dhaki u Bangladešu. Na toj *web* stranici nalazi se zlonamjerni kod kojim se provodi *drive-by* napad i isporučuje zlonamjerni paket s kripto *ransomware* programskim kodom na korisnička računala bez znanja korisnika i bez potrebe za administratorskim pravima kako bi zarazio korisničko računalo. Na računala se instalira kripto *ransomware* porodice *Filecoder*. Jedina naznaka zaraze računala zlonamjernim kodom vidljiva je u

[4] CTB Locker ransomware (Curve-Tor-Bitcoin Locker) koji koristi TOR za skrivanje komunikacije s C&C
[5] RDP - Remote Desktop Protocol

kratkotrajnom sumnjivom mrežnom prometu tijekom kojega zaražena računala generiraju mrežni promet s desetak tisuća UDP (*User Datagram Protocol*) paketa veličine 64 do 100 bajta, kakvim se najčešće koristi *DDOS* (*Distributed Denial-of-service*) napad. Međutim, svrha generiranja takva prometa u promatranom slučaju nije u potpunosti razjašnjena nego je samo pokazatelj potencijalne zaraze sustava. Nakon generiranja spomenutoga mrežnoga prometa, *Filecoder* kripto *ransomware* mirovat će na svim zaraženim strojevima do srijede 25. veljače 2015. u 11.30 sati, kada će se pokrenuti zlonamjerni kod na zaraženim poslovnim sustavima diljem Hrvatske i šifrirati golema količina korisničkih podataka u samo nekoliko minuta.



Slika 3. Vremenska crta napada *Filecoder ransomwarea* s označenim razdobljima napada na sustave u RH[6]

S obzirom na to da *Filecoder* kripto *ransomware* primarno cilja na poslovne subjekte, ovaj *ransomware* zlonamjerni kod cilja na dijeljene direktorije jer je pretpostavka da poslovni korisnici dijele poslovne dokumente na dijeljenim mrežnim lokacijama. Lokalno pohranjeni podatci na korisničkim računalima koji nisu dijeljeni nisu kriptirani, premda se radilo o tipovima datoteka čije je ekstenzije *Filecoder* kripto *ransomware* kod na dijeljenim lokacijama šifrirao. Također, bitno je napomenuti da je za uspješno pristupanje i šifriranje podataka na dijeljenim mrežnim lokacijama ovoj inačici *Filecoder* kripto *ransomwarea* preduvjet da te mrežne lokacije budu dijeljene na korisničkim računalima, a novijim inačicama kripto *ransomware* zlonamjernoga koda to nije potrebno, nego je dovoljno da korisnik ima prava čitanja i pisanja po mrežnim lokacijama.

Šteta nad podatcima kod jednoga analiziranoga poslovnoga korisnika iznosila je više od 80 Gb podataka na sedam različitih korisničkih računala te na 40-ak Gb podataka dvaju poslužitelja na kojima su bili dijeljeni zajednički korisnički podatci.

Nakon kriptiranja mrežno i lokalno pohranjenih podataka sva skeniranja pokazivala su čiste i od zlonamjernih programa sigurne sustave. Razlog lažno negativnih rezultata skeniranja zaraženih sustava ima uporište u činjenici da se radilo o *zero-day* napadu tj. o napadu dotad nepoznate inačice zlonamjernoga programa, koji iz toga razloga nije bio identificiran niti od jednoga proizvođača sigurnosnih programa. Napadnuti korisnici koristili su se ažurnim Microsoftovim inačicama: *Endpoint Client protection* 2012, *Windows Defender*, *Malwarebytes Anti-malware* i ESET *Nod32*.

B. Analiza napada

Analizirana su 23 napada na 11 poslovnih subjekata i institucija u Republici Hrvatskoj te na nekoliko pojedinaca. Glavni vektor napada kripto *ransomwarea* bila je usluga e-pošte. Sustav je bio kompromitiran i zaražen trojanskim konjem, koji zahtjeva akciju korisnika kako bi zarazio računalo, bilo da se radi o *drive-by* tipu zaraze, koja preusmjerava preglednik na zlonamjernim programom zaraženu stranicu, ili o otvaranju zaraženoga privitka s izvršnom datotekom u e-poruci. Ova vrsta zlonamjernoga koda podjednako napada male i velike informatičke sustave, a prema prikupljenim podatcima, ni sofisticirana višeslojna vatrozidna i sigurnosna rješenja, poput onih u financijskim institucijama, nisu bila pošteđena napada. Stoga je u svim promatranim slučajevima glavna crta obrane bio sustav za izradu sigurnosnih kopija (*Backup*). Niti u jednom od analiziranih primjera napada ne dolazi do zaraze samih poslužitelja nego samo korisničkih računala koja su pogonjena Microsoft Windows operativnim sustavima. Najčešće napadnuti operativni sustavi su Windows XP i Windows 7. To je logično s obzirom na to da je riječ o najraširenijim inačicama MS-ovih operativnih sustava. Sva napadnuta računala imala su antivirusnu zaštitu, a u većini slučajeva sustavi su bili zaštićeni vatrozidom.

Iz prikupljenih podataka vidljivo je da su tvorci kripto *ransomware* zlonamjernoga koda od žrtava tražili sličnu otkupninu, između 0,5 BTC-a i 3 BTC-a, koja je uglavnom bila izražena u protuvrijednosti američkoga dolara ili eura. Prosječna otkupnina iznosila je oko 2 BTC-a. Pozitivno je – što je vidljivo iz prikupljenih podataka, da niti jedna napadnuta tvrtka ili institucija nije platila traženu otkupninu, čime se dugoročno obeshrabruje one koji se služe kripto *ransomware* vrstom financijske iznude. Otkupninu su platila dva pojedinca iz analizirane skupine kojima su bila zaražena osobna a ne poslovna računala. Žrtvama koje su platile traženu otkupninu podatci su bili dekriptirani, čemu se zapravo ne treba čuditi s obzirom na to da je u interesu tvoraca *ransomwarea* da budu dosljedni kako bi ohrabrili postojeće i buduće žrtve da plate otkupninu za otete podatke.

Budući da je u gotovo svim promatranim napadima vektor zaraze bila usluga elektroničke pošte te da antivirusna rješenja nisu mogla prepoznati i zaustaviti kripto *ransomware* napade, poželjno je obranu sustava

[6] *http://www.virusradar.com/en/Win32_Filecoder/chart/history*

temeljiti na pravilno konfiguriranom vatrozidu koji je u mogućnosti nadzirati SSL (*Secure Sockets Layer*) promet kako bi uspješno blokirao sve e-poruke koje sadržavaju privitke s izvršnim ekstenzijama kao i odlaznu komunikaciju prema *C&C* poslužiteljima kibernetičkih kriminalaca. Kao dodatni oblik zaštite sustava od malicioznoga koda, preporučljivo je implementirati *Sandbox* sigurnosne mehanizme, koji mogu „detoniranjem" u kontroliranom okružju testirati potencijalno zlonamjeran programski kod ili maliciozne *web* stranice.

IV.ZAŠTITA OD KRIPTO *RANSOMWARE* ZLONAMJERNOGA KODA

Iako ne postoji idealno sigurnosno rješenje i sigurnosne politike koje bi omogućile ne samo stopostotnu sigurnost računalnih sustava poslovnih i privatnih subjekata nego istodobno i nesmetan korisnički rad, potrebno je napraviti odgovarajuće preventivne aktivnosti koje svaki korisnik i tvrtka / institucija mogu poduzeti kako bi se zaštitili od ove vrlo opasne vrste zlonamjernoga programskoga koda.

A. Sigurnosne aplikacije i vatrozidna zaštita

Antivirusna rješenja teško identificiraju nove kripto *ransomware* zlonamjerne programe obzirom na to da takvi programi na računalu zapravo ne poduzimaju (iz perspektive aktivnosti diskovnoga podsustava) neuobičajene radnje. Kripto *ransomware* zapravo samo pristupa podatkovnom sustavu te provodi radnje čitanja i pisanja. Međutim, rezultati novijih istraživanja pokazali su da ipak postoje načini razlikovanja zlonamjerne aktivnosti diskovnoga podsustava od normalne aktivnosti [17]. Tako analize pokazuju da u slučaju kripto *ransomware* napada diskovni podsustav pokazuje povišenu i vrlo veliku aktivnost, poput velikoga broja sličnih poziva I/O funkcija kao što su enkripcijska funkcija i funkcija brisanja datoteka u vrlo kratkom vremenu. Također je moguće pomnim nadzorom *MFT* (*Master File Table*) tablice detektirati kreiranje, enkripciju i brisanje datoteka. Kod enkripcije datoteka vidljiv je velik broj *MFT* zapisa sa šifriranim sadržajem u $DATA atributu šifriranih datoteka koje ne dijele istu putanju direktorija. Budući da gotovo svi kripto *ransomware* zlonamjerni programi tijekom napada stvaraju ovakvu aktivnost diskovnoga sustava [17], moguće je (za razliku od standardnih metoda detektiranja zlonamjernoga koda, koje uglavnom funkcioniraju samo za poznate vrste kripto *ransomwarea*) stvoriti znatno moćnije aplikacije za detekciju ove vrste zlonamjernih programa koje bi mogle detektirati i blokirati čak i nove, nepoznate vrste kripto *ransomwarea*.

Uz praćenje aktivnosti diskovnoga podsustava, dodatnu razinu zaštite predstavlja korištenje posebno nadziranih mamaca u obliku datoteka za zlonamjerni program najinteresantnijim diskovnim lokacijama, koje su za kripto *ransomware* glavni ciljevi. [17] Te datoteke mamci (*decoy*) bile bi u stalnoj komunikaciji s *anti-ransomware* sigurnosnom aplikacijom, pa ako bi se i dogodio pokušaj enkripcije tih datoteka, sigurnosna aplikacija automatski bi blokirala pristup diskovnom podsustavu.

Iako je antivirusna industrija na početku bila zatečena i nespremna za napade kripto *ransomwarea*, stvari su se znatno promijenile na bolje, tako da u ovom trenutku postoje mnoga aplikativna rješenja za zaštitu računalnoga sustava od kripto *ransomwarea*. Jedna od najpopularnijih aplikacija za zaštitu od ove vrste zlonamjernih programa svakako je *Cryptomonitor,* koji preporučuje i Nacionalni CERT [18]. *Cryptomonitor* štiti računalni sustav ne samo nadziranjem *CryptoAPI-ja* nego i *Entrapment* i *Count* zaštitnim metodama [19], koje predstavljaju prethodno opisane pristupe zaštite diskovnoga podsustava od enkripcije podataka. Korisno je napomenuti da je u trenutku pisanja ovoga rada profesionalnu inačicu *Cryptomonitor* aplikacije preuzela i preimenovala tvrtka Malwarebytes. Nova aplikacija također je potpuno besplatna i uključuje sve nabrojene vrste zaštite sustava pod novim imenom *Malwarebytes Anti-ransomware*.

Zaštita i nadzor mrežnoga prometa također ima vrlo značajnu ulogu u prevenciji napada svih vrsta zlonamjernoga koda, uključujući i kripto *ransomware*. Stoga je preporučljivo primijeniti inspekciju *HTTP, FTP, IMAP, SMTP, POP3, CIFS/Netbios i TCP* mrežnih tokova [20] i onemogućiti transfer Zip datoteka zaštićenih zaporkom, *MS Office* dokumenata koji sadržavaju makro naredbe (VBA 5 ili više) i datoteka s izvršnim ili arhivskim ekstenzijama zip, .jar, .tar, .7z, .msi, .com, .exe, .scr, .bat, .js, .jse, .vb, .vbe, .wsf, .wsh, .cmd i drugih. U vatrozidu je također preporučljivo blokirati pristup poznatim *C&C* poslužiteljima kojima se koriste zlonamjerni programi i onemogućiti *TOR* (*The Onion Router*) komunikaciju kojom se neki *ransomware* zlonamjerni programi koriste za komunikaciju sa svojim *C&C* poslužiteljima kako bi pribavili ključeve potrebne za enkripciju žrtvinih podataka. [21]

Nije na odmet postaviti prikazivanje skrivenih (poznatih) ekstenzija datoteka kako bi se lakše primijetile ekstenzije poput *.exe. Korisnik također može zabraniti pokretanje datoteka s pokretačkim ekstenzijama iz mapa poput *Temp, Appdata* ili *LocalAppData*, kojima se neki kripto *ransomware* zlonamjerni programi koriste kao ishodišnima za svoje zlonamjerne izvršne datoteke. S obzirom na to da se neki *ransomware* alati oslanjaju na Windows *RDP* (*Remote Desktop Protocol*) kako bi zarazili računalo, preporučljivo je taj servis ugasiti ako ga korisnik ne upotrebljava. Vrlo je bitno da se poslužiteljske i korisničke aplikacije i operativni sustavi održavaju redovitim nadogradnjama na aktualne programske inačice. Korištenje programske potpore nadograđene na aktualnu inačicu smanjuje rizik od infekcije računalnoga sustava

iskorištavanjem kojega od sigurnosnih propusta u operativnom sustavu ili aplikacijama. Uz sve navedeno, kao oblik proaktivne zaštite preporučuje se koristiti kompleksne i slojevite sigurnosne aplikacije renomiranih antivirusnih kompanija na poslužiteljima i korisničkim računalima, koje radi svoje ekspertize i resursa mogu brže i kvalitetnije reagirati na pojavu novih, nepoznatih tipova razmatranoga zlonamjernoga koda.

B. Sigurnosne kopije

Sustavi za izradu sigurnosnih kopija zadnja su ali najvažnija crta obrane od potencijalnoga katastrofičnoga gubitka podataka. Stoga ne treba isticati koliko je važno implementirati sustav za izradu sigurnosnih kopija računalnoga sustava i korisničkih podataka jer ti sustavi ne štite korisnike i tvrtke samo od gubitaka podataka uzrokovanih kripto *ransomware* napadom nego i od svih ostalih izvora uništenja podataka, uključujući i elementarne nepogode i prirodne katastrofe.

C. Poslovna sigurnosna politika

Kao što smo vidjeli u prethodnim poglavljima, glavni vektori zaraze poslovnih i privatnih informatičkih sustava raznim tipovima zlonamjernih programa, uključujući i kripto *ransomware,* najčešće se ostvaruju *drive-by* napadom, zaraženim privitkom elektroničke pošte i korištenjem *crackova* za piratizaciju popularnih aplikacija i igara. Svim spomenutim tipovima napada potrebna je potrebna asistencija needuciranoga zaposlenika kako bi kripto *ransomware* zarazio računalni sustav. Upravo se zato napadači masovno koriste društvenim inženjeringom u svrhu širenja svojih zlonamjernih programa. Stoga je ključan moment implementacija sustava upravljanja informacijskom sigurnosti (*ISMS - Information Security Management System*) standarda ISO 27001/27002 te podizanje općega informatičkoga znanja zaposlenika. Potrebno je organizirati sigurnosne treninge i seminare na kojima će ih se upoznati s prijetnjama na internetu, prevencijom kompromitacije sustava, sigurnim radom na računalu i odgovarajućom reakcijom u slučaju napada na računalni sustav. [22]

V. ZAKLJUČAK

Prema rezultatima ove analize, pojava kripto *ransomware* zlonamjernih programa, koja je pogodila cijeli svijet u protekle tri godine, pokazala se kao prijetnja s ozbiljnim posljedicama za privatne a osobito za poslovne korisnike i njihove IT odjele. Sigurnosne kompanije i dalje zaostaju za inventivnošću, agilnošću i visokom motiviranošću tvoraca *ransomwarea,* stoga je vrlo vjerojatno da će se trend širenja zaraze računala s poznatim i novim inačicama kripto *ransomwarea* ne samo nastaviti nego i povećati. U tome nema ništa neobično s obzirom na to da su računalni kriminalci pronašli vrlo unosan način zarade – financijskom iznudom, prepoznavši značaj privatnih i poslovnih digitalnih informacija za suvremeno tržište. Jedini način da se privatni i poslovni subjekti zaštite od opasnosti kakvu predstavlja kripto *ransomware* zlonamjerni programi nedvojbeno je sveobuhvatna strategija kojom bi se trebale objediniti poslovne sigurnosne politike, sigurnosne aplikacije i hardveri, uz podizanje ne samo informatičke pismenosti korisnika nego i ukupne svijesti o novim opasnostima digitalnoga doba na primjerenu razinu.

LITERATURA

[1] Symantec, Evolution of Ransomware, 2015.

[2] Wikipedia, https://en.wikipedia.org/wiki/Ransomware , 08.12.2015.

[3] Kotov V. , Rajpal M., Bromium Report, Understanding Crypto-Ransomware, 2014.

[4] Comparitech, https://www.comparitech.com/blog/information-security/the-history-of-ransomware/ 17.02.2016.

[5] McAfee, Cryptolocker Update RevD , 2015.

[6] Sophos, https://www.sophos.com/support/knowledgebase/120797.aspx , 2014.

[7] NetMarketShare, https://www.netmarketshare.com/operating-system-market-share.aspx?qprid=10&qpcustomd=0 , 10.01.2016.

[8] Scott J., Spaniel D. , Cerber & Keranger: The latest Examples of Weaponized Encryption, 2016.

[9] Lipovsky R., Štefanko L.; Braniša G., The Rise of Android Ransomware, 2016.

[10] Sophos, https://blogs.sophos.com/2014/03/26/how-malware-works-anatomy-of-a-drive-by-download-web-attack-infographic/ 12.12.2015.

[11] BoxCryptor, https://www.boxcryptor.com/en/blog/malware-in-email-attachments , 27.02.2016.

[12] MalwareBytes, https://www.malwarebytes.org/articles/hacking-your-head-how-cyber-criminals-use-social-engineering/ 13.02.2016.

[13] Xing X., Meng W., Lee B. Weinsberg U., Sheth A., Understanding Malvertising Through Ad-Injecting Browser Extensions 2015.

[14] Symantec, http://www.symantec.com/connect/blogs/ransomware-extorting-money-panic-and-pressure , 22.12.2015.

[15] David E. Sanger: Confront and Conceal: Obama's Secret Wars and Surprising Use of American Power, Crown Publishing Group, New York 2012.

[16] McAfee, Combating ransomware 2016.

[17] Kharraz A., Robertson W., Balzarotti D., Bilge L., Kirda E., Cutting the Gordian Knot: A Look Under the Hood of Ransomware Attacks, 2014.

[18] Nacionalni CERT, www.cert.hr/node/26023 22.01.2016.

[19] Bleeping Computer, http://www.bleepingcomputer.com/forums/t/572146/cryptomonitor-stop-all-known-crypto-ransomware-before-it-encrypts-your-data/ 14.1.2016.

[20] Dell, http://en.community.dell.com/techcenter/security/network-mobile-email/w/wiki/7895.best-practices-to-protect-against-cryptowall-and-cryptolocker-sw12434 14.1.2016.

[21] Lumension, Ransomware: The New Normal in Malware, 2015.

[22] Alessandrini A, Ransomware Hostage Rescue Manual, 2015.


ERICSSON
GRADSKI ŽIVOT
SADA, SVUDA OKO NAS.
Ukoliko želimo, u umreženom društvu prednosti života u gradu su svuda oko nas. Spajanjem izoliranih mjesta s najnaseljenijima više ne trebamo živjeti u gradu da bi uživali u njegovim dobrobitima.
Širom svijeta naša tehnologija i usluge omogućuju promjene i pomažu ljudima koji te promjene pokreću. Prilike su sada svuda oko nas - iskoristite ih!
www.ericsson.com
www.ericsson.hr

UVIJEK DOBIVATE VIŠE

SPOJITE SVE USLUGE U VAŠEM DOMU U **MAGENTU 1** I UĐITE U SVIJET DODATNIH BESPLATNIH POGODNOSTI:

- ✓ razgovarajte koliko god želite fiksnim telefonom prema svima, a mobitelom unutar HT mreže
- ✓ doživite vrhunsko iskustvo surfanja na još bržem fiksnom i 4G mobilnom internetu
- ✓ iskoristite 500 besplatnih minuta za razgovore s fiksnoga telefona prema fiksnim brojevima u zemljama EU-a



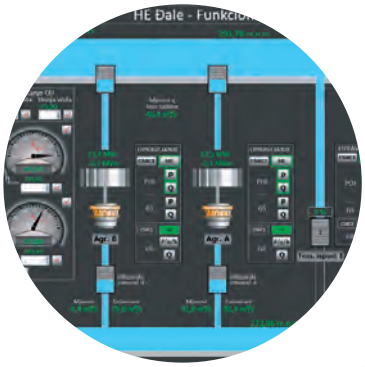
Otkrijte još više na **hrvatskitelekom.hr**, T prodajnom mjestu ili pozivom na **0800 9000**.
Za vaš posao potražite **Magenta 1 Business**.

T..

ŽIVJETI ZAJEDNO

tradition.knowledge.responsibility.

tradicija.
znanje.
odgovornost.









KONČAR

www.koncar.hr

sponsors

golden sponsors

KONČAR

ERICSSON
Ericsson Nikola Tesla

T

silver sponsor

INFODOM

bronze sponsors

in2

STORM
GRUPA

Hewlett Packard
Enterprise

OiV

HEP

sponsors

vip

ALTPRO d.o.o.
ISTRAŽIVANJE - RAZVOJ - PROIZVODNJA

mTEH mjerne tehnologije d.o.o.

KING ICT
INFORMATION & COMMUNICATION TECHNOLOGIES

SELMET

Microsoft

PROCADD
www.ib-procadd.si

nomen d.o.o.
MES and ERP Systems since 1994

DANIELI AUTOMATION

NOKIA

MIPRO
Kružna 8/II, HR-51000 Rijeka, Croatia

phone: +385 51 423 984
mipro@mipro.hr • www.mipro.hr

ISBN 978-953-233-087-8